MONASH
BUSINESS
SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

Auxiliary Likelihood-Based Approximate Bayesian Computation in State Space Models

Gael M. Martin, Brendon P.M. McCabe, David T. Frazier, Worapree Maneesoonthorn and Christian P. Robert

April 2016

Working Paper 09/16

Auxiliary Likelihood-Based Approximate Bayesian Computation in State Space Models*

Gael M. Martin†, Brendan P.M. McCabe‡, David T. Frazier§,

Worapree Maneesoonthorn¶ and Christian P. Robert‖

April 27, 2016

Abstract

A new approach to inference in state space models is proposed, using approximate Bayesian computation (ABC). ABC avoids evaluation of an intractable likelihood by matching summary statistics computed from observed data with statistics computed from data simulated from the true process, based on parameter draws from the prior. Draws that produce a 'match' between observed and simulated summaries are retained, and used to estimate the inaccessible posterior; exact inference being feasible only if the statistics are sufficient. With no reduction to sufficiency being possible in the state space setting, we pursue summaries via the maximization of an auxiliary likelihood function. We derive conditions under which this auxiliary likelihood-based approach achieves Bayesian consistency and show that - in a precise limiting sense - results yielded by the auxiliary maximum likelihood estimator are replicated by the auxiliary score. Particular attention is given to a structure in which the state variable is driven by a continuous time process, with exact inference typically infeasible in this case due to intractable transitions. Two models for continuous time stochastic volatility are used for illustration, with auxiliary likelihoods constructed by applying computationally efficient filtering methods to discrete time approximations. The extent to which the conditions for consistency are satisfied is demonstrated in both cases, and the accuracy of the proposed technique when applied to a square root volatility model also demonstrated numerically. In multiple parameter settings a separate treatment of each parameter, based on integrated likelihood techniques, is advocated as a way of avoiding the curse of dimensionality associated with ABC methods.

Keywords: Likelihood-free methods, latent diffusion models, Bayesian consistency, asymptotic sufficiency, unscented Kalman filter, stochastic volatility.

JEL Classification: C11, C22, C58

*This research has been supported by Australian Research Council Discovery Grant No. DP150101728.

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

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

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

¶Melbourne Business School, University of Melbourne, Australia.

‖University of Paris Dauphine, Centre de Recherche en Économie et Statistique, and University of Warwick.

1 Introduction

The application of Approximate Bayesian computation (ABC) (or likelihood-free inference) to models with intractable likelihoods has become increasingly prevalent of late, gaining attention in areas beyond the natural sciences in which it first featured. (See Beaumont, 2010, Csillary *et al.*, 2010; Marin *et al.*, 2011, Sisson and Fan, 2011 and Robert, 2015, for reviews.) The technique circumvents direct evaluation of the likelihood function by selecting parameter draws that yield pseudo data - as simulated from the assumed model - that matches the observed data, with the matching based on summary statistics. If such statistics are sufficient (and if an arbitrarily small tolerance is used in the matching) the selected draws can be used to produce a posterior distribution that is exact up to simulation error; otherwise, an estimate of the *partial* posterior, where the latter reflects the information content of the set of summary statistics, is the only possible outcome.

The choice of statistics for use within ABC, in addition to techniques for determining the matching criterion, are clearly of paramount importance, with much recent research having been devoted to devising ways of ensuring that the information content of the chosen set of statistics is maximized, in some sense; e.g. Joyce and Marjoram (2008), Wegmann *et al.* (2009), Blum (2010), Fearnhead and Prangle (2012) and Frazier *et al.* (2015). In this vein, Drovandi *et al.* (2011), Gleim and Pigorsch (2013), Creel and Kristensen (2015), Creel *et al.,* (2015) and Drovandi *et al.* (2015), produce statistics via an *auxiliary* model selected to approximate the features of the true data generating process. This approach mimics, in a Bayesian framework, the principle underlying the frequentist method of indirect inference (II) (Gouriéroux *et al.* 1993, Smith, 1993) using, as it does, the approximating model to produce feasible inference about an intractable true model. Whilst the price paid for the approximation in the frequentist setting is a possible reduction in efficiency, the price paid in the Bayesian case is posterior inference that is conditioned on statistics that are not sufficient for the parameters of the true model, and which amounts to only partial inference as a consequence.

Our paper continues in this spirit, but with focus given to the application of auxiliary model-based ABC methods in the state space model (SSM) framework. Whilst ABC methods have been proposed in this setting, *inter alia*, Jasra *et al.,* 2010, Dean *et al.*, 2014, Martin *et al.*, 2014, Calvet and Czellar, 2015a, 2015b, Yildirim *et al.*, 2015), such methods use ABC principles (without summarization) to estimate either the likelihood function or the smoothed density of the states, with established techniques (e.g. maximum likelihood or (particle) Markov chain Monte Carlo) then being used to conduct inference on the static parameters themselves. (Jasra, 2015, provides an extensive review of this literature, including existing theoretical results, as well as providing comprehensive computational

insights.)

Our aim, in contrast, is to explore the use of ABC alone and as based on some form of summarization, in conducting inference on the static parameters in SSMs. We begin by demonstrating that reduction to a set of sufficient statistics of fixed dimension relative to the sample size is *infeasible* in such models. That is, one is precluded from the outset from using ABC (based on summary statistics) to conduct exact finite sample inference in SSMs; *only* partial inference is feasible via this route. Given the difficulty of characterizing the nature of posterior inference that conditions on non-sufficient statistics, we motivate the use of ABC here by means of a different criterion. To wit, we give conditions under which ABC methods are *Bayesian consistent* in the state space setting, in the sense of producing draws that yield a degenerate distribution at the true vector of static parameters in the (sample size) limit. To do this we adopt the auxiliary likelihood approach to produce the summaries. This is entirely natural when considering the canonical case where continuous time SSMs (for which the likelihood function is typically unavailable) are approximated by discretized versions, and auxiliary likelihoods subsequently constructed. Use of maximum likelihood to estimate the auxiliary parameters also allows asymptotic sufficiency to be invoked, thereby ensuring that - for large samples at least - maximum information is extracted from the auxiliary likelihood in producing the summaries.

We give particular emphasis to two non-linear stochastic volatility models in which the state is driven by a continuous time diffusion. Satisfaction of the full set of conditions for Bayesian consistency is shown to hold for the model driven by a (latent) Ornstein-Ulenbeck process when a discrete time linear Gaussian approximation is adopted and the auxiliary likelihood function is evaluated by the Kalman filter (KF). For the second example, in which a square root volatility process is assumed - and the auxiliary likelihood associated with the Euler discretization is evaluated via the augmented unscented Kalman filter (AUKF) (Julier *et al.*, 1995, 2000) - all conditions other than an identification condition can be theoretically verified.

We also illustrate that to the order of accuracy that is relevant in establishing the theoretical properties of an ABC technique, a selection criterion based on the score of the auxiliary likelihood - evaluated at the maximum likelihood estimator (MLE) computed from the observed data - yields equivalent results to a criterion based directly on the MLE itself. This equivalence is shown to hold in both the exactly and over-identified cases, and independently of any (positive definite) weighting matrix used to define the two alternative distance measures, and implies that the proximity to asymptotic sufficiency yielded by using the auxiliary MLE in an ABC algorithm will be replicated by the use of the auxiliary score. Given the enormous gain in speed achieved by avoiding optimization of the auxiliary likelihood at each replication of ABC, this is a critical result from a computational perspective.

Finally, we briefly address the issue of dimensionality that plagues ABC techniques in multiple parameter settings. (See Blum, 2010, and Nott *et al.*, 2014). Specifically, we demonstrate numerically the improved accuracy that can be achieved by matching individual parameters via the corresponding scalar score of the integrated auxiliary likelihood, as an alternative to matching on the multi-dimensional score statistic as suggested, for example, in Drovandi *et al.* (2015).

The paper proceeds as follows. In Section 2 we briefly summarize the basic principles of ABC as they would apply in a state space setting, including the role played by summary statistics and sufficiency. We demonstrate the lack of finite sample sufficiency reduction in a SSM, using the linear Gaussian model for illustration. In Section 3, we then proceed to demonstrate the theoretical properties of the auxiliary likelihood approach to ABC, including the conditions under which Bayesian consistency holds, in very general settings. The sense in which inference based on the auxiliary MLE is replicated by inference based on the auxiliary score is also detailed. In Section 4 we then consider the auxiliary likelihood approach explicitly in the non-linear state space setting, using the two continuous time latent volatility models for illustration. Numerical accuracy of the proposed method - as applied to data generated artificially from the square root volatility model - is then assessed in Section 5. Existence of known (non-central chi-squared) transition densities means that the exact likelihood function/posterior distribution is available for the purpose of comparison. The accuracy of the auxiliary likelihood-based ABC posterior estimate is compared with: 1) an ABC estimate that uses a (weighted) Euclidean metric based on statistics that are sufficient for an observed autoregressive model of order one; and 2) an ABC estimate that exploits the dimension-reduction technique of Fearnhead and Prangle (2012), applied to this latter set of summary statistics. The auxiliary likelihood-based method is shown to provide the most accurate estimate of the exact posterior in almost all cases documented. Critically, numerical evidence is produced that Bayesian consistency holds when the auxiliary score is used to generate the matching statistics, in contrast to the mixed evidence for the alternative ABC methods. Section 6 concludes. Technical proofs are included in an appendix to the paper.

2 Auxiliary likelihood-based ABC in state space models

2.1 Outline of the basic approach

The aim of ABC is to produce draws from an approximation to the posterior distribution of a vector of unknowns, $\boldsymbol{\theta}$, given the T-dimensional vector of observed data

$\mathbf{y} = (y_1, y_2, ..., y_T)'$,

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

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

Algorithm 1 ABC algorithm

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

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}^i)\} \leq \varepsilon, \tag{1}$$

where $\boldsymbol{\eta}(.)$ is a (vector) statistic, $d\{.\}$ is a distance criterion, and, given N, the tolerance level ε is chosen as small as the computing budget allows

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

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

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

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

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

Adaptations of the basic rejection scheme have involved post-sampling corrections of the draws using kernel methods (Beaumont *et al.*, 2002, Blum 2010, Blum and François, 2010), or the insertion of Markov chain Monte Carlo (MCMC) and/or sequential Monte Carlo (SMC) steps (Marjoram *et al.*, 2003, Sisson *et al.,* 2007, Beaumont *et al.*, 2009, Toni *et al.*, 2009, and Wegmann *et al.*, 2009), to improve the accuracy with which $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ is estimated, for any given number of draws. Focus is also given to choosing $\boldsymbol{\eta}(.)$ and/or $d\{.\}$ so as to render $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ a closer match to $p(\boldsymbol{\theta}|\mathbf{y})$, in some sense; see Joyce and Marjoram (2008), Wegmann *et al.*, Blum (2010) and Fearnhead and Prangle (2012). In the latter vein, Drovandi *et al.* (2011) argue, in the context of a specific biological model, that the use of $\boldsymbol{\eta}(.)$ comprised of the MLEs of the parameters of a well-chosen approximating model,

may yield posterior inference that is conditioned on a large portion of the information in the data and, hence, be close to exact inference based on $p(\boldsymbol{\theta}|\mathbf{y})$. (See also Gleim and Pigorsch, 2013, Creel and Kristensen, 2015, Creel *et al.,* 2015, and Drovandi *et al.*, 2015, for related work.) It is the spirit of this approach that informs the current paper, but with our attention given to rendering the approach feasible in a general state space framework that encompasses a large number of the models that are of interest to practitioners, including continuous time models.

Our focus then is on the application of ABC in the context of a general SSM with measurement and transition distributions,

$$p(y_t|x_t, \boldsymbol{\phi}) \tag{3}$$

$$p(x_t|x_{t-1}, \boldsymbol{\phi}) \tag{4}$$

respectively, where $\boldsymbol{\phi}$ is a p-dimensional vector of static parameters, elements of which may characterize either the measurement or state relation, or both. For expositional simplicity, and without loss of generality, we consider the case where both y_t and x_t are scalars. In financial applications it is common that both the observed and latent processes are driven by continuous time processes, with the transition distribution in (4) being unknown (or, at least, computationally challenging) as a consequence. Bayesian inference would then typically proceed by invoking (Euler) discretizations for both the measurement and state processes and applying MCMC- or SMC-based techniques (potentially with some ABC principles embedded within, as highlighted in the Introduction), with such methods being tailor-made to suit the features of the particular (discretized) model at hand.

The aim of the current paper is to use ABC principles to conduct inference about (3) and (4) using an approximation to the (assumed intractable) likelihood function. The full set of unknowns constitutes the augmented vector $\boldsymbol{\theta} = (\boldsymbol{\phi}', \mathbf{x}_c')'$ where, in the case when x_t evolves in continuous time, $\mathbf{x}_c$ represents the infinite-dimensional vector comprising the continuum of unobserved states over the sample period. However, to fix ideas, we define $\boldsymbol{\theta} = (\boldsymbol{\phi}', \mathbf{x}')'$, where $\mathbf{x} = (x_1, x_2, ..., x_T)'$ is the T-dimensional vector comprising the time t states for the T observation periods in the sample.[1] Implementation of the algorithm thus involves simulating from $p(\boldsymbol{\theta})$ by simulating $\boldsymbol{\phi}$ from the prior $p(\boldsymbol{\phi})$, followed by simulation of x_t via the process for the state, conditional on the draw of $\boldsymbol{\phi}$, and subsequent simulation of artificial data z_t conditional on the draws of $\boldsymbol{\phi}$ and the state variable. Crucially, our attention is given to inference about $\boldsymbol{\phi}$ only; hence, only draws of $\boldsymbol{\phi}$ are retained (via the selection criterion) and those draws used to produce an estimate of the marginal posterior, $p(\boldsymbol{\phi}|\mathbf{y})$, and with sufficiency (or, more pertinently, lack thereof) to be viewed as relating to $\boldsymbol{\phi}$ only. Hence, from this point onwards, when we reference a vector of summary statistics,

[1]For example, in a continuous time stochastic volatility model such values may be interpreted as end-of-day volatilities.

$\boldsymbol{\eta}(\mathbf{y})$, it is the information content of that vector with respect to $\boldsymbol{\phi}$ that is of importance, and the asymptotic behaviour of $p_\varepsilon(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$ with reference to the true $\boldsymbol{\phi}_0$ that is under question. Similarly, in the numerical illustration in Section 5, it is the proximity of the particular (kernel-based estimate of) $p_\varepsilon(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$ explored therein to the exact $p(\boldsymbol{\phi}|\mathbf{y})$ that is documented. We comment briefly on state inference in Section 6.

Before outlining the proposed methodology for the model in (3) and (4) in Section 3, we highlight a key observation that provides some motivation for our particular approach, namely that reduction to sufficiency in finite samples is not possible in state space settings. We use a linear Gaussian state space model to illustrate this result, as closed-form expressions are available in this case; however, as highlighted at the end of the section, the result is, in principle, applicable to any SSM.

2.2 Lack of finite sample sufficiency reduction

When the cardinality of the set of sufficient statistics is small relative to the sample size a significant reduction in complexity is achieved and in the case of ABC, conditioning on the sufficient statistics leads to no loss of information, and the method produces a simulation-based estimate of the exact posterior. The difficulty that arises is that only distributions that are members of the exponential family (EF) possess sufficient statistics that achieve a reduction to a fixed dimension relative to the sample size. In the context of the general SSM described by (3) and (4) the effective use of sufficient statistics is problematic. For any t it is unlikely that the marginal distribution of y_t will be a member of the EF, due to the vast array of non-linearities that are possible, in either the measurement or state equations, or both. Moreover, even if y_t *were* a member of the EF for each t, to achieve a sufficiency reduction it is required that the *joint* distribution of $\mathbf{y} = \{y_t; t = 1, 2, ..., T\}$ also be in the EF. For example, even if y_t were Gaussian, it does not necessarily follow that the joint distribution of $\mathbf{y}$ will achieve a sufficiency reduction. The most familiar example of this is when $\mathbf{y}$ follows a Gaussian moving average (MA) process and consequently only the whole sample is sufficient.

Even the simplest SSMs generate MA-like dependence in the data. Consider the linear Gaussian SSM, expressed in regression form as

$$y_t = x_t + e_t \tag{5}$$

$$x_t = \delta + \rho x_{t-1} + v_t, \tag{6}$$

where the disturbances are respectively independent $N(0, \sigma_e^2 = 1)$ and $N(0, \sigma_v^2)$ variables. In this case, the joint distribution of the vector of y_t's (which are marginally normal and members of the EF) is $\mathbf{y} \sim \mathbf{N}(\mu\boldsymbol{\iota}, \sigma_x^2 (r_{SN}\mathbf{I} + \mathbf{V}))$, where $r_{SN} = \sigma_e^2/\sigma_x^2$ is the inverse of the signal-to-noise (SN) ratio, $\mu = \delta/(1-\rho)$, $\boldsymbol{\iota}$ is the T-dimensional vector of 1's and $\mathbf{V}$ is the

familiar Toeplitz matrix associated with an autoregressive (AR) model of order 1. The matrix $\mathbf{V}^{-1}$ has a tri-diagonal form, illustrated here without loss of generality for the case of $T = 5$:

$$\mathbf{V}^{-1} = \begin{bmatrix} 1 & -\rho & 0 & 0 & 0 \\ -\rho & \rho^2+1 & -\rho & 0 & 0 \\ 0 & -\rho & \rho^2+1 & -\rho & 0 \\ 0 & 0 & -\rho & \rho^2+1 & -\rho \\ 0 & 0 & 0 & -\rho & 1 \end{bmatrix}.$$

In general $\mathbf{V}^{-k}$ is $(2k+1)$-diagonal.

To construct the sufficient statistics we need to evaluate $(r_{SN}\mathbf{I} + \mathbf{V})^{-1}$, which appears in the quadratic form of the multivariate normal density, with the structure of $(r_{SN}\mathbf{I} + \mathbf{V})^{-1}$ determining the way in which sample information about the parameters is accumulated and, hence, the sufficiency reduction that is achievable. To illustrate this, we write

$$(r_{SN}\mathbf{I} + \mathbf{V})^{-1} = \mathbf{V}^{-1} - r_{SN}\mathbf{V}^{-2} + r_{SN}^2\mathbf{V}^{-3} - ..., \tag{7}$$

with the expression reducing to the result for an *observed* AR(1) process when $r_{SN} = 0$. In this case the sufficient statistics are thus calculated from the quadratic form of the normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}^{-1}$. Using the conventional row/column matrix notation to express a general quadratic form $\mathbf{y}'\mathbf{A}\mathbf{y}$ as

$$\mathbf{y}'\mathbf{A}\mathbf{y} = \sum_{t=1}^{T} a_{t,t}y_t^2 + 2\sum_{s>t}\sum_{t=1}^{T} a_{s,t}y_s y_t, \tag{8}$$

it is instructive to re-express the right-hand-side of (8) as

$$\sum_{t=1}^{T} a_{t,t}y_t^2 + 2\sum_{t>1}^{T} a_{t,t-1}y_t y_{t-1} + 2\sum_{t>2}^{T} a_{t,t-2}y_t y_{t-2} + ... + 2\sum_{t>T-1}^{T} a_{t,t-T+1}y_t y_{t-T+1} \tag{9}$$

noting that the last term in the expansion is equivalent to $2a_{T,1}y_T y_1$. When $\mathbf{A} = \mathbf{V}^{-1}$, $a_{t,t-k} = 0$ for $t > k > 2$ and only the first 2 terms in (9) are present as a consequence. Since $\mathbf{V}^{-1}$ has constant terms along the diagonals except for end effects (i.e. the different first and last rows), the sufficient statistics thus comprise

$$s_1 = \sum_{t=2}^{T-1} y_t,\ s_2 = \sum_{t=2}^{T-1} y_t^2,\ s_3 = \sum_{t=2}^{T} y_t y_{t-1},\ s_4 = y_1 + y_T,\ s_5 = y_1^2 + y_T^2. \tag{10}$$

For $r_{SN} \neq 0$ however, as characterizes a state space model with measurement error, the extra terms in the expansion in (7) come into play. Taking the first-order approximation, for example, $\mathbf{A} = \mathbf{V}^{-1} - r_{SN}\mathbf{V}^{-2}$ is 5-diagonal, $a_{t,t-k} = 0$ for $t > k > 3$, and the first 3 terms in (9) remain. As a consequence, the dimension of the set of sufficient statistics increases

by 1. As the SN ratio declines, and r_{SN} increases as a result, higher-order terms are required to render the approximation in (7) accurate, and the dimension of the sufficient set increases correspondingly. Thus, for any $r_{SN} \neq 0$, the structure of $(r_{SN}\mathbf{I} + \mathbf{V})^{-1}$ is such that information in the sample of size T does not accumulate, and reduction to a sufficient set of statistics of dimension smaller than T is not feasible.

This same qualitative problem would also characterize any SSM nested in (3) and (4), with the only difference being that, in any particular case there would not necessarily be an analytical link between the SN ratio and the lack of sufficiency associated with any finite set of statistics calculated from the observations. The quest for an accurate ABC technique in a state space setting as based on an arbitrary set of statistics is thus not well-founded and this, in turn, motivates the search for summary measures via the application of MLE to an auxiliary likelihood. As well as enabling the extraction of the maximum information from the auxiliary likelihood via the asymptotic sufficiency of the auxiliary MLE, standard regularity conditions on the approximating likelihood function are able to be invoked for the purpose of establishing Bayesian consistency of the ABC posterior.

3 Auxiliary likelihood-based ABC

3.1 'Approximate' asymptotic sufficiency

Asymptotic Gaussianity of the MLE for the parameters of (3) and (4) (under regularity) implies that the MLE satisfies the factorization theorem and is thereby asymptotically sufficient for the parameters of that model. (See Cox and Hinkley, 1974, Chp. 9 for elucidation of this matter.) Denoting the log-likelihood function by $L(\mathbf{y}; \boldsymbol{\phi})$, maximizing $L(\mathbf{y}; \boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$ yields $\widehat{\boldsymbol{\phi}}$, which could, in principle, be used to define $\boldsymbol{\eta}(.)$ in an ABC algorithm. For large enough T (and for as $\varepsilon \to 0$) the algorithm would thus produce draws from the exact posterior. Indeed, in arguments that mirror those adopted by Gallant and Tauchen (1996) and Gouriéroux *et al.* (1993) for the efficient method of moments (EMM) and II estimators respectively, Gleim and Pigorsch (2013) demonstrate that if $\boldsymbol{\eta}(.)$ is chosen to be the MLE of an auxiliary model that 'nests' the true model in some well-defined way, asymptotic sufficiency will still be achieved; see also Gouriéroux and Monfort (1995) on this point.

Of course, if the SSM in question is such that the exact likelihood is accessible, the model is likely to be tractable enough to preclude the need for treatment via ABC. Further, the quest for asymptotic sufficiency via a *nesting* auxiliary model conflicts with the quest for an accurate non-parametric estimate of the posterior using the ABC draws, given that the dimension of the parameter set in the auxiliary model is, by construction, likely to be large. Hence, in practice, the appropriate goal in using the auxiliary likelihood approach

to ABC in the SSM context is to define a sensible *parsimonious* approximation to the true model in (3) and (4), for which the associated likelihood function can be evaluated with computational ease and speed. Heuristically, if the approximating model is 'accurate enough' as a representation of the true model, such an approach will yield, via the ABC algorithm, an estimate of the posterior distribution that is conditioned on a statistic that is 'close to' being asymptotically sufficient for $\boldsymbol{\phi}$. We certainly make no attempt in this paper to formalize this statement in any way. Nevertheless, we do view the notion of asymptotic sufficiency of the auxiliary MLE as being a intuitively compelling characteristic of the auxiliary likelihood-based approach to ABC, and the numerical results presented later provide some support for its importance in practice. More critically, however, pursuing the auxiliary likelihood route enables us to draw on regularity as it pertains to likelihood functions, and maximization thereof, to prove the (Bayesian) consistency of the resultant ABC posterior and, hence, the baseline accuracy of the inferences produced via this route.

3.2 Bayesian consistency and ABC

In the ABC setting, Bayesian consistency essentially requires that as $T \to \infty$ and $\varepsilon \to 0$, the estimated posterior based on the selected draws from $p_\varepsilon(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$ concentrates around the true parameter value generating the data; see, for example, Frazier *et al.* (2015) and the references therein. With a slight abuse of terminology, from this point onwards we denote the 'ABC posterior' by $p_\varepsilon(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$, recognizing that the quantity produced via ABC is actually the kernel-based density estimate constructed from a given number of draws, N, from $p_\varepsilon(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$ as defined in (2).

In what follows, we use the following notation throughout the remainder of the paper. For a d-dimensional vector $\mathbf{X}$ we denote the Euclidean norm of $\mathbf{X}$ as $\|\mathbf{X}\|$, and we let $E[\mathbf{X}]$ denote the expectation of $\mathbf{X}$ under the true probability distribution. Let "$\xrightarrow{P}$" denote convergence in probability, where P denotes a generic probability measure, and $O_P(a_n)$, $o_P(b_n)$ and plim have the standard connotations. Denote by $\mathbf{z}(\boldsymbol{\phi}^i)$ the ith vector of pseudo data, where the dependence of $\mathbf{z}(\boldsymbol{\phi}^i)$ on the ith random draw $\boldsymbol{\phi}^i$ from the prior $p(\boldsymbol{\phi})$ is made explicit.

For a given auxiliary model, with parameters $\boldsymbol{\beta} \in \mathbf{B} \subset \mathbb{R}^d$, $d \geq p$, and sample log-likelihood function $L_a(\mathbf{y};\boldsymbol{\beta})$, ABC can use as summary statistics for inference on $\boldsymbol{\phi}$ the maximizers of $L_a(\cdot;\boldsymbol{\beta})$, based on $\mathbf{y}$ and $\mathbf{z}(\boldsymbol{\phi}^i)$, which we represent respectively by

$$\widehat{\boldsymbol{\beta}}(\mathbf{y}) = \arg\max_{\boldsymbol{\beta}\in\mathbf{B}} L_a(\mathbf{y};\boldsymbol{\beta}) \text{ and } \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i)) = \arg\max_{\boldsymbol{\beta}\in\mathbf{B}} L_a(\mathbf{z}(\boldsymbol{\phi}^i);\boldsymbol{\beta}).$$

Using $\widehat{\boldsymbol{\beta}}(\mathbf{y})$ and $\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))$ as summary statistics, we can take as the distance criterion in (1),

$$d\{\boldsymbol{\eta}(\mathbf{y}),\boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\phi}^i))\} = \sqrt{\left[\widehat{\boldsymbol{\beta}}(\mathbf{y})-\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))\right]' \boldsymbol{\Omega} \left[\widehat{\boldsymbol{\beta}}(\mathbf{y})-\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))\right]}, \tag{11}$$

where $\boldsymbol{\Omega}$ is some positive definite matrix.

The intuition behind Bayesian consistency of ABC based on $\boldsymbol{\eta}(\mathbf{y}) = \widehat{\boldsymbol{\beta}}(\mathbf{y})$ follows from the following sequence of arguments. Firstly, under mild regularity conditions, as $T \to \infty$ the criterion in (11) should satisfy

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\phi}^i))\} \xrightarrow{P} \sqrt{\left[\boldsymbol{\beta}_0 - \mathbf{b}(\boldsymbol{\phi}^i)\right]' \boldsymbol{\Omega} \left[\boldsymbol{\beta}_0 - \mathbf{b}(\boldsymbol{\phi}^i)\right]}, \tag{12}$$

where

$$\boldsymbol{\beta}_0 = \arg\max_{\boldsymbol{\beta} \in \mathbf{B}} \left\{ \operatorname*{plim}_{T\to\infty} (1/T) L_a(\mathbf{y}; \boldsymbol{\beta}) \right\}$$

and, for any $\boldsymbol{\phi}^i \in \mathbf{B}$,

$$\mathbf{b}(\boldsymbol{\phi}^i) = \arg\max_{\boldsymbol{\beta} \in \mathbf{B}} \left\{ \operatorname*{plim}_{T\to\infty} (1/T) L_a(\mathbf{z}(\boldsymbol{\phi}^i); \boldsymbol{\beta}) \right\}.$$

Secondly, under identification conditions, $\boldsymbol{\phi}^i = \boldsymbol{\phi}_0$ is the only value that satisfies $\boldsymbol{\beta}_0 = \mathbf{b}(\boldsymbol{\phi}^i)$ (where $\boldsymbol{\phi}_0$ is the parameter generating the observed data) and, as a consequence, the only value that satisfies

$$d\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi}^i)\} = \sqrt{\left[\boldsymbol{\beta}_0 - \mathbf{b}(\boldsymbol{\phi}^i)\right]' \boldsymbol{\Omega} \left[\boldsymbol{\beta}_0 - \mathbf{b}(\boldsymbol{\phi}^i)\right]} = 0. \tag{13}$$

Hence, as $T \to \infty$, the only value of $\boldsymbol{\phi}^i$ satisfying $d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\phi}^i))\} \leq \varepsilon$ for any $\varepsilon \geq 0$ is $\boldsymbol{\phi}^i = \boldsymbol{\phi}_0$, and so for well-behaved $\widehat{\boldsymbol{\beta}}(\mathbf{y})$, as $T \to \infty, \varepsilon \to 0$ the ABC algorithm will only select draws arbitrarily close to $\boldsymbol{\phi}_0$.

Put formally, the ABC posterior, based on $\boldsymbol{\eta}(\mathbf{y}) = \widehat{\boldsymbol{\beta}}(\mathbf{y})$, will be Bayesian consistent if, for $\Psi_\delta(\boldsymbol{\phi}_0) := \{\boldsymbol{\phi} \in \boldsymbol{\Phi} : \|\boldsymbol{\phi} - \boldsymbol{\phi}_0\| > \delta\}$ and any $\delta > 0$, $\int_{\Psi_\delta(\boldsymbol{\phi}_0)} p_\varepsilon(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))d\boldsymbol{\phi} = o_P(1)$, as $T \to \infty$ and $\varepsilon \to 0$. To formally establish Bayesian consistency for $\boldsymbol{\eta}(\mathbf{y}) = \widehat{\boldsymbol{\beta}}(\mathbf{y})$ and the distance in (11), we require the following assumptions.

Assumption A:

(A1) The parameter spaces $\mathbf{B} \subset \mathbb{R}^d$ and $\boldsymbol{\Phi} \subset \mathbb{R}^p$ are compact.

(A2) For any $\boldsymbol{\phi}^i \in \boldsymbol{\Phi}$, $\{z_t(\boldsymbol{\phi}^i)\}_{t=1}^T$ is stationary and ergodic.

(A3) The map $(\boldsymbol{\phi}^i, \boldsymbol{\beta}) \mapsto L_a(\mathbf{z}(\boldsymbol{\phi}^i); \boldsymbol{\beta})$ is continuous in $(\boldsymbol{\phi}^i, \boldsymbol{\beta})$.[2]

(A4) For each $\boldsymbol{\beta} \in \mathbf{B}$, and $\boldsymbol{\phi}^i \in \boldsymbol{\Phi}$, $E[|\ell(z_t(\boldsymbol{\phi}^i); \boldsymbol{\beta})|] < \infty$, where $\ell(z_t(\boldsymbol{\phi}^i); \boldsymbol{\beta})$ is the t-th contribution to $L_a(\mathbf{z}(\boldsymbol{\phi}^i); \boldsymbol{\beta})$.

[2] In certain cases the sample paths of the simulated $\mathbf{z}(\boldsymbol{\phi}^i)$, may not be continuous in $\boldsymbol{\phi}^i$, invalidating Assumption **(A3)**. In these situations an alternative assumption, known as first-moment continuity (see, for example, Hansen, 2012), can often be applied.

(A5) $L_\infty(\boldsymbol{\phi}^i;\boldsymbol{\beta}) := \text{plim}_{T\to\infty}(1/T)L_a(\mathbf{z}(\boldsymbol{\phi}^i);\boldsymbol{\beta})$ exists and has unique maximum $\mathbf{b}(\boldsymbol{\phi}^i) = \arg\max_{\boldsymbol{\beta}\in\mathbf{B}} L_\infty(\boldsymbol{\phi}^i;\boldsymbol{\beta})$, where $\boldsymbol{\beta}_0 = \mathbf{b}(\boldsymbol{\phi}_0) = \arg\max_{\boldsymbol{\beta}\in\mathbf{B}} L_\infty(\boldsymbol{\phi}_0;\boldsymbol{\beta})$.

Assumption I:

(I1) The prior $p(\boldsymbol{\phi})$ is continuous and $p(\boldsymbol{\phi}_0) > 0$.

(I2) The mapping $\boldsymbol{\phi} \mapsto \mathbf{b}(\boldsymbol{\phi})$ is one-to-one; that is, for every $\boldsymbol{\beta} \in \mathbf{B}$ the equation $\boldsymbol{\beta} = \mathbf{b}(\boldsymbol{\phi})$ has a unique solution in $\boldsymbol{\phi}$.

Remark 1: Under correct specification of the model generating the data $\mathbf{y}$, Assumptions **(A1)-(A5)** ensure that $\sup_{\boldsymbol{\beta}\in\mathbf{B}} |(1/T)L_a(\mathbf{y};\boldsymbol{\beta}) - L_\infty(\boldsymbol{\phi}_0;\boldsymbol{\beta})| = o_P(1)$, for $L_\infty(\boldsymbol{\phi}_0;\boldsymbol{\beta})$ defined in **(A5)**, and that $\|\widehat{\boldsymbol{\beta}}(\mathbf{y}) - \boldsymbol{\beta}_0\| = o_P(1)$. In addition, Assumptions **(A1)-(A5)** are enough to ensure that $\sup_{\boldsymbol{\phi}^i\in\boldsymbol{\Phi}} \|\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i)) - \mathbf{b}(\boldsymbol{\phi}^i)\| = o_P(1)$. The uniform convergence of $\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))$ to $\mathbf{b}(\boldsymbol{\phi}^i)$ is crucial as it ensures that the simulated paths $\mathbf{z}(\boldsymbol{\phi}^i)$, and the subsequent $\boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\phi}^i))$, are well-behaved over $\boldsymbol{\Phi}$.

Remark 2: The distance in (11) essentially mimics the Wald criterion used in the II technique.[3] Similar to II, in our Bayesian analyses, in which (11) is used to produce ABC draws, $\boldsymbol{\Omega}$ can also be defined as the sandwich form of a variance-covariance estimator (Gleim and Pigorsch, 2013, and Drovandi *et al.*, 2015), or as the inverse of the (estimated) variance-covariance matrix for $\boldsymbol{\beta}$, evaluated at $\widehat{\boldsymbol{\beta}}(\mathbf{y})$ (Drovandi *et al.,* 2011). In these cases it is more useful to denote the weighting matrix by $\widehat{\Omega}(\mathbf{y}, \widehat{\boldsymbol{\beta}}(\mathbf{y}))$ and Bayesian consistency then requires, for some positive definite $\Omega_\infty(\boldsymbol{\beta}_0)$, $\|\widehat{\Omega}(\mathbf{y}, \widehat{\boldsymbol{\beta}}(\mathbf{y})) - \Omega_\infty(\boldsymbol{\beta}_0)\|_* \xrightarrow{P} 0$, where, for $\boldsymbol{\Omega}$ an $n \times m$ matrix, $\|\boldsymbol{\Omega}\|_* = \sqrt{\text{Trace}(\boldsymbol{\Omega}'\boldsymbol{\Omega})}$.

Remark 3: Assumption **(I1)** ensures that the prior used within ABC places non-zero probability on the truth, and is standard in the analysis of Bayesian consistency. Assumption **(I2)** ensures that the limit distance in (13) is zero if and only if $\boldsymbol{\phi} = \boldsymbol{\phi}_0$.

The following theorem formally establishes Bayesian consistency of the ABC posterior $p_\varepsilon(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$. The proof is contained in Appendix A.1.

Theorem 1 *For all $\delta > 0$, if Assumptions **(A)** and **(I)** are satisfied, then for $\varepsilon = o(1)$ and $\varepsilon^{-1}\sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \|\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i)) - \mathbf{b}(\boldsymbol{\phi}^i)\| = o_P(1)$,*

$$\int_{\Psi_\delta(\boldsymbol{\phi}_0)} p_\varepsilon(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))d\boldsymbol{\phi} \xrightarrow{P} 0, \text{ for } \boldsymbol{\eta}(\mathbf{y}) = \widehat{\boldsymbol{\beta}}(\mathbf{y}), \text{ as } T \to \infty,$$

where $\Psi_\delta(\boldsymbol{\phi}_0) := \{\boldsymbol{\phi} \in \boldsymbol{\Phi} : \|\boldsymbol{\phi} - \boldsymbol{\phi}_0\| > \delta\}$.

[3] In practice the implementation of II may involve the use of a simulated sample in the computation of $\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))$ that is a multiple of the size of the empirical sample.

Remark 4: Our focus on Bayesian consistency only requires a weak set of sufficient conditions on the auxiliary likelihood that can be readily verified in many cases, as illustrated throughout the remainder. In particular, the tenants of Theorem 1 are weaker than those considered in the ABC literature where asymptotic behaviour of ABC point estimates is the focus; see, for example, Li and Fearnhead (2015) and Creel *et al.* (2015). For instance, nothing about our conditions requires the summaries to satisfy a central limit theorem, which allows for highly dependent data. We refer the reader to Frazier *et al.* (2015) for a treatment of ABC consistency as it relates to conditioning on general summary statistics, and in completely general model frameworks.

3.3 Auxiliary score-based ABC

With large computational gains, $\boldsymbol{\eta}(.)$ in (1) can be defined using the score of the auxiliary likelihood. That is, the score vector associated with the approximating likelihood function, when evaluated at the simulated data, and with $\widehat{\boldsymbol{\beta}}(\mathbf{y})$ substituted for $\boldsymbol{\beta}$, will be closer to zero the 'closer' is the simulated data to the true. Hence, the distance criterion in (1) for an ABC algorithm can be based on $\boldsymbol{\eta}(.) = \mathbf{S}(.;\widehat{\boldsymbol{\beta}}(\mathbf{y}))$, where

$$\mathbf{S}(\mathbf{z}(\boldsymbol{\phi}^i);\boldsymbol{\beta}) = T^{-1}\frac{\partial L_a(\mathbf{z}(\boldsymbol{\phi}^i);\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}. \tag{14}$$

This yields the selection rule

$$d\{\boldsymbol{\eta}(\mathbf{y}),\boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\phi}^i))\} = \sqrt{\left[\mathbf{S}(\mathbf{z}(\boldsymbol{\phi}^i);\widehat{\boldsymbol{\beta}}(\mathbf{y}))\right]' \boldsymbol{\Sigma} \left[\mathbf{S}(\mathbf{z}(\boldsymbol{\phi}^i);\widehat{\boldsymbol{\beta}}(\mathbf{y}))\right]} \leq \varepsilon, \tag{15}$$

where $\mathbf{S}(\mathbf{y};\widehat{\boldsymbol{\beta}}(\mathbf{y})) = 0$ is invoked, and $\boldsymbol{\Sigma}$ denotes an arbitrary positive definite weighting matrix. Implementation of ABC via (15) is faster (by orders of magnitude) than the approach based upon $\boldsymbol{\eta}(.) = \widehat{\boldsymbol{\beta}}(.)$, due to the fact that maximization of the auxiliary likelihood is required only once, in order to produce $\widehat{\boldsymbol{\beta}}(.)$ from the observed data $\mathbf{y}$. All other calculations involve simply the evaluation of $\mathbf{S}(.;\widehat{\boldsymbol{\beta}}(\mathbf{y}))$ at the simulated data, with a numerical differentiation technique invoked to specify $\mathbf{S}(.;\widehat{\boldsymbol{\beta}}(\mathbf{y}))$, when not known in closed-form.

Once again in line with the proof of the consistency of the relevant frequentist (EMM) estimator, the Bayesian consistency result in Section 3.2 could be re-written in terms $\boldsymbol{\eta}(.) = \mathbf{S}(.;\widehat{\boldsymbol{\beta}}(\mathbf{y}))$, upon the addition of a differentiability condition regarding $L_a(\mathbf{z}(\boldsymbol{\phi}^i);\boldsymbol{\beta})$ and the assumption that $\boldsymbol{\beta} = \mathbf{b}(\boldsymbol{\phi})$ is the unique solution to the limiting first-order condition, $\partial L_\infty(\boldsymbol{\phi},\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \text{plim}_{T\to\infty}\, T^{-1}\partial L_a(\mathbf{z}(\boldsymbol{\phi});\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, with the convergence uniform in $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$. In brief, given that $\widehat{\boldsymbol{\beta}}(\mathbf{y}) \xrightarrow{P} \mathbf{b}(\boldsymbol{\phi}_0)$, as $T \to \infty$ the limiting value of the choice criterion in (15) is

$$d\{\mathbf{b}(\boldsymbol{\phi}_0),\mathbf{b}(\boldsymbol{\phi}^i)\} = \sqrt{\left[\partial L_\infty(\boldsymbol{\phi}^i;\mathbf{b}(\boldsymbol{\phi}_0))/\partial\boldsymbol{\beta}\right]' \boldsymbol{\Sigma} \left[\partial L_\infty(\boldsymbol{\phi}^i;\mathbf{b}(\boldsymbol{\phi}_0))/\partial\boldsymbol{\beta}\right]}.$$

Thus, once again, as $T \to \infty$, and irrespective of the form of $\mathbf{\Sigma}$, the only value of $\boldsymbol{\phi}^i$ satisfying (15) for any $\varepsilon \geq 0$ is $\boldsymbol{\phi}^i = \boldsymbol{\phi}_0$, and the ABC posterior estimate will concentrate around $\boldsymbol{\phi}_0$ accordingly.

Hence, under regularity conditions, Bayesian consistency will be maintained when using the score of an auxiliary likelihood. However, a remaining pertinent question concerns the impact on sufficiency (or, more precisely, on *the proximity to asymptotic sufficiency*) associated with using the score instead of the auxiliary MLE itself. In practical terms this question can be re-phrased as: do the selection criteria based on $\mathbf{S}(.;\widehat{\boldsymbol{\beta}}(\mathbf{y}))$ and $\widehat{\boldsymbol{\beta}}(\mathbf{y})$ yield identical draws of $\boldsymbol{\phi}$? If the answer is yes, then, unambiguously, for large enough T and for $\varepsilon \to 0$, the score- and MLE-based ABC criteria will yield equivalent estimates of the exact posterior $p(\boldsymbol{\phi}|\mathbf{y})$, with the accuracy of those (equivalent) estimates dependent, of course, on the nature of the auxiliary likelihood itself.

For any auxiliary likelihood (satisfying identification and regularity conditions) with unknown parameter vector $\boldsymbol{\beta}$, we expand the (scaled) score function in (14), evaluated at $\widehat{\boldsymbol{\beta}}(\mathbf{y})$, around the point $\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))$,

$$\mathbf{S}(\mathbf{z}(\boldsymbol{\phi}^i);\widehat{\boldsymbol{\beta}}(\mathbf{y})) = \mathbf{S}(\mathbf{z}(\boldsymbol{\phi}^i);\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))) + \mathbf{D}\left[\widehat{\boldsymbol{\beta}}(\mathbf{y}) - \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))\right] = \mathbf{D}\left[\widehat{\boldsymbol{\beta}}(\mathbf{y}) - \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))\right], \tag{16}$$

where

$$\mathbf{D} = T^{-1}\frac{\partial^2 L_a(\mathbf{z}(\boldsymbol{\phi}^i);\widetilde{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i)))}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} \tag{17}$$

and $\widetilde{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))$ denotes an (unknown, and coordinate-specific) intermediate value between $\widehat{\boldsymbol{\beta}}(\mathbf{y})$ and $\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))$. Hence, using (16), the criterion in (15) becomes

$$\begin{aligned} & \sqrt{\left[\mathbf{S}(\mathbf{z}(\boldsymbol{\phi}^i);\widehat{\boldsymbol{\beta}}(\mathbf{y}))\right]' \mathbf{\Sigma} \left[\mathbf{S}(\mathbf{z}(\boldsymbol{\phi}^i);\widehat{\boldsymbol{\beta}}(\mathbf{y}))\right]} \\ = & \sqrt{\left[\widehat{\boldsymbol{\beta}}(\mathbf{y}) - \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))\right]' \mathbf{D}'\mathbf{\Sigma}\mathbf{D} \left[\widehat{\boldsymbol{\beta}}(\mathbf{y}) - \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))\right]} \leq \varepsilon. \end{aligned} \tag{18}$$

Subject to standard conditions regarding the second derivatives of the auxiliary likelihood, the matrix $\mathbf{D}$ in (17) will be of full rank for $\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))$ close to $\widehat{\boldsymbol{\beta}}(\mathbf{y})$. As a consequence, and given the positive definiteness of $\mathbf{\Sigma}$, $\mathbf{D}'\mathbf{\Sigma}\mathbf{D}$ will be a positive definite matrix that is some function of $\boldsymbol{\phi}^i$. Hence, whilst for any $\varepsilon > 0$, the presence of $\mathbf{D}$ affects selection of $\boldsymbol{\phi}^i$, as $\varepsilon \to 0$, $\boldsymbol{\phi}^i$ will be selected via (18) if and only if $\widehat{\boldsymbol{\beta}}(\mathbf{y})$ and $\widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}^i))$ are equal. Similarly, irrespective of the form of the (positive definite) weighting matrix in (11), the MLE criterion will produce these same selections. This result obtains no matter what the dimension of $\boldsymbol{\beta}$ relative to $\boldsymbol{\phi}$, i.e. no matter whether the true parameters are exactly or over-identified by the parameters of the auxiliary likelihood. This result thus goes beyond the comparable result regarding the II/EMM estimators (see, for e.g. Gouriéroux and Monfort, 1996), in that the equivalence is independent of the form of weighting matrix used *and* the degree of over-identification that prevails.

Of course, in practice ABC is implemented with $\varepsilon > 0$, at which point the two ABC criteria will produce different draws. However, for the models entertained in this paper, preliminary investigation has assured us that the difference between the ABC estimates of the posteriors yielded by the alternative criteria is negligible for small enough ε. Hence, in the numerical section we operate solely with the score-based approach as the computationally feasible method of extracting both consistency and approximate asymptotic sufficiency in the state space setting.

4 Auxiliary likelihood-based ABC for two latent volatility models

In this section we give consideration to the application of auxiliary likelihood-based ABC in the specific context of the general non-linear non-Gaussian SSM defined in (3) and (4), with the primary goal being to illustrate the extent to which the conditions for Bayesian consistency, as presented in Theorem 1, are satisfied in practically relevant examples. Specifically, we explore the consistency of ABC as applied to two continuous time latent volatility models, using plausible approximating models and giving emphasis to computationally feasible evaluation of the auxiliary likelihoods.

4.1 Ornstein-Ulenbeck stochastic volatility model

Assume that the (logarithmic) return at time t, r_t, has mean zero with volatility governed by a continuous-time Ornstein-Ulenbeck (OU) process for the logarithm of its variance V_t,

$$r_t = V_t^{1/2}\epsilon_t, \tag{19}$$

$$d\ln(V_t) = \kappa(d - \ln(V_t))dt + \sigma dW_t, \tag{20}$$

where W_t is a Brownian motion. We observe a discrete sequence of returns from the structural model (19)-(20), and our goal is to conduct Bayesian inference on the parameters governing the dynamics of volatility.

Defining $x_t = \ln(V_t)$, on the interval $[t, t+\Delta]$, the law-of-motion for dx_t can be exactly discretized to yield

$$x_{t+\Delta} = d(1 - e^{-\Delta\kappa}) + x_t e^{-\Delta\kappa} + \sigma\sqrt{\frac{(1 - e^{-2\Delta\kappa})}{2\kappa}}\nu_t, \tag{21}$$

where $\nu_t \sim_{i.i.d} N(0,1)$ and, for some M, φ, we have $M \geq \kappa, d, \sigma \geq \varphi > 0$. Taking $\Delta = 1$, we may thus analyze the model in (19)-(20) by considering the discrete time version,

$$\ln(r_t^2) = \omega + x_t + \zeta_t \tag{22}$$

$$x_t = d(1 - e^{-\kappa}) + x_{t-1}e^{-\kappa} + \sigma\sqrt{\frac{(1 - e^{-2\kappa})}{2\kappa}}\nu_t \equiv \gamma + \rho x_{t-1} + \sigma_v\nu_t, \tag{23}$$

where $\gamma = d(1-\rho)$, $\rho = \exp(-\kappa)$, $\sigma_v = \sigma\sqrt{(1-\rho^2)/2\kappa}$, $\omega = -1.27$, and with

$$\zeta_t = \ln(\epsilon_t^2) - \omega \tag{24}$$

a mean-zero log-chi-squared random variable with variance $\sigma_\zeta^2 = \pi^2/2$. Before proceeding we write the model in a more compact, yet equivalent, form

$$y_t = \ln(r_t^2) - \gamma^* = x_t + \zeta_t, \tag{25}$$

$$x_t = \rho x_{t-1} + \sigma_v \nu_t, \tag{26}$$

where $\gamma^* = \omega + \gamma/(1-\rho)$. Given the exact discretization of the state in (20), (25)-(26) form an equivalent representation of the model in (19)-(20), and so, hereafter, we refer to the model in (25)-(26) as the true model under analysis (using the acronym: SV-OU), characterized by the unknown structural parameters d, κ and σ. Using $\overline{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$, an implied posterior for d can easily be constructed by appropriately scaling the observed and simulated data to have unit variance, and using the accepted draws for κ, σ, obtained from an ABC-based procedure. Therefore, we treat γ^* as known and concentrate on verification of the Assumptions in Theorem 1 for $\boldsymbol{\phi} = (\kappa, \sigma)'$.

To implement an auxiliary likelihood-based ABC algorithm, we adopt the following linear Gaussian approximation of (25)-(26),

$$y_t = x_t + e_t, \tag{27}$$

$$x_t = \beta_1 x_{t-1} + \beta_2 v_t, \tag{28}$$

with $e_t \sim N(0, \sigma_\zeta^2)$, $v_t \sim N(0,1)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. The approximation thus invokes the (incorrect) assumption that ζ_t in equation (24) is $\zeta_t \sim N(0, \sigma_\zeta^2)$, whilst retaining the correct specification for the conditional mean of y_t, namely x_t. Given that the unknown parameters appear only in the conditional mean, verifying the conditions for Bayesian consistency is simplified, with the parameters β_1 and β_2 in the approximating model being continuous and one-to-one transformations of the corresponding structural parameters, as follows,[4]

$$\beta_1 = e^{-\kappa};\ \beta_2 = \sigma\sqrt{\frac{1-\exp(-2\kappa)}{2\kappa}}.$$

The log-likelihood for the auxiliary model, $L_a(\mathbf{y}; \boldsymbol{\beta})$, can be readily constructed using the Kalman filter (hereafter, KF), and the auxiliary MLE, $\widehat{\boldsymbol{\beta}}(\mathbf{y})$, obtained by maximizing $L_a(\mathbf{y}; \boldsymbol{\beta})$. ABC could then be based on either the general distance in (11), or that in (18).

As the following corollary demonstrates, Bayesian consistency of ABC based on this particular auxiliary likelihood is achieved. The proof of this corollary is detailed in Appendix A.2.

[4]The continuity and one-to-one nature of the transformation defining the auxiliary parameters ensures compactness of the auxiliary parameter space $\mathbf{B}$.

Corollary 2 *For the SV-OU model in* (25)-(26) *and true value* $\boldsymbol{\phi}_0 = (\kappa_0, \sigma_0)'$*, the model in* (27)-(28)*, with auxiliary likelihood* $L_a(\mathbf{y}; \boldsymbol{\beta})$ *constructed via the KF, yields Bayesian consistent inference for* $\boldsymbol{\phi}_0$ *with* $\boldsymbol{\eta}(\mathbf{y}) = \widehat{\boldsymbol{\beta}}(\mathbf{y})$ *as* $T \to \infty$ *and* $\varepsilon \to 0$*.*

4.2 Square root stochastic volatility model

Assume, again, we observe mean-zero returns and now consider the square root model for V_t,

$$\begin{aligned} r_t &= V_t^{1/2}\epsilon_t, \\ dV_t &= (\delta - \alpha V_t)dt + \sigma_v\sqrt{V_t}dW_t, \end{aligned}$$

with W_t again a Brownian motion. We restrict the structural parameters as $2\delta \geq \sigma_v^2$ to ensure positive volatility, and for some M, φ, we impose $M \geq \sigma_v, \alpha, \delta \geq \varphi > 0$. With these restrictions, V_t is mean reverting and as $t \to \infty$, V_t approaches a steady state gamma distribution, with $E[V_t] = \delta/\alpha$ and $var(V_t) = \sigma_v^2\delta/2\alpha^2$. The transition density for V_t, conditional on V_{t-1}, is

$$p(V_t|V_{t-1}) = c\exp(-u-v)\left(\frac{v}{u}\right)^{q/2} I_q(2(uv)^{1/2}),$$

where $c = 2\alpha/\sigma_v^2(1-\exp(-\alpha))$, $u = cV_{t-1}\exp(-\alpha)$, $v = cV_t$, $q = \frac{2\delta}{\sigma_v^2} - 1$, and $I_q(.)$ is the modified Bessel function of the first kind of order q. The conditional distribution function is non-Central chi-square, $\chi^2(2cV_t; 2q+2, 2u)$, with $2q+2$ degrees of freedom and non-centrality parameter $2u$. In the same spirit as above (and maintaining comparable notation to adopted therein, but for $x_t := V_t$), we take squares and logarithms of the measurement equation, leading to the model

$$\begin{aligned} y_t &= \ln(r_t^2) = \ln(x_t) + \zeta_t \quad (29) \\ dx_t &= (\delta - \alpha x_t)dt + \sigma_v\sqrt{x_t}dW_t, \quad (30) \end{aligned}$$

where ζ_t is, again, the log-chi-squared random variable defined in (24). We view (29)-(30) as the true model under analysis and refer to it hereafter as the SV-SQ model.

To implement an auxiliary likelihood-based ABC algorithm, we adopt a Gaussian approximation for ζ_t in (29) and an Euler discretization for (30), yielding the approximating model,

$$\begin{aligned} y_t &= \ln(x_t) + e_t \quad (31) \\ x_t &= \beta_1 + \beta_2 x_{t-1} + \beta_3\sqrt{x_{t-1}}v_t, \quad (32) \end{aligned}$$

where $e_t \sim N(0, \sigma_\zeta^2)$, v_t is a truncated Gaussian variable with lower bound, $v_t > \frac{-\beta_1}{\beta_3}$, and we define the auxiliary parameters as $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$. Similar parameter restrictions to

those imposed on the structural parameters $\boldsymbol{\phi}$ are required of the elements of $\boldsymbol{\beta}$: $M \geq \beta_1$, $\beta_3 \geq \varphi > 0$, $\varphi \leq \beta_2 \leq 1 - \varphi$, and $2\beta_1 \geq \beta_3^2$. Importantly, and in contrast to the situation that obtains in the previous section, the discretization in (32) is not exact; hence $\boldsymbol{\phi} \neq \boldsymbol{\beta}$, and the precise link between the two sets of parameters impacts (in particular) on the identification condition for consistency, **I2**. Moreover, non-linearities characterize both (31) and (32); hence, and also in contrast with the previous example, the KF is not a feasible method for evaluating the auxiliary likelihood function, $L_a(\mathbf{y}; \boldsymbol{\beta})$. Therefore, we turn to the augmented unscented KF (AUKF) as an alternative means of evaluating the likelihood of this approximation. General pseudo code detailing implementation of the AUKF is given in Algorithm 2, with more detailed implementation instructions given in Appendix A.3.1.

Algorithm 2 General AUKF algorithm

1: Initialize the system in (31) and (32) with a matrix of sigma-points $X_{a,0}$ and a vector of fixed weights; see, Appendix A.3.1 for the definition of these sigma-points and weights;
2: **while** $t \leq T$ **do**
3: Propagate $X_{a,t-1}$ through (32) to obtain x_t sigma points for time t;
4: Using simple weighted sums of the x_t sigma points, generate the predicted mean and variance for x_t;
5: Use the predicted mean and variance to generate a new matrix of sigma points $X_{a,t}$;
6: Propagate $X_{a,t}$ through (31) to obtain y_t sigma points for time t;
7: Using simple weighted sums of the y_t sigma points, generate the predicted mean and variance for y_t;
8: Use the predicted mean and variance to form a Gaussian conditional density for y_t;
9: Using the predicted mean and variance for y_t and KF up-dating, produce the filtered mean and variance for x_t, given the observation of y_t, and up-date the sigma points $X_{a,t}$ accordingly;
10: Set $t = t + 1$;
11: **end while**
12: $L_a(\mathbf{y}; \boldsymbol{\beta})$ is the log-product of the increments in Step 8.

Given $L_a(\mathbf{y}; \boldsymbol{\beta})$ as evaluated via the AUKF, once again ABC could be based on either distance: (11) or (18). However, unlike in the case of the SV-OU model, the precise form of the auxiliary likelihood function depends on the first-order Euler discretization of the continuous-time state process and the particular specifications used to implement the AUKF. For the AUKF specification detailed in Appendix A.3.1, we state the following corollary, the proof of which is given in Appendix A.3.2:

Corollary 3 *For the SV-SQ model in* (29)-(30) *and true value* $\boldsymbol{\phi}_0 = (\delta_0, \alpha_0, \sigma_{v,0})'$*, the model in* (31) *and* (32)*, with auxiliary likelihood* $L_a(\mathbf{y}; \boldsymbol{\beta})$ *constructed via the AUKF filter, and with* $\boldsymbol{\eta}(\mathbf{y}) = \widehat{\boldsymbol{\beta}}(\mathbf{y})$*, satisfies Assumption (**A**).*

Bayesian consistent inference for ϕ_0 however, is also dependent on the satisfaction of Assumption (**I**). Whilst **I1** is trivially satisfied via the specification of a sensible prior, **I2** does not appear to be amenable to analytical investigation, or verification, given the nature of the auxiliary likelihood, as numerically evaluated using the AUKF, and the role played by the true and auxiliary parameters therein. Hence, we choose to investigate consistency for this model numerically as part of the exercise that follows.

5 Numerical assessment of the auxiliary likelihood-based ABC method

We undertake a numerical exercise in which the accuracy of the auxiliary likelihood-based method of ABC is compared with that of ABC methods based on a set of summary statistics chosen without explicit reference to an auxiliary likelihood. We conduct this exercise using the SV-SQ model in (29)-(30) as the example, with the auxiliary score produced by evaluating the likelihood function of the approximating model (defined by (31) and (32)) using the AUKF in the manner described above. We assess the accuracy of all ABC posterior estimates with reference to the exact (marginal) posteriors for a given (finite) sample size, as well as illustrating the behavior of the estimated densities as the sample size increases. In particular, we illustrate that despite the difficulty of formally establishing the identification condition (Assumption **I2)** for the auxiliary likelihood-based ABC approach, Bayesian consistency would appear to hold. In contrast, there is no clear evidence that the summary statistic-based ABC estimates yield consistent inference. We consider both the case where a single parameter (only) is unknown (and dimensionality thus plays no role), and the case where two, and then all three parameters of the model are unknown. In Section 5.2 we propose a dimension reduction technique for the multi-parameter case, based on marginalization of the auxiliary likelihood. All results are then documented in Section 5.3.

5.1 Data generation and computational details

For the purpose of this illustration we simulate artificially an 'empirical' sample of size T from the model in (29) and (30), with the parameters set to values that yield observations on both r_t and V_t that match the characteristics of (respectively) daily returns and daily values of realized volatility (constructed from 5 minute returns) for the S&P500 stock index over the 2003-2004 period: namely, $\alpha = 0.1$; $\delta = 0.004$; $\sigma_v = 0.062$. This relatively calm period in the stock market is deliberately chosen as a reference point, as the inclusion of price and volatility jumps, and/or a non-Gaussian conditional distribution in the model would be an empirical necessity for any more volatile period, such as that witnessed

during the 2008/2009 financial crisis, for example. The aim of this exercise being to assess the accuracy of the alternative ABC methods in a non-linear state space setting, it is important to have access to the exact posterior, and the SV-SQ model - without additional distributional complexities - enables this posterior to be accessed, via the deterministic non-linear filtering method of Ng *et al.* (2013). In brief, the method of Ng *et al.* represents the recursive filtering and prediction distributions used to define the exact likelihood function as the numerical solutions of integrals defined over the support of ζ_t in (29), with deterministic integration used to evaluate the relevant integrals, and the *exact* transitions in (30) used in the specification of the filtering and up-dating steps. Whilst lacking the general applicability of the ABC-based method proposed here, this deterministic filtering method is ideal for the particular model used in this illustration, and can be viewed as producing a very accurate estimate of the exact density, without any of the simulation error that would be associated with an MCMC-based comparator, for instance. We refer the reader to Ng *et al.* for more details of the technique; see also Kitagawa (1987).[5] The likelihood function, evaluated via this method, is then multiplied by a uniform prior that imposes the restrictions: $0 < \alpha < 1$; δ, $\sigma_v^2 > 0$ and $2\delta \geq \sigma_v^2$. The three marginal posteriors are then produced via deterministic numerical integration (over the parameter space), with a very fine grid on $\boldsymbol{\phi}$ being used to ensure accuracy. For ease of interpretation we report the posterior results for $\rho = 1 - \alpha$, where values of ρ close to unity signify a very persistent volatility process.

We compare the performance of the score-based technique with that of more conventional ABC methods based on summary statistics that may be deemed to be a sensible choice in this setting. For want of a better choice we propose a set of summary statistics that are sufficient for an observable AR(1) process (under Gaussianity), as given in (10). Two forms of distances are used. Firstly, we apply the conventional Euclidean distance, with each summary statistic also weighted by the inverse of the variance of the values of the statistic across the ABC draws. That is, we define

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\phi}^i))\} = [\sum_{j=1}^{5} (s_j^i - s_j^{obs})^2 / var(s_j)]^{1/2} \tag{33}$$

for ABC iteration $i = 1., 2, ..., N$, where $var(s_j)$ is the variance (across i) of the s_j^i, and s_j^{obs} is the observed value of the jth statistic. Secondly, we use a distance measure proposed

[5] We note that the application of this filter in Ng *et al.* is to a non-parametric representation of the measurement error. In the current setting, in which ζ_t is specified parametrically, the known form of the distribution of ζ_t is used directly in the evaluation of the relevant integrals. We refer the reader to Section 2.2 of that paper for a full description of the algorithm. Preliminary experimentation with the number of grid points used in the deterministic integration was undertaken in order to ensure that the resulting estimate of the likelihood function/posterior stabilized, with 100 grid points underlying the final results documented here. As an additional check we also evaluated the exact (normalized) likelihood function using a bootstrap particle filter, based on 50,000 particle draws. The filtering-based estimate was indistinguishable from the grid-based estimate and, hence, is not reproduced here.

in Fearnhead and Prangle (2012) which, as made explicit in Blum *et al.* (2013), is a form of dimension reduction method. We explain this briefly as follows. Given the vector of observations $\mathbf{y}$, the set of summary statistics in (10) are used to produce an estimate of $E(\phi_j|\mathbf{y})$, $j = 1, 2, 3$, which, in turn, is used as the summary statistic in a subsequent ABC algorithm. The steps of the Fearnhead and Prangle (FP) procedure (as modified for this context) for selection of the scalar parameter ϕ_j, $j = 1, 2, 3$, are as given in Algorithm 3.

Algorithm 3 FP ABC algorithm

1: Simulate $\boldsymbol{\phi}^i$, $i = 1, 2, ..., N$, from $p(\boldsymbol{\phi})$

2: Simulate $\mathbf{x}^i = (x_1^i, x_2^i, ..., x_T^i)'$ from (30) using the exact transitions, and pseudo data, $\mathbf{z}^i$ using $p(\mathbf{z}|\mathbf{x}^i)$

3: Given $\mathbf{z}^i$, construct

$$\mathbf{s}^i = \left[s_1^i, s_2^i, s_3^i, s_4^i, s_5^i\right]' \tag{34}$$

4: For $\boldsymbol{\phi}_j = (\phi_j^1, \phi_j^2, ..., \phi_j^N)'$, $\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \mathbf{s}^1 & \mathbf{s}^2 & \cdots & \mathbf{s}^N \end{bmatrix}'$ and $\boldsymbol{\phi}_j = E[\boldsymbol{\phi}_j|\mathbf{Z}] + \mathbf{e} = \mathbf{X}\begin{bmatrix} \alpha & \boldsymbol{\gamma}' \end{bmatrix}' + \mathbf{e}$, where $\mathbf{Z} = \left[\mathbf{z}^1, \mathbf{z}^2, ..., \mathbf{z}^N\right]$ and $\boldsymbol{\gamma}$ is of dimension (5×1), use OLS to estimate $E[\boldsymbol{\phi}_j|\mathbf{Z}]$ as $\widehat{E}[\boldsymbol{\phi}_j|\mathbf{Z}] = \widehat{\alpha} + \begin{bmatrix} \mathbf{s}^1 & \mathbf{s}^2 & \cdots & \mathbf{s}^N \end{bmatrix}'\widehat{\boldsymbol{\gamma}}$

5: For $\eta(\mathbf{z}^i) = \widehat{E}(\phi_j|\mathbf{z}^i) = \widehat{\alpha} + \mathbf{s}^{i'}\widehat{\boldsymbol{\gamma}}$ and $\eta(\mathbf{y}) = \widehat{E}(\phi_j|\mathbf{y}) = \widehat{\alpha} + \mathbf{s}^{obs'}\widehat{\boldsymbol{\gamma}}$, where $\mathbf{s}^{obs}$ denotes the vector of summary statistics in (34) calculated from the vector of observed returns, use:

$$d\{\eta(\mathbf{y}), \eta(\mathbf{z}^i)\} = \left|\widehat{E}(\phi_j|\mathbf{y}) - \widehat{E}(\phi_j|\mathbf{z}^i)\right| = \left|\mathbf{s}^{i'}\widehat{\boldsymbol{\gamma}} - \mathbf{s}^{obs'}\widehat{\boldsymbol{\gamma}}\right| \tag{35}$$

as the selection criterion for ϕ_j.

The score-based method uses the distance measure in (15). The weighting matrix $\boldsymbol{\Sigma}$ is set equal to the Hessian-based estimate of the variance-covariance matrix of the (joint) MLE of $\boldsymbol{\beta}$, evaluated at the MLE computed from the observed data, $\widehat{\boldsymbol{\beta}}(\mathbf{y})$. For the case where a single parameter only is unknown, the absolute value of the relevant scalar score is used to define (15).

5.2 Dimension reduction via integrated likelihood techniques

An ABC algorithm induces two forms of approximation error. Firstly, and most fundamentally, the use of a vector of summary statistics $\boldsymbol{\eta}(\mathbf{y})$ to define the selection criterion in (1) means that a simulation-based estimate of $p(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$ is the outcome of the exercise. Only if $\boldsymbol{\eta}(\mathbf{y})$ is sufficient for $\boldsymbol{\phi}$ is the partial posterior density $p(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$ equivalent to the exact posterior $p(\boldsymbol{\phi}|\mathbf{y})$. Secondly, the partial posterior density itself, $p(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$, will be estimated with error, due to both the use of a non-zero tolerance, ε, and the use of a finite set of draws, N to estimate $p(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$, for any given ε. Typically ε is chosen such that, for a given computational burden (i.e. a given value of N), a certain (small) proportion of draws of $\boldsymbol{\phi}^i$ are selected, with attempts then made to reduce the error associated with

non-parametric density estimation of $p(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$ via post-sampling corrections of the draws (Beaumont *et al.*, 2002, Blum, 2010, Blum and François, 2010). Critically, as highlighted by Blum, the accuracy of the estimate of $p(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$ (for any given ε and N) will be less, the larger the dimension of $\boldsymbol{\eta}(\mathbf{y})$. This 'curse of dimensionality' obtains even when the parameter $\boldsymbol{\phi}$ is a scalar, and relates solely to the dimension of $\boldsymbol{\eta}(\mathbf{y})$. As elaborated on further by Nott *et al.* (2014), this problem is exacerbated as the dimension of $\boldsymbol{\phi}$ itself increases, firstly because an increase in the dimension of $\boldsymbol{\phi}$ brings with it a concurrent need for an increase in the dimension of $\boldsymbol{\eta}(\mathbf{y})$ and, secondly, because the need to estimate a multi-dimensional density (for $\boldsymbol{\phi}$) brings with it its own problems related to dimension.

As a potential solution to the inaccuracy induced by the dimensionality of the problem, Nott *et al.* (2014) suggest allocating (via certain criteria) a subset of the full set of summary statistics to each element of $\boldsymbol{\phi}$, ϕ_j, $j = 1, 2, ..., p$, using kernel density techniques to estimate each marginal density, $p(\phi_j|\boldsymbol{\eta}_j(\mathbf{y}))$, and then using standard techniques to retrieve a more accurate estimate of the joint posterior, $p(\boldsymbol{\phi}|\boldsymbol{\eta}(\mathbf{y}))$, if required. However, the remaining problem associated with the (possibly still high) dimension of each $\boldsymbol{\eta}_j(\mathbf{y})$, in addition to the very problem of defining an appropriate set $\boldsymbol{\eta}_j(\mathbf{y})$ for each ϕ_j, remains unresolved.[6]

The principle advocated in this paper is to exploit the information content in the MLE of the parameters of an auxiliary likelihood, $\boldsymbol{\beta}$, to yield summary measures on which to condition. Within this framework, the dimension of $\boldsymbol{\beta}$ determines the dimension of $\boldsymbol{\eta}(\mathbf{y})$ and the curse of dimensionality thus prevails for high-dimensional $\boldsymbol{\beta}$. However, in this case a solution is available, at least when the dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are equivalent and there is a natural link between the elements of the two parameter vectors, which is clearly so for the model investigated in the numerical exercise, in which we produce an approximation by discretization of the latent diffusion. In this case then, integrating the auxiliary likelihood function with respect to all parameters other than β_j and then producing the score of this function with respect to β_j (as evaluated at the integrated MLE from the observed data, $\widehat{\beta}_j(\mathbf{y})$), yields, by construction, an obvious scalar statistic for use in selecting draws of ϕ_j and, hence, a method for estimating $p(\phi_j|\mathbf{y})$.[7] If the marginal posteriors only are of interest, then all p marginals can be estimated in this way, with p applications of $(p-1)$-dimensional integration required at each step within ABC to produce the relevant score statistics. Importantly, we do not claim here that the 'proximity' to sufficiency (for $\boldsymbol{\phi}$) of the vector statistic $\boldsymbol{\eta}(\mathbf{y})$, translates into an equivalent relationship between the score of the integrated likelihood function and the corresponding scalar parameter, nor that the associated product density is coherent with a joint probability distribution. If the joint

[6] See Blum *et al.* (2013) for further elaboration on the dimensionality issue in ABC and a review of current approaches for dealing with the problem.

[7] For a general discussion of the use of integrated likelihood methods in statistics see Berger et al. (1999).

posterior (of the full vector $\boldsymbol{\phi}$) is of particular interest, the sort of techniques advocated by Nott *et al.* (2014), amongst others, can be used to yield joint inference from the estimated marginal posteriors.

Put more formally, let $\boldsymbol{\beta}_{-j} = (\beta_1, ..., \beta_{j-1}, \beta_{j+1}, ..., \beta_p)'$ be a $(p-1)$-dimensional parameter vector of auxiliary parameters and $\mathbf{B}_{-j} \subset \mathbb{R}^{(p-1)}$ the parameter space associated with $\boldsymbol{\beta}_{-j}$. For $p(\boldsymbol{\beta}_{-j}|\beta_j)$ the conditional prior probability of $\boldsymbol{\beta}_{-j}$, define the integrated likelihood $L_a^I(\mathbf{y}; \beta_j)$ as

$$L_a^I(\mathbf{y}; \beta_j) = \int_{\mathbf{B}_{-j}} L_a(\mathbf{y}; \boldsymbol{\beta}) p(\boldsymbol{\beta}_{-j}|\beta_j) d\boldsymbol{\beta}_{-j}.$$

For a given auxiliary model and conditional prior specification, $L_a^I(\mathbf{y}; \beta_j)$ can be used to obtain convenient scalar summary statistics for use in estimating $p(\phi_j|\mathbf{y})$ via ABC. In the specific setting where the auxiliary model is a discretization of the continuous time model, there is a natural link between the elements of the two parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$, and summary statistics for ABC can be based on integrated MLEs of the auxiliary likelihood $L_a(\mathbf{y}; \boldsymbol{\beta})$, defined as

$$\widehat{\beta}_j = \arg\max_{\beta_j} L_a^I(\mathbf{y}; \beta_j) \text{ and } \widehat{\beta}_j(\boldsymbol{\phi}) = \arg\max_{\beta_j} L_a^I(\mathbf{z}(\boldsymbol{\phi}); \beta_j),$$

or integrated scores evaluated at the integrated MLE $\widehat{\beta}_j$, defined as

$$S^I(\mathbf{y}; \widehat{\beta}_j) = \frac{\partial \log\left(L_a^I(\mathbf{y}; \beta_j)\right)}{\partial \beta_j}|_{\beta_j = \widehat{\beta}_j} \text{ and } S^I(\mathbf{z}(\boldsymbol{\phi}); \widehat{\beta}_j) = \frac{\partial \log\left(L_a^I(\mathbf{z}(\boldsymbol{\phi}); \beta_j)\right)}{\partial \beta_j}|_{\beta_j = \widehat{\beta}_j}.$$

Marginal ABC-based posterior estimates of $p(\phi_j|\mathbf{y})$ are then obtained by 'selecting' draws of ϕ_j^i according to the selection criterion in (1), with $\eta(\mathbf{y})$ given by either of the *univariate* summaries $\widehat{\beta}_j$ or $S^I(\mathbf{y}; \widehat{\beta}_j)$. More generally, for a well-chosen auxiliary model, there is likely a *qualitative* link between the auxiliary and structural parameters (e.g. location, scale, tail behavior, persistence) that can be exploited to decide which univariate summary to use as the selection statistic for any given ϕ_j.

5.3 Numerical results

5.3.1 Finite sample accuracy

In order to abstract initially from the impact of dimensionality on the ABC methods, we first report results for each single parameter of the SV-SQ model, keeping the remaining two parameters fixed at their true values, and for a relatively small sample size of $T = 500$. Three ABC-based estimates of the relevant exact (univariate) posterior, based on a uniform prior, are produced in this instance, with all estimates produced by applying kernel smoothing methods to the accepted draws for each ABC algorithm. Three matching

statistics are used, respectively: 1) the (uni-dimensional) score based on the auxiliary likelihood function as evaluated via the AUKF (ABC-score); 2) the summary statistics in (10), matched via the Euclidean distance measure in (33) (ABC-summ stats); and 3) the summary statistics in (10), matched via the FP distance measure in (35) (ABC-FP). We produce representative posterior (estimates) in each case, to give some visual idea of the accuracy (or otherwise) that is achievable via the ABC methods. We then summarize accuracy by reporting the average, over 50 runs of ABC, of the root mean squared error (RMSE) of each ABC-based estimate of the exact posterior for a given parameter, computed as:

$$RMSE = \sqrt{\frac{1}{G}\sum_{g=1}^{G}(\widehat{p}_g - p_g)^2}, \tag{36}$$

where $\widehat{p}_g$ is the ordinate of the ABC density estimate and p_g the ordinate of the exact posterior density, at the gth grid-point used to produce the plots. All single parameter results are documented in Panel A of Table 1, with (average) RMSEs for a given parameter reported as a ratio to that of ABC-score.

Figure 1, Panel A reproduces the exact posterior of (the single unknown parameter) ρ and the three ABC-based estimates, for a single run of ABC. As is clear, the auxiliary score-based ABC estimate (denoted by 'ABC - score method' in the key) provides a very accurate estimate of the exact posterior, using only 50,000 replications of the simplest accept/reject ABC algorithm, and fifteen minutes of computing time on a desktop computer. In contrast, the ABC method based on the summary statistics, combined using a Euclidean distance measure, performs very poorly. Whilst the dimensional reduction technique of Fearnhead and Prangle (2012), applied to this same set of summary statistics, yields some improvement, it does not match the accuracy of the score-based method. Comparable graphs are produced for the single parameters δ and σ_v in Panels B and C respectively of Figure 1, with the remaining pairs of parameters (ρ and σ_v, and ρ and δ respectively) held fixed at their true values. In the case of δ, the score-based method provides a reasonably accurate representation of the shape of the exact posterior and, once again, a better estimate than both summary statistic methods, both of which illustrate a similar degree of accuracy (to each other) in this instance. For the parameter σ_v, none of the three ABC methods provide a particularly accurate estimate of the exact posterior.

The RMSE results recorded in Panel A of Table 1 confirm the qualitative nature of the single-run graphical results. For both ρ and δ, the score-based ABC method produces the most accurate estimate of the exact posterior of all comparators. In the case of σ_v the Fearnhead and Prangle (2012) method yields the most accurate estimate, but there is really little to choose between all three methods.

The results recorded in Panels B to D highlight that when either two or three parameters are unknown the score-based ABC method produces the most accurate density

Figure 1: Posterior densities for each single unknown parameter of the model in (29) and (30), with the other two parameters set to their true values. As per the key, the graphs reproduced are the exact posterior in addition to the three ABC-based estimates. The exact posterior is evaluated using the grid-based non-linear filter of Ng *et al.* (2013). The ABC-based estimates are produced by applying kernel smoothing methods to the accepted draws. All results are based on a sample size of $T = 500$.

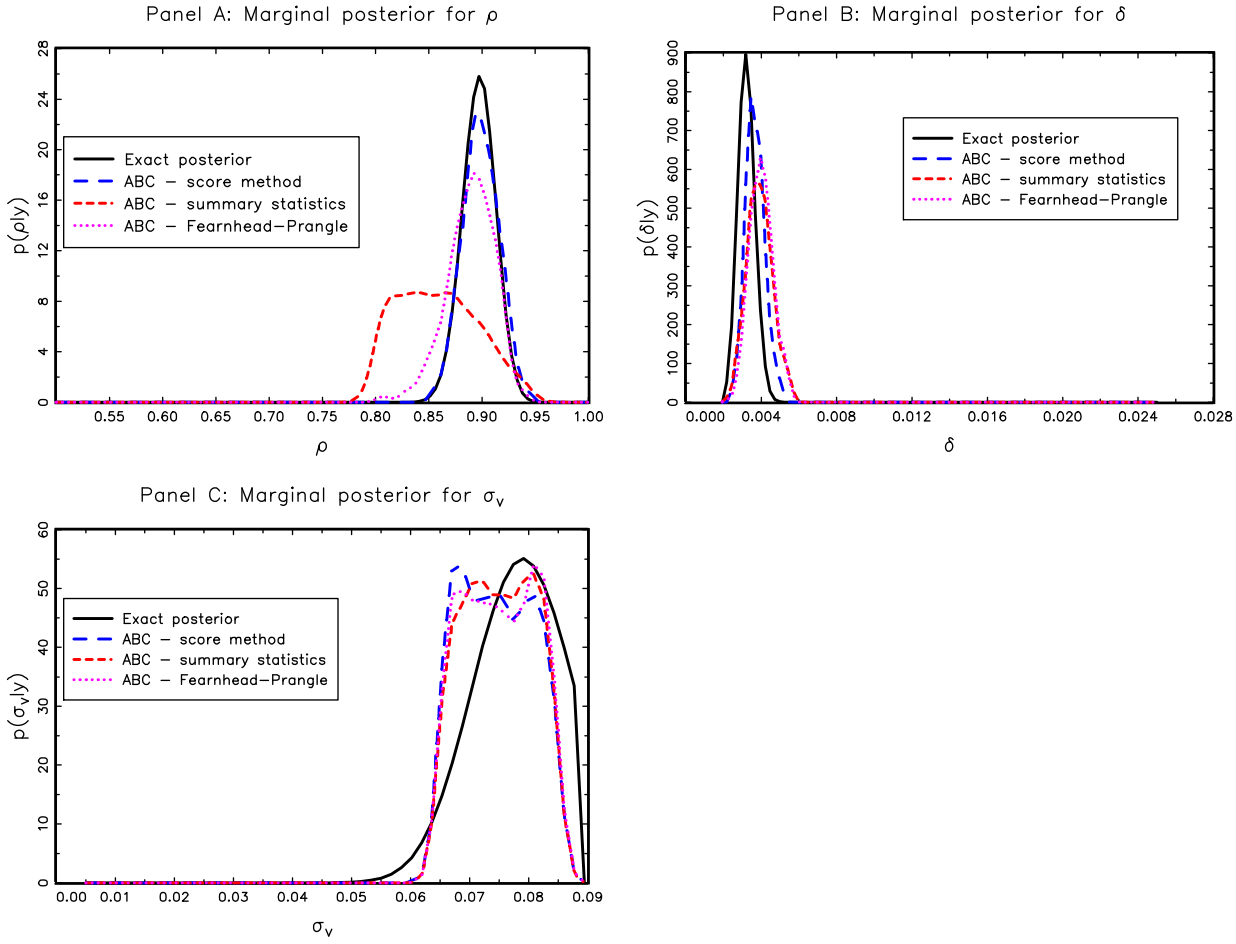

Table 1: Average RMSE of an estimated marginal posterior over 50 runs of ABC (each run using 50,000 replications); recorded as a ratio to the (average) RMSE for the (integrated) ABC score method. 'Score' refers to the ABC method based on the score of the AUKF model; 'SS' refers to the ABC method based on a Euclidean distance for the summary statistics in (10); 'FP' refers to the Fearnhead and Prangle ABC method, based on the summary statistics in (10). For the single parameter case, the (single) score method is documented in the row denoted by 'Int Score', whilst in the multi-parameter case, there are results for both the joint (Jt) and integrated (Int) score methods. The bolded figure indicates the approximate posterior that is the most accurate in any particular instance. The sample size is $T = 500$.

	Panel A One unknown			Panel B Two unknowns		Panel C Two unknowns		Panel D Three unknowns		
	ρ	δ	σ_v	ρ	σ_v	ρ	δ	ρ	δ	σ_v
ABC Method										
Jt Score	-	-	-	**0.873**	1.843	1.613	1.689	**0.408**	1.652	1.015
Int Score	**1.000**	**1.000**	1.000	1.000	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**
SS	5.295	1.108	1.044	5.079	2.263	4.129	1.986	2.181	1.582	1.093
FP	1.587	1.254	**0.926**	1.823	2.286	3.870	2.416	1.877	2.763	1.101

estimates in *all* cases, with the integrated likelihood technique described in Section 5.2 (recorded as 'Int. Score' in the table) yielding a further improvement in accuracy in five out of the seven cases, auguring quite well for this particular approach to dimension reduction.

5.3.2 Large sample performance

In order to document numerically the extent to which the ABC posteriors become increasingly concentrated around the true parameters (or otherwise) as the sample size increases, we complete the numerical demonstration by recording in Table 2 the average probability mass (over the 50 runs of ABC) within a small interval around the true parameter, for all three ABC posterior estimates. For the purpose of illustration we record these results for the (three) single unknown parameter cases only, using the ABC kernel density ordinates to estimate the relevant probabilities, via rectangular integration. The boundaries of the interval used for a given parameter (recorded at the top of the table) are determined by the grid used to numerically estimate the kernel density.

The results in Panel A (for $T = 500$) accord broadly with the qualitative nature of the single run results recorded graphically in Figure 1, with the score-based method producing superior results for ρ and δ and there being little to choose between the (equally inaccurate)

Table 2: Posterior mass in given intervals around the true parameters, averaged over 50 runs of ABC using 50,000 replications. 'Score' refers to the ABC method based on the score of the AUKF model; 'SS' refers to the ABC method based on a Euclidean distance for the summary statistics in (10); 'FP' refers to the Fearnhead and Prangle ABC method, based on the summary statistics in (10). The bolded figure indicates the largest (average) posterior mass for each case. Results in Panel A are for $T = 500$ and those in Panel B for $T = 2000$. One parameter at a time is treated as unknown.

	Panel A: $T = 500$			Panel B: $T = 2000$		
	ρ	δ	σ_v	ρ	δ	σ_v
True value:	$\rho_0 = 0.9$	$\delta_0 = 0.004$	$\sigma_{\nu 0} = 0.062$	$\rho_0 = 0.9$	$\delta_0 = 0.004$	$\sigma_{\nu 0} = 0.062$
Interval:	(0.88, 0.92)	(0.003, 0.005)	(0.052, 0.072)	(0.88, 0.92)	(0.003, 0.005)	(0.052, 0.072)
ABC Method						
Score	**0.88**	**0.90**	**0.44**	**0.94**	**1.00**	0.85
SS	0.28	0.84	**0.44**	0.24	0.78	**0.87**
FP	0.76	0.89	0.41	0.61	0.91	**0.87**

estimates in the case of σ_v. Importantly, when the sample size increases the score-based density displays clear evidence of increased concentration around the true parameter value, for all three parameters, providing numerical evidence that the identification condition I2 holds. For the two alternative methods however, this is not uniformly the case, with increased concentration occurring for σ_v only. Theoretical results in Frazier *et al.* (2015) indeed highlight the fact that Bayesian consistency is far from assured for any given set of summary statistics; hence the lack of numerical support for consistency when ABC is based on somewhat arbitrarily chosen conditioning statistics is not completely surprising. We refer to the reader to that paper for further discussion of this general point.

6 Conclusions and discussion

This paper has explored the application of approximate Bayesian computation in the state space setting. Certain fundamental results have been established, namely the lack of reduction to finite sample sufficiency and (under regularity) the Bayesian consistency of the auxiliary likelihood-based method. The (limiting) equivalence of ABC estimates produced by the use of both the maximum likelihood and score-based matching statistics

has also been demonstrated. The idea of tackling the dimensionality issue that plagues the application of ABC in high dimensional problems via an integrated likelihood approach has been proposed. The approach has been shown to yield some benefits in the particular numerical example explored in the paper. However, a much more comprehensive analysis of different non-linear settings (and auxiliary models) would be required for a definitive conclusion to be drawn about the trade-off between the gain to be had from marginalization and the loss that may stem from integrating over an *inaccurate* auxiliary likelihood.

Indeed, the most important challenge that remains, as is common to the related frequentist techniques of indirect inference and efficient methods of moments, is the specification of a computationally efficient and accurate approximation. Given the additional need for parsimony, in order to minimize the number of statistics used in the matching exercise, the principle of aiming for a large nesting model, with a view to attaining full asymptotic sufficiency, is not an attractive one. We have illustrated the use of one simple approximation approach based on an auxiliary likelihood constructed via the unscented Kalman filter. The relative success of this approach in the particular example considered, certainly in comparison with methods based on other more *ad hoc* choices of summary statistics, augurs well for the success of score-based methods in the non-linear setting. Further exploration of approximation methods in other non-linear state space models is the subject of on-going research.

Finally, we note that despite the focus of this paper being on inference about the static parameters in the state space model, there is nothing to preclude marginal inference on the states being conducted, at a second stage. Specifically, conditional on the (accepted) draws used to estimate $p(\boldsymbol{\phi}|\mathbf{y})$, existing filtering and smoothing methods (including the recent methods, referenced earlier, that exploit ABC at the filtering/smoothing level) could be used to yield draws of the states, and (marginal) smoothed posteriors for the states produced via the usual averaging arguments. With the asymptotic properties of both approaches established (under relevant conditions), of particular interest would be a comparison of both the finite sample accuracy and the computational burden of the hybrid ABC-based methods that have appeared in the literature, with that of the method proposed herein, in which $p(\boldsymbol{\phi}|\mathbf{y})$ is targeted more directly via ABC principles alone.

References

[1] Beaumont, M.A. 2010. Approximate Bayesian Computation in Evolution and Ecology, *Annual Review of Ecology, Evolution, and Systematic,* 41, 379-406.

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

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

[4] Berger, James O.; Liseo, Brunero; Wolpert, Robert L. Integrated Likelihood Methods for Eliminating Nuisance Parameters. Statist. Sci. 14 (1999), no. 1, 1-28

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

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

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

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

[9] Calvet, C. and Czellar, V. 2015a. Accurate Methods for Approximate Bayesian Computation Filtering. *Journal of Financial Econometrics* 13, 798-838.

[10] Calvet, C. and Czellar, V. 2015b. Through the Looking Glass: Indirect Inference via Simple Equilibria. *Journal of Econometrics* 185, 343-358.

[11] Cornuet, J-M., Santos, F., Beaumont, M.A., Robert, C.P., Marin, J-M., Balding, D.J., Guillemard, T. and Estoup, A. 2008. Inferring Population History with DIY ABC: a User-friendly Approach to Approximate Bayesian Computation, *Bioinformatics* 24, 2713-2719.

[12] Cox, D.R. and Hinkley, D.V. 1974. *Theoretical Statistics*, Chapman and Hall, London.

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

[14] Creel, M., Gao, J., Hong, H. and Kristensen, D. 2015. Bayesian Indirect Inference and the ABC of GMM. *http://arxiv.org/abs/1512.07385.*

[15] Dean, T., Singh, S., Jasra, A. and Peters, G. 2014. Parameter Inference for Hidden Markov Models with Intractable Likelihoods, *Scand. J. Statist.* 41, 970-987.

[16] Drovandi, C.C., Pettitt, A.N. and Faddy, M.J. 2011. Approximate Bayesian Computation Using Indirect Inference, *JRSS(C),* 60 1 - 21.

[17] Drovandi, C.C., Pettitt, A.N. and Lee, A. 2015. Bayesian Indirect Inference Using a Parametric Auxiliary Model. *Statistical Science,* Vol. 30, No. 1, 72-95.

[18] 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: 419–474.

[19] Frazier D.T., Martin G.M., and Robert, C.P., 2015. On Consistency of Approximate Bayesian Computation, *http://arxiv.org/abs/1508.05178.*

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

[21] Ghosal, Subhashis; Ghosh, Jayanta K.; Samanta, Tapas. On Convergence of Posterior Distributions. *Ann. Statist.* 23 (1995), no. 6, 2145-2152.

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

[23] Gouriéroux, C. and Monfort, A. 1995. *Statistics and Econometric Models.* CUP.

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

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

[26] Hansen, L. P. 2012. Proofs for Large Sample Properties of Generalized Method of Moments Estimators, *Journal of Econometrics.* 170, 325-330.

[27] Heggland, K. and Frigessi, A. 2004. Estimating Functions in Indirect Inference, *JRSS(B)* 66, 447-462.

[28] Jasra, A. 2015. Approximate Bayesian Computation for a Class of Time Series Models. *International Statistical Review* 83, 405-435.

[29] Jasra, A, Singh, S, Martin, J., McCoy, E. 2010. Filtering via Approximate Bayesian Computation. Statistics and Computing 22, 1223-1237.

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

[31] Julier, S.J., Uhlmann, J.K. and Durrant-Whyte, H.F. 1995. A New Approach for Filtering Nonlinear Systems. *Proceedings of the American Control Conference*, 1628-1632.

[32] Julier, S.J., Uhlmann, J.K. and Durrant-Whyte, H.F. 2000. A New Method for the Nonlinear Transformation of Means and Covariances in Filters and Estimators, IEEE *Transactions on Automatic Control* 45, 477-481.

[33] Kitagawa, G. 1987. Non-Gaussian State Space Modeling of Nonstationary Time Series. *Journal of the American Statistical Association* 76, 1032-1064.

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

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

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

[37] Martin, J. S., Jasra, A., Singh, S. S., Whiteley, N., Del Moral, P. and McCoy, E. 2014. Approximate Bayesian Computation for Smoothing, *Stoch. Anal. Appl.* 32, 397-422.

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

[39] Ng, J., Forbes, C,S., Martin, G.M. and McCabe, B.P.M. 2013. Non-parametric Estimation of Forecast Distributions in Non-Gaussian, Non-linear State Space Models, *International Journal of Forecasting* 29, 411-430

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

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

[42] Robert, C.P. 2015. Approximate Bayesian Computation: a Survey on Recent Results, *Proceedings, MCQMC2014.*

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

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

[45] Smith, A.A. 1993. Estimating Non-linear Time Series Models Using Vector Autoregressions: Two Approaches, *Journal of Applied Econometrics* 8, 63-84.

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

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

[48] van der Vaart. 1998. *Asymptotic Statistics*. CUP.

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

[50] Yildirim, S, Singh, S.S, Dean, T. and Jasra, A. 2015. Parameter Estimation in Hidden Markov Models with Intractable Likelihoods via Sequential Monte Carlo. *JCGS*, 24, 846-865.

A Proofs of Technical Results

This appendix collects the proofs for the technical results stated in the paper. The proof of Theorem 1 relies on several intermediate results, which, in the name of space, have been detailed in a supplemental appendix, available at http://users.monash.edu.au/~gmartin/.

A.1 Proof of Theorem 1

To simplify the notation, in what follows we use $\widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}(\mathbf{y})$ and $\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}) := \widehat{\boldsymbol{\beta}}(\mathbf{z}(\boldsymbol{\phi}))$. The proof is broken into three parts. First, we demonstrate that the distance

$$d\left\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\phi}))\right\} = d\left\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} \equiv \sqrt{\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right)' \Omega \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right)}$$

converges uniformly in $\boldsymbol{\phi}$ to $d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\}$. The second piece focuses on showing that the only value for which $d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\} \leq \varepsilon$ as $\varepsilon \to 0$ and $T \to \infty$ is $\boldsymbol{\phi}_0$. After one and two have been established, we go on to show that all posterior mass concentrates asymptotically on sets of the form $\left\{\boldsymbol{\phi} \in \boldsymbol{\Phi} : d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\} = 0\right\}$.

We first demonstrate the uniform convergence of $d\{\widehat{\boldsymbol{\beta}},\, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\}$. The triangle inequality yields

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\phi}))\} \leq d\left\{\widehat{\boldsymbol{\beta}}, \mathbf{b}(\boldsymbol{\phi})\right\} + d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\}.$$

Consider $d\left\{\widehat{\boldsymbol{\beta}}, \mathbf{b}(\boldsymbol{\phi})\right\}$ and note that,

$$d\left\{\widehat{\boldsymbol{\beta}}, \mathbf{b}(\boldsymbol{\phi})\right\} \leq d\left\{\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0\right\} + d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\}.$$

The result $\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$ holds by noting the following two points: one, under Assumptions **(A1)-(A4)**, Proposition 1 in the supplemental appendix ensures that $L_a(\mathbf{y}; \boldsymbol{\beta})/T$ satisfies the uniform law of large numbers (ULLN)

$$\sup_{\boldsymbol{\beta} \in \mathbf{B}} |L_a(\mathbf{y}; \boldsymbol{\beta})/T - L_\infty(\boldsymbol{\phi}_0; \boldsymbol{\beta})| \xrightarrow{P} 0;$$

two, by **(A5)** $L_\infty(\boldsymbol{\phi}_0; \boldsymbol{\beta})$ has unique maximizer $\boldsymbol{\beta}_0 = \mathbf{b}(\boldsymbol{\phi}_0)$. From one and two we have, by standard arguments, see, e.g., Theorem 5.9 of van der Vaart (1998), $\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$.

Consistency of $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ implies

$$d\left\{\widehat{\boldsymbol{\beta}}, \mathbf{b}(\boldsymbol{\phi})\right\} \leq o_P(1) + d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\},$$

and so

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\phi}))\} \leq d\left\{\widehat{\boldsymbol{\beta}}, \mathbf{b}(\boldsymbol{\phi})\right\} + d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} \leq d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\} + d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} + o_P(1).$$

By definition,

$$d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} \leq \sup_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\}.$$

The first term on the right hand side of the above inequality is $o_P(1)$ if $\sup_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \|\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}) - \mathbf{b}(\boldsymbol{\phi})\| \xrightarrow{P} 0$. Now, note that, by Corollary 1 in the supplemental appendix, $\forall \delta > 0$, as $T \to \infty$

$$\sup_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \delta} |L(\mathbf{z}(\boldsymbol{\phi}); \boldsymbol{\beta})/T - L_\infty(\boldsymbol{\phi}; \boldsymbol{\beta})| \xrightarrow{P} 0, \tag{37}$$

and by **(A5)** the limit criterion $L_\infty(\boldsymbol{\phi}; \boldsymbol{\beta})$ has unique maximizer, with respect to $\boldsymbol{\beta}$, $\mathbf{b}(\boldsymbol{\phi})$. Using these two facts, we now show, under the maintained assumptions,

$$\sup_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \left\| \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}) - \mathbf{b}(\boldsymbol{\phi}) \right\| = o_P(1). \tag{38}$$

Define the following terms:

$$\begin{aligned} \widetilde{Q}(\boldsymbol{\phi}; \boldsymbol{\beta}) &= L_a(\mathbf{z}(\boldsymbol{\phi}); \boldsymbol{\beta})/T - L_\infty(\boldsymbol{\phi}; \mathbf{b}(\boldsymbol{\phi})), \\ \widetilde{Q}_\infty(\boldsymbol{\phi}; \boldsymbol{\beta}) &= L_\infty(\boldsymbol{\phi}; \boldsymbol{\beta}) - L_\infty(\boldsymbol{\phi}; \mathbf{b}(\boldsymbol{\phi})). \end{aligned}$$

Note that, by **(A3)**, for all $\delta > 0$, if $\sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \|\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}) - \mathbf{b}(\boldsymbol{\phi})\| > \delta$, there exists $\epsilon(\delta) > 0$, such that

$$\sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \|\widetilde{Q}_\infty(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}))\| > \epsilon(\delta).$$

From here, note that

$$\Pr\left(\sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \|\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}) - \mathbf{b}(\boldsymbol{\phi})\| > \delta\right) \leq \Pr\left(\sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \|\widetilde{Q}_\infty(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}))\| > \epsilon(\delta)\right).$$

The result in (38) then follows if $\sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \|\widetilde{Q}_\infty(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}))\| = o_P(1)$.

Uniformly in $\boldsymbol{\phi}$,

$$\begin{aligned}
\|\widetilde{Q}_\infty(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}))\| &\leq \|\widetilde{Q}_\infty(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})) - \widetilde{Q}(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}))\| + \|\widetilde{Q}(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}))\| \\
&= \|L_\infty(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})) - L_a(\mathbf{z}(\boldsymbol{\phi});\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}))/T\| + \|\widetilde{Q}(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}))\| \\
&\leq \sup_{\boldsymbol{\beta}\in\mathbf{B}} \|L_\infty(\boldsymbol{\phi};\boldsymbol{\beta}) - L_a(\mathbf{z}(\boldsymbol{\phi});\boldsymbol{\beta})/T\| + \|\widetilde{Q}(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}))\| \\
&\leq o_P(1) + \|\widetilde{Q}(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}))\|. \qquad (39)
\end{aligned}$$

The first inequality follows from the triangle inequality, the second from the definition of $\widetilde{Q}_\infty(\boldsymbol{\phi};\boldsymbol{\beta})$, $\widetilde{Q}(\boldsymbol{\phi};\boldsymbol{\beta})$, the third from the definition of sup, and the last from Corollary 1 in the supplemental appendix.

From (39), the result follows if

$$\sup_{\boldsymbol{\phi}\in\boldsymbol{\Theta}} \|\widetilde{Q}(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}))\| = o_P(1).$$

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

$$\begin{aligned}
\|\widetilde{Q}(\boldsymbol{\phi};\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}))\| &\leq \inf_{\boldsymbol{\beta}\in\mathbf{B}} \|\widetilde{Q}(\boldsymbol{\phi};\boldsymbol{\beta})\| + o_P(1) \\
&\leq \inf_{\boldsymbol{\beta}\in\mathbf{B}} \|\widetilde{Q}(\boldsymbol{\phi};\boldsymbol{\beta}) - \widetilde{Q}_\infty(\boldsymbol{\phi};\boldsymbol{\beta})\| + \inf_{\boldsymbol{\beta}\in\mathbf{B}} \|\widetilde{Q}_\infty(\boldsymbol{\phi};\boldsymbol{\beta})\| + o_P(1) \\
&\leq \sup_{\boldsymbol{\beta}\in\mathbf{B}} \|L_a(\mathbf{z}(\boldsymbol{\phi});\boldsymbol{\beta})/T - L_\infty(\boldsymbol{\phi};\boldsymbol{\beta})\| + 0 + o_P(1) \\
&\leq o_P(1). \qquad (40)
\end{aligned}$$

Combining equations (39) and (40) yields $\sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \left\|\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}) - \mathbf{b}(\boldsymbol{\phi})\right\| = o_P(1)$. It then follows that $d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} = o_P(1)$, uniformly in $\boldsymbol{\phi}$ and so

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}(\boldsymbol{\phi}))\} = d\left\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} \xrightarrow{P} d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\}.$$

The second portion of the proof is complete if we demonstrate, as $T \to \infty$ and $\varepsilon \to 0$, the only value of $\boldsymbol{\phi}$ which the ABC algorithm selects is $\boldsymbol{\phi} = \boldsymbol{\phi}_0$. From the definition of the algorithm and the triangle inequality, for T large enough,

$$\begin{aligned}
d\left\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} &\leq d\left\{\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0\right\} + d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} + d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\} \\
&\leq \varepsilon/3 + \varepsilon/3 + d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\}
\end{aligned}$$

A draw from the prior $p(\boldsymbol{\phi})$, will then be accepted if

$$d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\} \leq \varepsilon/3.$$

By Assumption **(I2)**, the only value of $\boldsymbol{\phi}$ such that $d\left\{\mathbf{b}(\boldsymbol{\phi}), \boldsymbol{\beta}_0\right\} \leq \varepsilon/3$ as $\varepsilon \to 0$ is $\boldsymbol{\phi} = \boldsymbol{\phi}_0$.

Now, the result follows if for any $\delta > 0$, as $T \to \infty$ and for some sequence $\varepsilon \to 0$,

$$\Pr\left(d\left\{\boldsymbol{\phi}_0, \boldsymbol{\phi}\right\} > \delta | d\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\} \leq \varepsilon\right) \to 0$$

in probability. First, consider the set

$$A_\epsilon(\delta') := \left\{(\mathbf{z}, \boldsymbol{\phi}) : \{d\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\} \leq \varepsilon\} \text{ and } \left\{d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\} > \delta'\right\}\right\}$$

and the probability

$$\Pr\left(d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\} > \delta' | d\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\} \leq \varepsilon\right).$$

For all $(\mathbf{z}, \boldsymbol{\phi}) \in A_\epsilon(\delta')$, we have, by the triangle inequality:

$$\delta' < d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\} \leq d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} + d\left\{\widehat{\boldsymbol{\beta}}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}\right\} + d\left\{\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0\right\}.$$

For T large enough, $d\left\{\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0\right\} \leq \varepsilon/3$ and we have

$$\delta' - (4/3)\varepsilon < d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\}.$$

Note that, for $\delta' \geq (5/3)\varepsilon$, and $P_{\boldsymbol{\phi}}(\cdot)$ the conditional distribution of $\mathbf{z}$ given $\boldsymbol{\phi}$,

$$\begin{aligned}
\Pr[A_\epsilon(\delta')] &\leq \int_\Phi P_\phi\left(d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\} > \delta'\right) dp(\boldsymbol{\phi}) \\
&\leq \int_\Phi P_\phi\left(d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} > \delta' - (4/3)\varepsilon\right) dp(\boldsymbol{\phi}),
\end{aligned}$$

Hence, we can conclude, for $\delta' = 4/3\varepsilon + s$ and $s \geq \varepsilon/3$,

$$\Pr\left(d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\} > \delta' | d\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\} \leq \varepsilon\right) \leq \frac{\int_\Phi P_{\boldsymbol{\phi}}\left(d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} > s\right) dp(\boldsymbol{\phi})}{\int_\Phi P_{\boldsymbol{\phi}}\left(d\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\} \leq \varepsilon\right) dp(\boldsymbol{\phi})}. \tag{41}$$

First, focus on the denominator in (41). By the triangle inequality

$$d\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\} \leq d\{\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0\} + d\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\} + d\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\}.$$

From $\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$ and the uniform convergence in equation (38) we have, for T large enough and any $\boldsymbol{\phi} \in \boldsymbol{\Phi}$,

$$d\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\} \leq \varepsilon/3 + \varepsilon/3 + d\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\}.$$

Then, for any any $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ with $d\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\} \leq \varepsilon/3$

$$\begin{aligned}\int_{\Phi} P_{\phi}\left(d\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\} \leq \varepsilon\right) dp(\boldsymbol{\phi}) &\geq \int_{d\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\} \leq \varepsilon/3} P_{\phi}\left(d\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\} \leq \varepsilon/3\right) dp(\boldsymbol{\phi}) \\ &\geq \frac{\Pr\left[d\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\} \leq \varepsilon/3\right]}{2} + o(1), \end{aligned} \tag{42}$$

where the last inequality follows from the uniform convergence in equation (38) and the dominated convergence theorem. By Assumption **I**, the prior probability $\Pr\left[d\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\} \leq \varepsilon\right] > 0$ for any $\varepsilon \to 0$. Therefore, using equation (42) within equation (41) yields

$$\Pr\left(d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\} > \delta' | d\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\} \leq \varepsilon\right) \leq \frac{2\int_{\Phi} P_{\phi}\left(d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} > s\right) dp(\boldsymbol{\phi}),}{\Pr\left[d\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\} \leq \varepsilon/3\right] + o(1)}. \tag{43}$$

The result follows if the numerator on the right-hand-side of (43) (equivalently, the numerator of (41)) converges to zero. Using (38), for any $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, $P_{\phi}\left(d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} > s\right) = o(1)$ for $s \to 0$ slower than $d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} \to 0$. Hence, with $\varepsilon \leq 3s$ as defined above, taking $\varepsilon = o(1)$, such that $\varepsilon^{-1} \sup_{\phi \in \Phi} d\left\{\mathbf{b}(\boldsymbol{\phi}), \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\right\} = o_P(1)$, and using (38) yields

$$\Pr\left(d\left\{\boldsymbol{\beta}_0, \mathbf{b}(\boldsymbol{\phi})\right\} > \delta' | d\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\phi})\} \leq \varepsilon\right) \to 0.$$

The result now follows from the arbitrary choice of δ', and the continuity and injectivity of the map $\boldsymbol{\phi} \mapsto \mathbf{b}(\boldsymbol{\phi})$. ■

A.2 SV-OU Example: Proof of Corollary 2

The result follows by verifying the Assumptions of Theorem 1. First, we consider verification of Assumption **A**, which requires establishing that $L_a(\mathbf{y}; \boldsymbol{\beta})$ and $L_a(\mathbf{z}(\boldsymbol{\phi}^i); \boldsymbol{\beta})$ are sufficiently well-behaved. To this end, with reference to (28), define the one-step-ahead predictors, $\hat{x}_{t|t-1} = \beta_1 \hat{x}_{t-1}$ and $P^x_{t|t-1} = \beta_1^2 P^x_{t-1} + \beta_2$, and for the filtering steps define

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + P^x_{t|t-1}\xi_t / f_{t|t}, \;\; P^x_{t|t} = P^x_{t|t-1} - (P^x_{t|t-1})^2 / f_{t|t}, \tag{44}$$

where $\xi_t = y_t - \hat{x}_{t|t-1}$, and $f_{t|t} = P^x_{t|t-1} + \sigma^2_{\zeta}$. Likewise, denote the versions of $\hat{x}_{t|t}$ and $P^x_{t|t}$ based on simulated data $\{z_t(\boldsymbol{\phi}^i)\}_{t=1}^T$ by $\hat{x}_{t|t}(\boldsymbol{\phi}^i)$ and $P^x_{t|t}(\boldsymbol{\phi}^i)$, where $\hat{x}_{t|t}(\boldsymbol{\phi}^i)$ and $P^x_{t|t}(\boldsymbol{\phi}^i)$ are defined equivalently to $\hat{x}_{t|t}$ and $P^x_{t|t}$ (respectively) in (44) with y_t replaced by $z_t(\boldsymbol{\phi}^i)$. To close the recursions, for simplicity, let the initial conditions be known and given by $a_0 = 0, P_0 = \beta_2/(1-\beta_1)$.

Using these definitions, the filtered auxiliary log-likelihood function is given by, up to a constant,

$$L_a(\mathbf{y}; \boldsymbol{\beta}) = \sum_{t=1}^{T} \ell(y_t; \boldsymbol{\beta}) = -\sum_{t=1}^{T} \frac{1}{2}\ln(f_{t|t}) + \frac{1}{2}(\xi_t^2 / f_{t|t}).$$

From these definitions we can now verify Assumptions **(A2)-(A5)** (Assumption **(A1)** can be verified by inspection) for the approximating model for the SV-OU model.

(A2) The condition $M \geq \kappa \geq \varphi > 0$ ensures $e^{-\kappa} < 1 - e^{-\varphi}$, which ensures x_t is stationary. The stated restrictions $M \geq \sigma \geq \varphi > 0$ then guarantees satisfaction of **(A2)**.

(A3) To determine the continuity of $(\boldsymbol{\phi}^i, \boldsymbol{\beta}) \mapsto L_a(\mathbf{z}(\boldsymbol{\phi}^i); \boldsymbol{\beta})$, i.e., satisfaction of **(A3)**, first note that continuity with respect to $\boldsymbol{\beta}$ follows from the definitions of $\hat{x}_{t|t}$, $P^x_{t|t}$ and $L_a(\mathbf{z}(\boldsymbol{\phi}^i); \boldsymbol{\beta})$. Continuity with respect to $\boldsymbol{\phi}^i$ follows from continuity, in $\boldsymbol{\phi}^i$, of $z_t(\boldsymbol{\phi}^i)$, $\hat{x}_{t|t}(\boldsymbol{\phi}^i)$, $P^x_{t|t}(\boldsymbol{\phi}^i)$, and the definition of $L_a(\mathbf{z}(\boldsymbol{\phi}^i); \boldsymbol{\beta})$.

(A4) Assumption **(A4)** follows from the stated restrictions on $\boldsymbol{\Phi}$ and the fact that $\boldsymbol{\beta}$ is a continuous bounded transformation of $\boldsymbol{\phi}$; i.e., $\boldsymbol{\beta} = (\exp(-\kappa), \sigma\sqrt{1 - \exp(-2\kappa)}/2\kappa)'$ and so if $\kappa \geq \varphi > 0$, $\beta_1 < 1$ and if $\sigma \geq \varphi$, $\beta_2 \geq \varphi$, which together imply $E_{\boldsymbol{\phi}^i}[|\ell(z_t(\boldsymbol{\phi}^i), \boldsymbol{\beta})|] < \infty$.

(A5) Satisfaction of **(A5)** requires that $\mathbf{b}(\boldsymbol{\phi}^i)$ uniquely maximize $L_\infty(\boldsymbol{\phi}^i, \boldsymbol{\beta}) = E_{\boldsymbol{\phi}^i}\left[\ell(z_t(\boldsymbol{\phi}^i); \boldsymbol{\beta})\right]$, where $E_{\boldsymbol{\phi}^i}[\cdot]$ denotes expectation under $\boldsymbol{\phi}$ and not $\boldsymbol{\phi}_0$. Primitive conditions guaranteeing this are as follows (see, e.g., Newey and McFadden (1994)): 1) $E_{\boldsymbol{\phi}^i}[|\ell(z_t(\boldsymbol{\phi}^i), \boldsymbol{\beta})|] < \infty$ for all $\boldsymbol{\beta} \in \mathbf{B}$; 2) $\ell(z_t(\boldsymbol{\phi}^i); \boldsymbol{\beta}) \neq \ell\left(z_t(\boldsymbol{\phi}^i); \tilde{\boldsymbol{\beta}}\right)$ for all $\boldsymbol{\beta} \neq \tilde{\boldsymbol{\beta}}$. Condition **1)** is satisfied by the restrictions on $\boldsymbol{\Phi}$ and (**A4**). Condition **2**) is satisfied by the uniqueness of the Kalman recursion for this particular model.

The result now follows if our prior $p(\boldsymbol{\phi})$ places non-zero probability on $\boldsymbol{\phi}_0$ and if the one-to-one condition in Assumption **I2** is satisfied. Any $p(\boldsymbol{\phi})$ that places positive probability on $[\varphi, M]^2$ satisfies Assumption **I1**.

To verify Assumption **I2** first define:

$$h(\phi) = \begin{pmatrix} f(\kappa) \\ g(\kappa, \sigma) \end{pmatrix} = \begin{pmatrix} e^{-\kappa} \\ \sigma\sqrt{\frac{1-e^{-2\kappa}}{-2\log(e^{-\kappa})}} \end{pmatrix},$$

where, by construction, $\boldsymbol{\beta} = h(\boldsymbol{\phi})$, with $h(\cdot)$ one-to-one. Hence, by the invariance property of the maximum likelihood principle, $\widehat{\boldsymbol{\beta}}(\mathbf{y}) = h\left(\widehat{\boldsymbol{\phi}}(\mathbf{y})\right)$, where

$$\widehat{\boldsymbol{\phi}}(\mathbf{y}) := \arg\max_{\phi \in \boldsymbol{\Phi}} L_a(\mathbf{y}; h^{-1}(\boldsymbol{\beta})) \equiv \arg\max_{\phi \in \boldsymbol{\Phi}} \bar{L}_a(\mathbf{y}; \boldsymbol{\phi})$$

and the result follows.

A.3 SV-SQ Example

A.3.1 Detailed Implementation of the AUKF

Given the assumed invariance (over time) of both e_t and v_t in (31) and (32) respectively, the sigma points needed to implement the AUKF are determined as:

$$e^1 = E(e_t);\ e^2 = E(e_t) + a_e\sqrt{var(e_t)};\ e^3 = E(e_t) - b_e\sqrt{var(e_t)}$$

and

$$v^1 = E(v_t);\ v^2 = E(v_t) + a_v\sqrt{var(v_t)};\ v^3 = E(v_t) - b_v\sqrt{var(v_t)}$$

respectively, and propagated at each t through the relevant non-linear transformations, $h_t(.)$ and $k_t(.)$. The values a_e, b_e, a_v and b_v are chosen according to the assumed distribution of e_t and v_t, with a Gaussian assumption for both variables yielding values of $a_e = b_e = a_v = b_v = \sqrt{3}$ as being 'optimal'. Different choices of these values are used to reflect higher-order distributional information and thereby improve the accuracy with which the mean and variance of the non-linear transformations are estimated; see Julier *et al.* (1995; 2000) for more details. Restricted supports are also managed via appropriate truncation of the sigma points. The same principles are applied to produce the mean and variance of the time varying state x_t, except that the sigma points need to be recalculated at each time t to reflect the up-dated mean and variance of x_t as each new value of y_t is realized.

In summary, the steps of the AUKF applied to evaluate the likelihood function of (31) and (32) are as follows:

1. Use the (assumed) marginal mean and variance of x_t, along with the invariant mean and variance of v_t and e_t respectively, to create the (3×7) matrix of augmented sigma points for $t = 0$, $X_{a,0}$, as follows. Define:

$$E(X_{a,0}) = \begin{bmatrix} E(x_t) \\ E(v_t) \\ E(e_t) \end{bmatrix},\ P_{a,0} = \begin{bmatrix} var(x_t) & 0 & 0 \\ 0 & var(v_t) & 0 \\ 0 & 0 & var(e_t) \end{bmatrix}, \tag{45}$$

and $\sqrt{P_{a,0}}_j$ as the jth column of the Cholesky decomposition (say) of $P_{a,0}$. Given the diagonal form of $P_{a,0}$ (in this case), we have

$$\sqrt{P_{a,0}}_1 = \begin{bmatrix} \sqrt{var(x_t)} \\ 0 \\ 0 \end{bmatrix};\ \sqrt{P_{a,0}}_2 = \begin{bmatrix} 0 \\ \sqrt{var(v_t)} \\ 0 \end{bmatrix};\ \sqrt{P_{a,0}}_3 = \begin{bmatrix} 0 \\ 0 \\ \sqrt{var(e_t)} \end{bmatrix}.$$

The seven columns of $X_{a,0}$ are then generated by

$$E(X_{a,0});\ E(X_{a,0}) + a_j\sqrt{P_{a,0}}_j \text{ ; for } j = 1,2,3;\ E(X_{a,0}) - b_j\sqrt{P_{a,0}}_j \text{ ; for } j = 1,2,3,$$

where $a_1 = a_x$, $a_2 = a_v$ and $a_3 = a_e$, and the corresponding notation is used for b_j, $j = 1,2,3$.

2. Propagate the $t = 0$ sigma points through the transition equation as $X_{x,1} = k_1(X_{a,0}, \boldsymbol{\beta})$ and estimate the predictive mean and variance of x_1 as:

$$E(x_1|y_0) = \sum_{i=1}^{7} w_i X_{x,1}^i \tag{46}$$

$$var(x_1|y_0) = \sum_{i=1}^{7} w_i (X_{x,1}^i - E(x_1|y_0))^2, \tag{47}$$

where $X_{x,1}^i$ denotes the ith element of the (1×7) vector $X_{x,1}$ and w_i the associated weight, determined as an appropriate function of the a_j and b_j; see Ponomareva and Date (2010).

3. Produce a new matrix of sigma points, $X_{a,1}$, for $t = 1$ generated by

$$E(X_{a,1});\ E(X_{a,1}) + a_j\sqrt{P_{a,1}}_j \text{ ; for } j = 1,2,3;\ E(X_{a,1}) - b_j\sqrt{P_{a,1}}_j \text{ ; for } j = 1,2,3, \tag{48}$$

using the updated formulae for the mean and variance of x_t from (46) and (47) respectively, in the calculation of $E(X_{a,1})$ and $P_{a,1}$.

4. Propagate the $t = 1$ sigma points through the measurement equation as $X_{y,1} = h_1(X_{a,1}, \boldsymbol{\beta})$ and estimate the predictive mean and variance of y_1 as:

$$E(y_1|y_0) = \sum_{i=1}^{7} w_i X_{y,1}^i \tag{49}$$

$$var(y_1|y_0) = \sum_{i=1}^{7} w_i (X_{y,1}^i - E(y_1|y_0))^2, \tag{50}$$

where $X_{y,1}^i$ denotes the ith element of the (1×7) vector $X_{y,1}$ and w_i is as defined in Step 3.

5. Estimate the first component of the likelihood function, $p(y_1|y_0)$, as a Gaussian distribution with mean and variance as given in (49) and (50) respectively.

6. Given observation y_1 produce the up-dated filtered mean and variance of x_t via the usual KF up-dating equations:

$$\begin{aligned} E(x_1|y_1) &= E(x_1|y_0) + M_1(y_1 - E(y_1|y_0)) \\ var(x_1|y_1) &= var(x_1|y_0) - M_1^2 var(y_1|y_0), \end{aligned}$$

where:

$$M_1 = \frac{\sum_{i=1}^{7} w_i (X_{x,1}^i - E(x_1|y_0))(X_{y,1}^i - E(y_1|y_0))}{var(y_1|y_0)}$$

and the $X_{x,1}^i$, $i = 1, 2, ..., 7$ are as computed in Step 3.

7. Continue as for Steps 2 to 6, with the obvious up-dating of the time periods and the associated indexing of the random variables and sigma points, and with the likelihood function evaluated as the product of the components produced in each implementation of Step 5, and the log-likelihood produced accordingly.

A.3.2 Proof of Corollary 3

Define $X_{a,0}$ to be the (3×7) matrix of initial sigma-points, as referenced in Appendix A.3.1, where $X_{a,0}(j,i)$ is the element in the j-th row and i-th column of $X_{a,0}$. For fixed weights $\{w_i\}_{i=1}^7$, with $\sum_{i=1}^7 w_i = 1,\ w_i > 0,\ i = 1,..,7$, we initialize the system by propagating $X_{a,0}$ through the state equation (32). To build the auxiliary likelihood using the AUKF, then define the predicted mean and variance of x_t as

$$\begin{aligned}
\hat{x}_{t|t-1} &= \sum_{i=1}^{7} w_i k\left(X_{a,0}(1,i), X_{a,0}(2,i), \boldsymbol{\beta}\right), \\
P^x_{t|t-1} &= \sum_{i=1}^{7} w_i \left[k\left(X_{a,0}(1,i), X_{a,0}(2,i), \boldsymbol{\beta}\right) - \hat{x}_{t|t-1}\right]^2, \\
k\left(X_{a,0}(1,i), X_{a,0}(2,i), \boldsymbol{\beta}\right) &= \beta_1 + \beta_2 X_{a,0}(1,i) + \beta_3 \left(\sqrt{X_{a,0}(1,i)}\right) X_{a,0}(2,i),
\end{aligned}$$

where the sigma points reflect the positivity of the variance. From $\hat{x}_{t|t-1}$ and $P^x_{t|t-1}$ the new matrix of sigma points is produced, $X_{a,1}$, following (48). Define the predicted mean and variance for the observed y_t, based on the $X_{a,t-1}$ matrix of sigma-points, as

$$\begin{aligned}
\hat{y}_{t|t-1} &= \sum_{i=1}^{7} w_i h\left(X_{a,t-1}(1,i), X_{a,t-1}(3,i), \boldsymbol{\beta}\right), \\
P^y_{t|t-1} &= \sum_{i=1}^{7} w_i \left[h\left(X_{a,t-1}(1,i), X_{a,t-1}(3,i), \boldsymbol{\beta}\right) - \hat{y}_{t|t-1}\right]^2, \\
h\left(X_{a,t-1}(1,i), X_{a,t-1}(3,i), \boldsymbol{\beta}\right) &= \log\left(X_{a,t-1}(1,i)\right) + X_{a,t-1}(3,i).
\end{aligned}$$

For $\zeta_t = y_t - \hat{y}_{t|t-1}$, the augmented Kalman filtering steps for x_t are as follows:

$$\begin{aligned}
\hat{x}_{t|t} &= \hat{x}_{t|t-1} + M_{t|t}\zeta_t, \\
P^x_{t|t} &= P^x_{t|t-1} - M^2_{t|t} P^y_{t|t-1} \\
M_{t|t} &= \frac{\sum_{i=1}^7 w_i \left[k\left(X_{a,t1-}(1,i), X_{a,t-1}(2,i), \boldsymbol{\beta}\right) - \hat{x}_{t|t-1}\right]\left[h\left(X_{a,t-1}(1,i), X_{a,t-1}(3,i), \boldsymbol{\beta}\right) - \hat{y}_{t|t-1}\right]}{P^y_{t|t-1}}.
\end{aligned}$$

In accordance with the AUKF algorithm in the Appendix A.3.1, and noting the structure of the approximating model in (31) and (32), the time-t matrix of sigma points $X_{a,t}$ is given by

$$X_{a,t} = \begin{pmatrix} \hat{x}_{t|t} & \hat{x}_{t|t}+a_1\sqrt{P^x_{t|t}} & \hat{x}_{t|t} & \hat{x}_{t|t} & \hat{x}_{t|t}-b_1\sqrt{P^x_{t|t}} & \hat{x}_{t|t} & \hat{x}_{t|t} \\ \lambda^* & \lambda^* & \lambda^*+a_2\sqrt{var(v_t)} & \lambda^* & \lambda^* & \lambda^*-b_2\sqrt{var(v_t)} & \lambda^* \\ \gamma^* & \gamma^* & \gamma^* & \gamma^*+a_3\sqrt{\pi^2/2} & \gamma^* & \gamma^* & \gamma^*-b_3\sqrt{\pi^2/2} \end{pmatrix}, \quad (51)$$

where $a_1 = a_2 = a_3 = b_1 = b_2 = b_3 = \sqrt{3}$, $\gamma^* = -1.27$,

$$\lambda^* = \frac{\phi\left(\frac{-\beta_1}{\beta_3}\right)}{1 - \Phi\left(\frac{-\beta_1}{\beta_3}\right)} \text{ and } var(v_t) = \left[1 - \lambda\left(\frac{-\beta_1}{\beta_3}\right)\left(\lambda\left(\frac{-\beta_1}{\beta_3}\right) - \frac{-\beta_1}{\beta_3}\right)\right]. \tag{52}$$

Using the definitions $\hat{y}_{t|t-1}$, $P^y_{t|t-1}$, ζ_t, the auxiliary log-likelihood is given by

$$L_a(\mathbf{y}; \boldsymbol{\beta}) = \sum_{t=2}^{T} \ell(y_t|y_{t-1}; \boldsymbol{\beta}) = -\sum_{t=2}^{T} \ln(P^y_{t|t-1}) + \frac{1}{2}\frac{\zeta_t^2}{P^y_{t|t-1}},$$

where a Gaussian approximation for the components of the likelihood function is adopted at this point.

In what follows, denote the simulated versions of $\hat{x}_{t|t}$ and $P^x_{t|t}$ based on simulated data by $\hat{x}_{t|t}(\boldsymbol{\phi}^i)$ and $P^x_{t|t}(\boldsymbol{\phi}^i)$, where $\hat{x}_{t|t}(\boldsymbol{\phi}^i)$ and $P^x_{t|t}(\boldsymbol{\phi}^i)$ have the same definition as $\hat{x}_{t|t}$ and $P^x_{t|t}$ but with y_t replaced by z^i_t. Likewise, let $\hat{y}_{t|t-1}(\boldsymbol{\phi}^i)$, $P^y_{t|t-1}(\boldsymbol{\phi}^i)$ denote the simulated counterparts of $\hat{y}_{t|t-1}$ and $P^y_{t|t-1}$.

From these definitions we can now verify Assumptions **(A2)-(A5)** (Assumption **(A1)** can be verified by inspection) for the SV-SQ model in (29)-(30).

(A2) The stated restrictions on $\boldsymbol{\Phi}$ guarantee the satisfaction of **(A2)**.

(A3) Continuity of $(\boldsymbol{\phi}^i, \boldsymbol{\beta}) \mapsto L_a(\mathbf{z}(\boldsymbol{\phi}^i); \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ follows from the definitions of $\hat{x}_{t|t}$, $P^x_{t|t}$ and $L_a(\mathbf{z}(\boldsymbol{\phi}^i); \boldsymbol{\beta})$. Continuity with respect to $\boldsymbol{\phi}$ follows from continuity, in $\boldsymbol{\phi}$, of $z_t(\boldsymbol{\phi}^i)$, $\hat{x}_{t|t}(\boldsymbol{\phi}^i)$, $P^x_{t|t}(\boldsymbol{\phi}^i)$, and the definition of $L_a(\mathbf{z}(\boldsymbol{\phi}^i); \boldsymbol{\beta})$.

(A4) Assumption **(A4)** follows from the stated restrictions on $\mathbf{B}$ and $\boldsymbol{\Phi}$.

(A5) Primitive conditions guaranteeing **(A5)** are as follows: 1) $E_{\boldsymbol{\phi}^i}[|\ell(z_t(\boldsymbol{\phi}^i), \boldsymbol{\beta})|] < \infty$ for all $\boldsymbol{\beta} \in \mathbf{B}$; 2) $\ell(z_t(\boldsymbol{\phi}^i); \boldsymbol{\beta}) \neq \ell\left(z_t(\boldsymbol{\phi}^i); \tilde{\boldsymbol{\beta}}\right)$ for all $\boldsymbol{\beta} \neq \tilde{\boldsymbol{\beta}}$. Condition 1) is satisfied by the restrictions on $\boldsymbol{\Phi}$ and $\mathbf{B}$. For Condition 2) to be satisfied, the AUKF recursions must be unique in $\boldsymbol{\beta}$. Uniqueness of the AUKF recursions requires that the matrix of sigma-points be unique in $\boldsymbol{\beta}$ for each $t \geq 1$. Denote by $X_{a,t}(\boldsymbol{\beta})$ the (3×7) matrix of sigma-points in (51) constructed for a given $\boldsymbol{\beta}$. Focusing on the elements of $X_{a,t}(\boldsymbol{\beta})$ due to $\hat{x}_{t|t}$, by the Kalman recursions for this model, $\hat{x}_{t|t}$ is a unique function of $\boldsymbol{\beta}$ and so $X_{a,t}(\boldsymbol{\beta}) \neq X_{a,t}(\tilde{\boldsymbol{\beta}})$ if $\boldsymbol{\beta} \neq \tilde{\boldsymbol{\beta}}$, and the result follows.[8]

[8] We focus on the elements within the Kalman recursion portion of $X_{a,t}(\boldsymbol{\beta})$, since $\lambda^*(\boldsymbol{\beta})$ is not one-to-one in the parameters β_1, β_3 and so there exists $\tilde{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$ such that $\lambda^*(\boldsymbol{\beta}) = \lambda^*(\tilde{\boldsymbol{\beta}})$.