MONASH University

MONASH BUSINESS SCHOOL

DEPARTMENT OF ECONOMETRICS AND BUSINESS STATISTICS

ISSN 1440-771X

WORKING PAPER SERIES

Variational Bayes in State Space Models: Inferential and Predictive Accuracy

David T. Frazier
Ruben Loaiza-Maya
Gael M. Martin

Working Paper No. 01/22
FEBRUARY 2022

Variational Bayes in State Space Models: Inferential and Predictive Accuracy

David T. Frazier, Rubén Loaiza-Maya and Gael M. Martin*
Department of Econometrics and Business Statistics, Monash University

February 24, 2022

Abstract

Using theoretical and numerical results, we document the accuracy of commonly applied variational Bayes methods across a range of state space models. The results demonstrate that, in terms of accuracy on fixed parameters, there is a clear hierarchy in terms of the methods, with approaches that do not approximate the states yielding superior accuracy over methods that do. We also document numerically that the inferential discrepancies between the various methods often yield only small discrepancies in predictive accuracy over small out-of-sample evaluation periods. Nevertheless, in certain settings, these predictive discrepancies can become meaningful over a longer out-of-sample period. This finding indicates that the invariance of predictive results to inferential inaccuracy, which has been an oft-touted point made by practitioners seeking to justify the use of variational inference, is not ubiquitous and must be assessed on a case-by-case basis.

Keywords: State space models; Variational inference; Probabilistic forecasting; Bayesian consistency; Scoring rules.

1 Introduction

A common class of models used for time series modelling and prediction is the class of state space models (SSMs). This class includes nonlinear structures, like stochastic volatility models, regime switching models, mixture models, and models with random dynamic jumps; plus linear structures, such as linear Gaussian unobserved component models. (See

*The authors gratefully acknowledge funding from the Australian Research Council.

Durbin and Koopman, 2001, Harvey *et al.*, 2004, and Giordani *et al.*, 2011, for extensive reviews).

The key feature of SSMs is their dependence on hidden, or latent, 'local' variables, or states, which govern the dependence of the observed data, in conjunction with a vector of unknown 'global' parameters. This feature leads to inferential challenges with, for example, the likelihood function for the global parameters being analytically unavailable, except in special cases. Whilst frequentist methods have certainly been adopted (see Danielsson and Richard, 1993, Ruiz, 1994, Andersen and Srensen, 1996, Gallant and Tauchen, 1996, Sandmann and Koopman, 1998, Bates, 2006, Ait-Sahalia and Kimmel, 2007, and Aït-Sahalia *et al.*, 2021, amongst others), it is arguable that Bayesian Markov chain Monte Carlo (MCMC) methods have become the most common tool for analysing general SSMs, with such techniques expanded in more recent times to accommodate particle filtering, via pseudo-marginal variants such as particle MCMC (PMCMC) (Andrieu *et al.*, 2011; Flury and Shephard, 2011). See Giordani *et al.* (2011) and Fearnhead (2011) for a detailed coverage of this literature, including the variety of MCMC-based algorithms adopted therein.

Whilst (P)MCMC methods have been transformative in the SSM field, they do suffer from certain well-known limitations. Most notably, they require that either the (complete) likelihood function is available in closed form or that an unbiased estimator of it is available. Such methods also do not necessarily scale well to high-dimensional problems; that is, to models with multiple observed and/or state processes. If the assumed data generating process (DGP) is intractable, inference can proceed using approximate Bayesian computation (ABC) (Dean *et al.*, 2014; Creel and Kristensen, 2015; Martin *et al.*, 2019), since ABC requires only simulation - not evaluation - of the DGP. However, ABC also does not scale well to problems with a large number of parameters (see, e.g., Corollary 1 in Frazier *et al.*, 2018 for details).

Variational Bayes (VB) methods (see Blei *et al.*, 2017 for a review) can be seen as a potential class of alternatives to either (P)MCMC- or ABC-based inference in SSMs. In particular, and in contrast to these methods, VB scales well to high-dimensional problems, using optimization-based techniques to effectively manage a large number of unknowns

(Tran *et al.*, 2017; Quiroz *et al.*, 2018; Koop and Korobilis, 2020; Chan and Yu, 2020; Loaiza-Maya *et al.*, 2021).

In this paper, we make three contributions to the literature on the application of VB to SSMs. The first contribution is to highlight the fundamental issue that lies at the heart of the use of VB in an SSM setting, linking this to an existing issue identified in the literature as the 'incidental parameter problem' (Neyman and Scott, 1948; Lancaster, 2000; Westling and McCormick, 2019). In brief, without due care, the application of VB to the local parameters in an SSM leads to a lack of Bayesian consistency for the global parameters. Moreover, in a class of common SSMs, we demonstrate analytically the impact of this inconsistency on the resulting state inference, and show that even in idealized settings inconsistent inference for the global parameters can lead to highly inaccurate inferences about the local parameters. The second contribution is to review some existing variational methods, and to link their prospects for consistency to the manner in which they do, or do not, circumvent the incidental parameter problem. Thirdly, we undertake a numerical comparison of several competing variational methods, in terms of both inferential and predictive accuracy. The key findings are that: i) correct management of the local variables leads to inferential accuracy that closely matches that of exact (MCMC-based) Bayes; ii) inadequate treatment of the local variables leads, in contrast, to noticeably less accurate inference; iii) predictive accuracy shows some robustness to inferential inaccuracy, but only for small sample sizes. Once the size of the sample is very large, the consistency (or otherwise) of a VB method impinges on predictive accuracy, with a clear ranking becoming evident across the methods for some DGPs; with certain VB methods unable to produce similar out-of-sample accuracy results to exact Bayes in some settings.

We believe that all three contributions serve as novel insights into the role of VB in SSMs, which may lead to best practice, if heeded.

Throughout the remainder, we make use of the following notational conventions. Generic p, g are used to denote densities, and π is used to denote posteriors conditioned only on data, and where the conditioning is made explicit depending on the situation. For any arbitrary collection of data $(z_1, \ldots, z_n)$, we abbreviate this collection as z_1^n. For a sequence a_n,

the terms $O_p(a_n)$, $o_p(a_n)$ and $\to_p$ have their usual meaning. Similarly, we let $\text{plim}_n X_n = c$ denote $X_n \to_p c$. We let $d(\cdot,\cdot)$ denote a metric on $\Theta \subseteq \mathbb{R}^{d_\theta}$. The proofs of all theoretical results, certain definitions, plus additional tables and figures, are included in the Supplementary Appendix.

2 State space models: exact inference

An SSM is a stochastic process consisting of the pair $\{(X_t, Y_t)\}$, where $\{X_t\}$ is a Markov chain taking values in the measurable space $(\mathcal{X}, \mathcal{F}_X, \mu)$, and $\{Y_t\}$ is a process taking values in a measure space $(\mathcal{Y}, \mathcal{F}_Y, \chi)$, such that, conditional on $\{X_t\}$, the sequence $\{Y_t\}$ is independent. The model is formulated through the following conditional and transition densities: for a vector of unknown random parameters θ taking values in the probability space $(\Theta, \mathcal{F}_\theta, P_\theta)$, where P_θ admits the density function p_θ,

$$Y_t|X_t,\theta \sim g_\theta(y_t|x_t) \tag{1}$$

$$X_{t+1}|X_t,\theta \sim \chi_\theta(x_{t+1},x_t), \tag{2}$$

where $\chi_\theta(\cdot,\cdot)$ denotes the transition kernel with respect to the measure μ. For simplicity, throughout the remainder we disregard the terms dependence on the initial measure ν and the invariant measure μ, when no confusion will result. The order-one Markov assumption for X_t is innocuous, and any finite (and known) Markov order can be accommodated via a redefinition of the state variables.

Given the independence of Y_t conditional on X_t, and the Markovian nature of $X_t|X_{t-1}$, the complete data likelihood is

$$p_\theta(y_1^n, x_1^n) = \nu(x_1) g_\theta(y_1|x_1) \prod_{t=2}^{n} \chi_\theta(x_t, x_{t-1}) g_\theta(y_t|x_t).$$

The (average) observed data log-likelihood is thus

$$\ell_n(\theta) := \frac{1}{n} \log p_\theta(y_1^n) = \frac{1}{n} \log \int p_\theta(y_1^n, x_1^n) dx_1^n, \tag{3}$$

and the maximum likelihood estimator (MLE) of θ is $\widehat{\theta}_n^{MLE} := \text{argmax}_{\theta\in\Theta}\, \ell_n(\theta)$. As is standard knowledge, $\ell_n(\theta)$ is available in closed form only for particular forms of $g_\theta(y_t|x_t)$

and $\chi_\theta(x_{t+1}, x_t)$; the canonical example being when (2) and (1) define a linear Gaussian state space model (LGSSM). Similarly, for $p(\theta)$ denoting the prior density, the exact (marginal) posterior for θ, defined as

$$\pi(\theta|y_1^n) = \int \pi(\theta, x_1^n|y_1^n)\mathrm{d}x_1^n, \text{ where } \pi(\theta, x_1^n|y_1^n) \propto p(y_1^n|x_1^n, \theta)p(x_1^n|\theta)p(\theta), \tag{4}$$

is available (e.g. via straightforward MCMC methods) only in limited cases, the LGSSM being one such case. In more complex settings and/or settings where either θ or $\{(X_t, Y_t)\}$, or both, are high-dimensional, accessing (4) can be difficult, with standard MCMC methods leading to slow mixing, and thus potentially unreliable inferences (Betancourt, 2018).

To circumvent these issues, recent research has suggested the use of variational methods for SSMs: these methods can be used to approximate either the log-likelihood function in (3) or the marginal posterior in (4), depending on the mode of inference being adopted. The focus of this paper, as already highlighted, is on variational *Bayes* and, in particular, on the accuracy of such methods in SSMs. However, as part of the following section we also demonstrate the asymptotic behaviour of frequentist variational point estimators of θ, as this result will ultimately help us interpret the behavior of the variational posterior in SSMs.

3 State space models: variational inference

3.1 Overview

The idea of VB is to produce an approximation to the joint posterior $\pi(x_1^n, \theta|y_1^n)$ in (4) by searching over a given family of distributions for the member that minimizes a user-chosen divergence measure between the posterior of interest and the family. This replaces the posterior sampling problem with one of optimization over the family of densities used to implement the approximation. We now review the use of variational methods in SSMs, paying particular attention to the Markovian nature of the states.

VB approximates the posterior $\pi(x_1^n, \theta|y_1^n)$ by minimizing the KL divergence between a

family of densities $\mathcal{Q}$, with generic element $q(x_1^n, \theta)$, and π:

$$\mathrm{KL}(q||\pi) = \int q(x_1^n, \theta) \log \frac{q(x_1^n, \theta)}{\pi(x_1^n, \theta|y_1^n)} \mathrm{d}x_1^n \mathrm{d}\theta. \tag{5}$$

Optimizing the KL divergence directly is not feasible since it depends on the unknown $\pi(x_1^n, \theta|y_1^n)$; the very quantity we are trying to approximate. However, minimizing the KL divergence between q and π is equivalent to maximizing the so-called variational evidence lower bound (ELBO):

$$\mathrm{ELBO}(q||\pi) := \int q(x_1^n, \theta) \log \frac{p(y_1^n|x_1^n, \theta)p(x_1^n|\theta)p(\theta)}{q(x_1^n, \theta)} \mathrm{d}x_1^n \mathrm{d}\theta, \tag{6}$$

which we can access. Hence, for a given class $\mathcal{Q}$, we may define the variational approximation as

$$\widehat{q} := \underset{q \in \mathcal{Q}}{\mathrm{argmax}}\ \mathrm{ELBO}(q||\pi).$$

The standard approach to obtaining $\widehat{q}$ is to consider a class of product distributions

$$\mathcal{Q} = \{q : q(x_1^n, \theta) = q_\theta(\theta) q_x(x_1^n|\theta)\},$$

with Q often restricted to be mean-field, i.e., θ_i independent of θ_j, $i \neq j$, and x_1^n independent of θ.

Regardless of the variational family adopted, $\mathrm{KL}(q||\pi)$, and hence $\mathrm{ELBO}(q||\pi)$, involve both θ and x_1^n. The product form of $\mathcal{Q}$ allows us to write:

$$\begin{aligned}
\mathrm{ELBO}(q||\pi) &= \int_\Theta \int_\mathcal{X} q_\theta(\theta) q_x(x_1^n|\theta) \log \frac{p(y_1^n|x_1^n, \theta)p(x_1^n|\theta)p(\theta)}{q_\theta(\theta) q_x(x_1^n|\theta)} \mathrm{d}x_1^n \mathrm{d}\theta \\
&= \int_\Theta \int_\mathcal{X} q_\theta(\theta) q_x(x_1^n|\theta) \log \left[\frac{p(y_1^n|x_1^n, \theta)p(x_1^n|\theta)}{q_x(x_1^n|\theta)} \right] \mathrm{d}x_1^n \mathrm{d}\theta - \mathrm{KL}[q_\theta(\theta)||p(\theta)],
\end{aligned}$$

where the last line follows from Fubini's theorem and the fact that $q_x(x_1^n|\theta)$, by assumption, is a proper density function, for all θ. Further, defining

$$\mathcal{L}_n(\theta) := \int_\mathcal{X} q_x(x_1^n|\theta) \log \frac{p(y_1^n|x_1^n, \theta)p(x_1^n|\theta)}{q_x(x_1^n|\theta)} \mathrm{d}x_1^n, \tag{7}$$

by Jensen's inequality

$$\log p_\theta(y_1^n) = \log \int_\mathcal{X} p_\theta(y_1^n, x_1^n) dx_1^n = \log \int_\mathcal{X} q_x(x_1^n|\theta) \left\{ \frac{p(y_1^n|x_1^n, \theta)p(x_1^n|\theta)}{q_x(x_1^n|\theta)} \right\} \mathrm{d}x_1^n \geq \mathcal{L}_n(\theta).$$

Thus $\mathcal{L}_n(\theta)$ can be viewed as an approximation (from below) to the observed data log-likelihood. Defining

$$\Upsilon_n(q) := \int_\Theta \{\log p_\theta(y_1^n) - \mathcal{L}_n(\theta)\} q_\theta(\theta) \mathrm{d}\theta, \tag{8}$$

the ELBO$(q||\pi)$ can then be expressed as

$$\mathrm{ELBO}(q||\pi) = \int_\Theta \log p_\theta(y_1^n) q_\theta(\theta) \mathrm{d}\theta - \Upsilon_n(q) - \mathrm{KL}[q_\theta(\theta)||p(\theta)]. \tag{9}$$

This representation decomposes ELBO$(q||\pi)$ into three components, two of which only depend on the variational approximation of the global parameters θ, and a third component, $\Upsilon_n(q)$, that Yang *et al.* (2020) refer to as the average (with respect to $q_\theta(\theta)$) "Jensen's gap", which encapsulates the error introduced by approximating the latent states using a given variational class. While the first and last term in the decomposition can easily be controlled by choosing an appropriate class for $q_\theta(\theta)$, it is the average Jensen's gap that ultimately determines the behavior of the variational approximation.

3.2 Consistency of variational point estimators

The decomposition in (9) has specific implications for variational inference in SSMs, which can be most readily seen by first considering the case where we only employ a variational approximation for the states, and consider point estimation of the parameters θ. In this case, we can think of the variational family as $\mathcal{Q} := \{q : q(\theta, x_1^n) = \delta_\theta \times q_x(x_1^n|\theta)\}$, where δ_θ is the Dirac delta function at θ, and we can then write

$$\frac{1}{n}\mathrm{ELBO}(\theta \times q_x||\pi) = \ell_n(\theta) - \frac{1}{n}\Upsilon_n(\theta, q_x) + \frac{1}{n}\log p(\theta),$$

where we abuse notation and represent functions with arguments δ_θ only by the parameter value $\theta \in \Theta$, and also make use of the short-hand notation q_x for $q_x(x_1^n|\theta)$. Define the variational point estimator as

$$(\widehat{\theta}_n, \widehat{q}_x) := \underset{\theta\in\Theta, \mathcal{Q}}{\mathrm{argmax}} \ \frac{1}{n}\mathrm{ELBO}(\theta \times q_x||\pi).$$

At a minimum, we would hope that the variational estimator $\widehat{\theta}_n$ converges to the same point as the MLE. To deduce the behavior of $\widehat{\theta}_n$, we employ the following high-level regularity conditions.

Assumption 3.1. *(i) The parameter space Θ is compact, and $0 < p(\theta) < \infty$. (ii) There exists a deterministic function $H(\theta)$, continuous for all $\theta \in \Theta$, and such that $\sup_{\theta\in\Theta} |H(\theta) - \ell_n(\theta)| = o_p(1)$. (iii) For some value $\theta_0 \in \Theta$, for all $\epsilon > 0$, there exists a $\delta > 0$ such that $H(\theta_0) \geq \sup_{d(\theta,\theta_0)>\delta} H(\theta) + \delta$.*

Low level regularity conditions that imply Assumption 3.1 are given in Douc *et al.* (2011). Since the main thrust of this paper is to deduce the accuracy of variational methods in SSMs, and not to focus on the technical details of the SSMs in particular, we make use of high-level conditions to simplify the exposition and reduce necessary technicalities that may otherwise obfuscate the main point.

The following result shows that consistency of $\widehat{\theta}_n$ (for θ_0) is guaranteed if the variational family for the states is 'good enough'.

Lemma 3.1. *Define $\kappa_n := \frac{1}{n} \inf_{q_x\in\mathcal{Q}_x} \Upsilon_n(\theta_0, q_x)$, and note that $\kappa_n \geq 0$. If Assumption 3.1 is satisfied, and if $\kappa_n = o_p(1)$, then $\widehat{\theta}_n \to_p \theta_0$.*

The above result demonstrates that for the variational point estimator $\widehat{\theta}_n$ to be consistent, the (infeasible) average Jensen's gap must converge to zero. Intuitively, this requires that the error introduced by approximating the states grows more slowly than the rate at which information accumulates in our observed sample, i.e., n. The condition $\kappa_n = o_p(1)$ is stated at the true value, θ_0, rather than at the estimated value, as it will often be easier to deduce satisfaction of the condition, or otherwise, at convenient points in the parameter space.

As the following example illustrates, even in the simplest SSMs, the scaled (average) Jensen's gap need not vanish in the limit, and can ultimately pollute the resulting inference on θ_0.

Example 3.1 (Linear Gaussian model)**.** Consider the following SSM,

$$X_{t+1} = \rho X_t + \sigma_0 \epsilon_t, \quad X_1 \sim \mathcal{N}\left(0, \sigma_0^2\right), \quad Y_t = \alpha X_t + \sigma_0 \eta_t,$$

with $\{\epsilon_t\}$ and $\{\eta_t\}$ independent sequences of i.i.d. standard normal random variables. We observe a sequence $\{Y_t\}$ from the above model, but the states $\{X_t\}$ are unobserved. Furthermore, consider that $\theta = (\rho, \alpha)'$ is unknown while σ_0 is known.

We make use of the autoregressive nature of the state process to approximate the posterior for $\pi(x_1^n|\theta, y_1^n)$ via the variational family: $\mathcal{Q}_x := \{\lambda \in [0,1) : q_\lambda(x_1^n|\sigma_0) = \mathcal{N}[x_1^n; 0, \nu(\lambda)\Phi_n(\lambda)]\}$, where $\nu(\lambda) = \frac{\sigma_0^2}{(1-\lambda^2)}$ and,

$$\Phi_n(\lambda) = \begin{pmatrix} 1 & \lambda & \lambda^2 & \cdots & \lambda^{n-1} \\ \lambda & 1 & \lambda & \cdots & \lambda^{n-2} \\ \lambda^2 & \lambda & 1 & \cdots & \lambda^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda^{n-1} & \lambda^{n-2} & \lambda^{n-3} & \cdots & 1 \end{pmatrix}$$

When evaluated at $\lambda = \rho_0$, Q_x is the actual (infeasible) joint distribution of the states, and thus should provide a reasonable approximation to the state posterior.

Lemma 3.2. *Let $\sigma_0 > 0$ and $0 \leq |\rho_0| < 1$, $0 \leq |\alpha_0| < M$. Assume the variational parameter defining $\mathcal{Q}_x$ is fixed at $\lambda = \rho_0$. (i) If $\rho_0 = 0$ and known, then the variational point estimator $\widehat{\alpha}$ is consistent if and only if $\alpha_0 = 0$. (ii) If $\alpha_0 = 0$ and known, then the variational point estimator for $\widehat{\rho}$ is consistent if and only if $\rho_0 = 0$.*

Lemma 3.2 demonstrates that even in this simplest of SSMs, variational inference is inconsistent in anything other than the most vacuous cases. In short, so long as there is weak dependence in states the estimator of α is inconsistent; alternatively, if there is no relationship between Y_t and X_t, i.e., $\alpha_0 = 0$, then the only way in general to obtain consistent inference for ρ_0 is if $\rho_0 = 0$!

3.3 Lack of Bayes consistency of the variational posterior

While the above results pertain to variational point estimators of θ_0, a similar result can be stated in terms of the so-called 'idealized' variational posterior. To state this result, we approximate the state posterior using the class of variational approximations,

$$q_x(x_1^n) := q_\lambda(x_1^n),$$

where $\lambda \in \Lambda$ denotes the vector of so-called 'variational parameters' that characterize the elements in $\mathcal{Q}$. With reference to (7), making the dependence of $q_\lambda(x_1^n)$ on the variational

parameter λ explicit leads to the criterion $L_n(\theta, \lambda)$, where $q_x(x_1^n)$ in (7) is replaced by $q_\lambda(x_1^n)$. Optimizing over λ for fixed θ yields the profiled criterion,

$$\widehat{L}_n(\theta) := \mathcal{L}_n[\theta, \widehat{\lambda}_n(\theta)] \equiv \sup_{\lambda \in \Lambda} \mathcal{L}_n(\theta, \lambda), \tag{10}$$

and the 'idealized' variational posterior for θ,

$$\widehat{q}(\theta | y_1^n) \propto \exp\left\{\widehat{L}_n(\theta)\right\} p(\theta).$$

We remark that, unlike with the frequentist optimization problem, the idealized VB posterior incorporates a component of Jensen's gap directly into the definition of that posterior. A sufficient condition for the 'VB ideal' to concentrate onto θ_0 is that θ_0 is the maximum of a well-defined limit counterpart to $\widehat{L}_n(\theta)$. However, there is no reason to suspect this is the case a priori.

The 'idealized' variational posterior $\widehat{q}(\theta | y_1^n)$ is a generalized posterior, in the sense of Bissiri *et al.* (2016), based on the profiled criterion function $\widehat{L}_n(\theta)$. Given that $\widehat{q}(\theta | y_1^n)$ is constructed from a profiled criterion, the 'idealized' variational posterior is then related to the frequentist profiled variational inference approach described in Westling and McCormick (2019). In their analysis, the authors view variational point estimators of the global parameters θ as M-estimators based on the profiled variational criterion function in (10). They then explore conditions and examples under which the variational point estimator, based on maximizing $\widehat{L}_n(\theta)$, do, or do not, deliver consistent estimates of θ_0.

While Westling and McCormick (2019) focus on consistency of variational point estimators, we study concentration of the 'idealized' posterior distribution $\widehat{q}(\theta | y_1^n)$. The following result shows that, under regularity conditions similar to those maintained in Westling and McCormick (2019), the 'idealized' variational posterior $\widehat{q}(\theta | y_1^n)$ is Bayes consistent for some value that may or may not coincide with θ_0.

Assumption 3.2. *(i) There exists a map* $\theta \mapsto \lambda(\theta) \in \Lambda$ *such that* $\sup_{\theta \in \Theta} \|\widehat{\lambda}_n(\theta) - \lambda(\theta)\| = o_p(1)$. *(ii) There exist a deterministic function* $\mathcal{L} : \Theta \times \Lambda \mapsto \mathbb{R}$ *and a* $\theta_\star \in \Theta$ *such that the following are satisfied: (a) for all* $\epsilon > 0$ *there exists some* $\delta > 0$ *such that* $\inf_{\theta \in d(\theta, \theta_\star) > \epsilon} [\mathcal{L}(\theta, \lambda(\theta)) - \mathcal{L}(\theta_\star, \lambda(\theta_\star))] \leq -\delta$*; (b)* $\sup_{\theta \in \Theta, \lambda \in \Lambda} |\mathcal{L}_n(\theta, \lambda)/n - \mathcal{L}(\theta, \lambda)| = o_p(1)$.

(iii) For any $\epsilon > 0$, $\int_\Theta \mathbf{1}\{\theta : \mathcal{L}(\theta, \lambda) - \mathcal{L}(\theta_\star, \lambda(\theta_\star)) < \epsilon\} p(\theta)\mathrm{d}\theta > 0$. *(iv) For all* n *large,* $\int_\Theta \exp\left\{\widehat{L}_n(\theta)\right\} p(\theta)\mathrm{d}\theta < \infty$.

Lemma 3.3. *Under Assumption 3.2, for any* $\epsilon > 0$, $\widehat{Q}\left(\{\theta \in \Theta : d(\theta, \theta_\star) > \epsilon\} | y_1^n\right) = o_p(1)$.

Assumption 3.2(2.b) implies that $L_n(\theta, \lambda)/n$ converges to $\mathcal{L}(\theta, \lambda)$ (uniformly in θ and λ); while part (2.a) is an identification condition and states that $\mathcal{L}(\theta, \lambda)$ is maximized at some $\theta_\star$, which may differ from θ_0. This identification condition makes clear that if $\theta_\star \neq \theta_0$, then $\mathcal{L}[\theta_\star, \lambda(\theta_\star)] > \mathcal{L}[\theta_0, \lambda(\theta_0)]$ and the idealized posterior for θ will not concentrate onto θ_0. This can be interpreted explicitly in terms of Jensen's gap as defined in (8) by recalling that under Assumption 3.1, $\ell_n(\theta_0) \to_p H(\theta_0)$, and by considering the limit of (the scaled) Jensen's gap evaluated at θ_0,

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\Upsilon_n\left(\theta_0, q_{\widehat{\lambda}_n(\theta_0)}\right) = H(\theta_0) - \mathcal{L}[\theta_0, \lambda(\theta_0)] \geq H(\theta_0) - \mathcal{L}[\theta_\star, \lambda(\theta_\star)] + \delta,$$

for some $\delta \geq 0$. If Assumption 3.2(2.a) is satisfied at $\theta_\star \neq \theta_0$, then $\delta > 0$, and $\kappa_n := \Upsilon_n(\theta_0, q_{\widehat{\lambda}_n(\theta_0)})/n \to_p C > 0$.

Taken together, Lemmas 3.1 and 3.3 show that, regardless of whether one conducts variational frequentist or Bayesian inference in SSMs, consistent inference for θ_0 will require that a version of Jensen's gap converges to zero. Moreover, as Example 3.1 has demonstrated, this is not likely to occur even in simple SSMs. The point is further exemplified in the follow example, where we explore the Bayesian consistency of the idealized VB posterior in the same linear Gaussian SSM.

Example 3.2 (Linear Gaussian model revisited)**.** Returning to the linear Gaussian SSM in Example 3.1, let us again consider the case where $\theta = (\alpha, \rho)'$ is unknown, while σ_0 is known, and we consider variational inference for θ using the idealized variational posterior. Our variational family for the vector of states is again taken to be $\mathcal{Q}_x$, which depends on the single variational parameter λ, and leads to a jointly Gaussian approximation with zero-mean and covariance matrix $\nu(\lambda)\Phi_n(\lambda)$ defined previously.

When $\sigma_0^2 = 1$, and known, the limit criterion $\mathcal{L}(\theta, \lambda)$ can be constructed analytically, and the mapping $\theta \mapsto \lambda(\theta)$, obtained by maximizing $\mathcal{L}(\theta, \lambda)$ with respect to λ for fixed θ,

calculated. For fixed θ, with $\rho \neq 0$, the mapping $\theta \mapsto \lambda(\theta)$ is given by (see the proof of Lemma 3.4 for details):

$$\lambda(\theta) = (\alpha^2 + \rho^2 - ((\alpha^2 + \rho^2 + \rho + 1)(\alpha^2 + \rho^2 - \rho + 1))^{1/2} + 1)/\rho.$$

In order for the idealized variational posterior $\widehat{q}(\theta|y_1^n)$ to concentrate onto θ_0, we require that the limit maximizer of $\mathcal{L}[\theta, \lambda(\theta)]$ coincide with θ_0 (see Lemma 3.3). The following result demonstrates that this does not occur in general.

Lemma 3.4. *Assume that $\rho \in [\underline{\rho}, \overline{\rho}]$, for some known $\underline{\rho} > 0$, but close to zero, and $\underline{\rho} < \overline{\rho} < 1$, and $\alpha \in [0, \overline{\alpha}]$, for some $\overline{\alpha} > 0$. Under the variational family $\mathcal{Q}_x$, we have*

$$\theta_\star = (\underline{\rho}, 0)' = \underset{\theta \in [\underline{\rho}, \overline{\rho}] \times [0, \overline{\alpha}]}{\operatorname{argmax}} \mathcal{L}[\theta, \lambda(\theta)].$$

Hence, if $\rho_0 > \underline{\rho}$, or $\alpha_0 > 0$, then $\widehat{Q}(\{\theta : d(\theta, \theta_0) > 0\}|y_1^n) > 0$ with probability converging to one.

The joint variational state approximation $q_\lambda(x_1^n)$ produces a closed-form marginal state approximation $q_\lambda(x_n)$, for any $n \geq 2$. Moreover, the marginal state posterior $\pi(x_n|\theta, y_1^n)$ is also known in closed form, for a given value of θ, and any $n \geq 2$. Given this, we can analytically evaluate the KL divergence between $q_\lambda(x_n)$ and $\pi(x_n|\theta, y_1^n)$, at any n, to characterize the accuracy of the resulting state approximation.

Corollary 3.1. *Under $\mathcal{Q}_x$, for any $n \geq 2$, we have $KL[\pi(x_n|\theta_0, y_1^n)||q_{\lambda(\theta_\star)}(x_n)] > 0$.*

The above result demonstrates that for any $n \geq 2$, the optimal variational state density is a biased approximation of the exact state density. Thus, even if θ_0 were known, and we only wished to conduct inference on x_1^n, the resulting variational approximation of the state density would ultimately deliver a poor approximation.

4 Implications

The above results suggest that VB methods can lead to inaccurate inference in the case of SSMs. In this section, we discuss further the implications of these results for inference on the global parameters, plus their implications for predictive accuracy.

4.1 Inference on global parameters

When conducting VB in SSMs, the need to approximate the posterior of x_1^n introduces a discrepancy between the exact posterior, and that which results from the VB approach. In this way, we can view the latent states x_1^n as *incidental or nuisance* parameters (see Lancaster, 2000, for a review), which are needed to make feasible the overall optimization problem, but which, in and of themselves, are not the object of interest. A similar point is made by Westling and McCormick (2019) in the case of independent states, and frequentist variational inference, where the authors demonstrate that inconsistency can occur, even in the case of independent observations, if delicate care is not taken with the choice of variational class for x_1^n.

However, the incidental parameter problem has not stopped researchers from using VB methods to conduct inference on θ in SSMs. While the general conclusions elucidated above apply, in principle, to *all such methods*, we next discuss two specific categories of VB methods in greater detail, and comment on their ability to deliver consistent inference for θ_0.

4.1.1 Integration approaches

A possible VB approach is to first 'integrate out' the latent states so that there is no need to perform joint inference on (θ, x_1^n). Such an approach can be motivated by the fact that if we take $q_x(x_1^n|\theta) = \pi(x_1^n|y_1^n, \theta)$, (i.e. take the variational approximation for x_1^n to be equivalent to the exact posterior for x_1^n conditional on θ), then we can rewrite $\mathrm{KL}(q||\pi)$ as

$$\begin{aligned}\mathrm{KL}(q||\pi) &= \int_\Theta \int_{\mathcal{X}} q_\theta(\theta)\pi(x_1^n|y_1^n,\theta) \log \frac{q_\theta(\theta)\pi(x_1^n|y_1^n,\theta)}{\pi(x_1^n|y_1^n,\theta)\pi(\theta|y_1^n)} \mathrm{d}x_1^n \mathrm{d}\theta \\ &= \mathrm{KL}[q_\theta||\pi(\theta|y_1^n)],\end{aligned}$$

with the final line exploiting the fact that $\pi(x_1^n|y_1^n, \theta)$ integrates to one for all θ. Thus, if we are able to use as our variational approximation for the states the actual (conditional) posterior, we can transform a variational problem for (θ, x_1^n) into a variational problem for θ alone.

The above approach is adopted by Loaiza-Maya *et al.* (2021), and is applicable in any

case where draws from $p(x_1^n|y_1^n, \theta)$ can be reliably and cheaply obtained, with the resulting draws then used to 'integrate out' the states via the above KL divergence representation. While the approach of Loaiza-Maya *et al.* (2021) results in the above simplification, the real key to their approach is that it can be used to unbiasedly estimate the gradient of $\text{ELBO}[q_\theta||\pi(\theta|y_1^n)]$ (equivalent, in turn, to the gradient of the joint ELBO in (6), by the above argument). This, in turn, allows optimization over q_θ to produce an approximation to the posterior $\pi(\theta|y_1^n)$. Indeed, such an approach can be applied in many SSMs, such as unobserved component models like the LGSSM, in which draws from $\pi(x_1^n|y_1^n, \theta)$ can be generated exactly via, for example, forward (Kalman) filtering and backward sampling (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994); or various nonlinear models (e.g. those featuring stochastic volatility), in which efficient Metropolis- Hastings-within-Gibbs algorithms are available (Kim *et al.*, 1998; Jacquier *et al.*, 2002; Primiceri, 2005; Huber *et al.*, 2020).

In cases where we are not able to sample readily from $\pi(x_1^n|y_1^n, \theta)$ it may still be possible to integrate out the states using particle filtering methods. To this end, assume that we can obtain an unbiased estimate of the observed data likelihood $p_\theta(y_1^n)$ using a particle filter, which we denote by $\widehat{p}_\theta(y_1^n)$. We follow Tran *et al.* (2017) and write $\widehat{p}_\theta(y_1^n)$ as $\widehat{p}(y_1^n|\theta, z)$ to make the estimator's dependence on the random filtering explicit through the dependence on a random variable z, with z subsequently defined by the condition $z = \log \widehat{p}(y_1^n|\theta, z) - \log p_\theta(y_1^n)$. For $g(z|\theta)$ denoting the density of $z|\theta$, Tran *et al.* (2017) consider VB for the augmented posterior

$$\pi(\theta, z|y_1^n) = \widehat{p}(y_1^n|\theta, z)g(z|\theta)p(\theta)/p(y_1^n) = p_\theta(y_1^n)\exp(z)g(z|\theta)p(\theta)/p(y_1^n) = \pi(\theta|y_1^n)\exp(z)g(z|\theta),$$

which, marginal of z, has the correct target posterior $\pi(\theta|y_1^n)$ due to the unbiasedness of the estimator $\widehat{p}(y_1^n|\theta, z)$. The authors refer to the resulting method as variational Bayes with an intractable likelihood function (VBIL). The VBIL posteriors can be obtained by considering a variational approximation to $\pi(\theta, z|y_1^n)$ that minimizes the KL divergence

between $q(\theta, z) = q_\theta(\theta)g(z|\theta)$ and $\pi(\theta, z|y_1^n)$:

$$\begin{aligned}
\mathrm{KL}[q(\theta, z)||\pi(\theta, z|y_1^n)] &= \int_\Theta \int_{\mathcal{Z}} q_\theta(\theta)g(z|\theta) \log \frac{q_\theta(\theta)g(z|\theta)}{\pi(\theta|y_1^n)\exp(z)g(z|\theta)} \mathrm{d}z\mathrm{d}\theta \\
&= \int_\Theta \int_{\mathcal{Z}} q_\theta(\theta)g(z|\theta) \log \frac{q_\theta(\theta)}{p_\theta(y_1^n)\exp(z)p(\theta)} \mathrm{d}z\mathrm{d}\theta + \log p(y_1^n) \\
&= -\int_\Theta q_\theta(\theta) \log p_\theta(y_1^n) \mathrm{d}\theta + \mathrm{KL}(q_\theta||p_\theta) + \Upsilon_n[q(\theta, z)] + \log p(y_1^n),
\end{aligned}$$

where in this case

$$\Upsilon_n[q(\theta, z)] = \int_\Theta \int_{\mathcal{Z}} q_\theta(\theta)g(z|\theta) \left\{\log p_\theta(y_1^n) - \log \widehat{p}(y_1^n|\theta, z)\right\} \mathrm{d}z\mathrm{d}\theta.$$

For fixed θ, $\mathbb{E}_z\left[\widehat{p}(y_1^n|\theta\, z)\right] = p_\theta(y_1^n)$, but in general $\log \widehat{p}(y_1^n|\theta, z)$ is a biased estimator of $\log p_\theta(y_1^n)$, from which it follows that $\Upsilon_n[q(\theta, z)] \geq 0$. However, in contrast to the general approximation of the states discussed in Section 3, which intimately relies on the choice of the approximating density $q_x(x_1^n|\theta)$, VBIL can achieve consistent inference on θ_0 by choosing an appropriate number of particles N in the production of $\widehat{p}(y_1^n|\theta, z)$.

To see this, we recall that a maintained assumption in the literature on PMCMC methods is that, for all n and N, the conditional mean and variance of the density $g(z|\theta)$ satisfy $\mathbb{E}[z|\theta] = -\gamma(\theta)^2/2N$, and $\mathrm{Var}\left[z|\theta\right] = \gamma(\theta)^2/N$, where $\gamma(\theta)^2$ is bounded uniformly over Θ; see, e.g., Assumption 1 in Doucet *et al.* (2015) and Assumption 1 in Tran *et al.* (2017). However, in general, N is assumed to be chosen so that $\mathbb{E}[z|\theta] = -\sigma^2/2$ and $\mathrm{Var}\left[z|\theta\right] = \sigma^2 > 0$, $0 < \sigma < \infty$. Note that, under this choice for N, for any $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr\left[\Upsilon_n[q(\theta_0, z)]/n > \varepsilon\right] = \lim_{n\to\infty} \Pr\left[-q_\theta(\theta_0)\mathbb{E}\left[z|\theta_0\right] > n\epsilon\right] = \lim_{n\to\infty} \Pr\left[q_\theta(\theta_0)\sigma^2/2 > n\varepsilon\right] = 0,$$

assuming $q_\theta(\theta_0), \sigma^2 < \infty$.

From this condition, we see that the VBIL inference problem is asymptotically the same as the VB inference problem for θ alone. Consequently, existing results on the posterior concentration of VB methods for θ alone can be used to deduce posterior concentration of the VBIL posterior for θ.

4.1.2 Structured approximations of the states

Yet another approach for dealing with variational inference in the presence of states is to consider a structured approximation that allows for a dynamic updating of the approxima-

tion for the posterior of the states. Such an approximation can be achieved by embedding in the class of variational densities an analytical filter, like the Kalman filter. Koop and Korobilis (2020) propose the use of the Kalman filter *within VB* (VBKF) as a means of approximating the posterior density of the states using Kalman recursions. In particular, the authors approximate the posterior $\pi(x_1^n|y_1^n,\theta)$ by approximating the relationship between X_t and X_{t-1}, which may in truth be non-linear in θ, by the random walk model $X_t = X_{t-1} + \epsilon_t$, with $\epsilon_t \sim i.i.d.N(0,\sigma_0^2)$, and then use Kalman filtering to update the states in conjunction with a linear approximation to the measurement equation. Using this formulation, the variational approximation is of the form $q(x_1^n,\theta) = q_\theta(\theta)q_x(x_1^n)$, where $q_x(x_1^n) \propto \prod_{k\geq 1}\exp(-\left\{x_k - \widehat{x}_{k|k}\right\}^2(1-\mathcal{K}_k)P_{k|k-1}/2)$ and where the terms $\mathcal{K}_k, P_{k|k-1}, \widehat{x}_{k|k}$ are explicitly calculated using the Kalman recursion: $\widehat{x}_{k|k} = \widehat{x}_{k|k-1} + \mathcal{K}_k(y_t^\star - \widehat{x}_{k|k-1})$, and where $\mathcal{K}_k$ is the Kalman gain, $P_{k|k-1}$ is the predicted variance of the state, and in the application of Koop and Korobilis (2020), $y_t^\star = \log(y_t^2)$.

While the solution proposed by the VBKF is likely to lead to better inference on the states, especially when x_1^n behaves like a random walk, ultimately we are still 'conducting inference' on x_1^n, and thus we still encounter the incidental parameter problem as a consequence. Indeed, taking as the variational family for x_1^n the Kalman filter approximation yields, at time $k \geq 1$, a conditionally normal density with mean $\hat{x}_{k|k}$ and variance $(1-\mathcal{K}_k)P_{k|k-1}$. Hence, we have a variational density that has the same structure as in Lemma 3.2, but which allows for a time varying mean and variance. Given this similarity, there is no reason to suspect that such an approach will yield inferences that are consistent. Indeed, further intuition can be obtained by noting that, in the VBKF formulation, the simplification of the state equation means that we disregard any dependence between the states and the values of θ that drive their dynamics.

The variational approach of Chan and Yu (2020) can be viewed similarly: the suggested algorithm assumes and exploits a particular dynamic structure for the states that allows for analytical (posterior) updates and thus leads to computationally simple estimates for the variational densities of $q_x(x_1^n)$. As with the VBKF approach, the assumed nature of the state process used by Chan and Yu (2020) to estimate $q_x(x_1^n)$ implies that, in general, it

is unlikely that Bayesian consistency can be achieved. Due to space restrictions, further discussion on the specifics of this approach are relegated to Sections A.1.1 and B.1.4 of the Supplementary Appendix.

4.2 VB-based prediction

VB provides, at best, an approximation to the posterior and, as a result, may well yield less accurate inferences than those produced by the exact posterior (see, e.g. Koop and Korobilis, 2020; Gunawan *et al.*, 2021). However, VB *can* perform admirably in predictive settings, see, e.g., Quiroz *et al.* (2018) and Frazier *et al.* (2021), amongst others, in the sense of replicating the out-of-sample accuracy achieved by exact predictives, when such comparators are available. (See Frazier *et al.*, 2019 for a comparable finding in the context of predictions based on ABC.) Therefore, even though the VB posterior may not necessarily converge to the true value θ_0, so long as the value onto which it is concentrating is not too far away from θ_0, it may be that VB-based predictions perform well in practice.

Recall the conditional density of Y_{n+1} given x_{n+1} and θ is $g_\theta(Y_{n+1}|x_{n+1})$, so that the predictive pdf for Y_{n+1} can be expressed as

$$\begin{aligned} p(Y_{n+1}|y_1^n) &= \int_\Theta \int_{\mathcal{X}} g_\theta(Y_{n+1}|x_{n+1}) \pi(x_1^{n+1}, \theta|y_1^n) \mathrm{d}x_1^{n+1} \mathrm{d}\theta \qquad (11) \\ &= \int_\Theta \int_{\mathcal{X}} \int_{\mathcal{X}} g_\theta(Y_{n+1}|x_{n+1}) \underbrace{p(x_{n+1}|x_n, y_1^n, \theta) p(x_1^n|y_1^n, \theta)}_{(1)} \underbrace{\pi(\theta|y_1^n)}_{(2)} \mathrm{d}x_{n+1} \mathrm{d}x_1^n \mathrm{d}\theta, \end{aligned}$$

where the last line follows from the Markovianity of the state transition equation (see equation (2)). In many large SSMs, using MCMC methods to estimate (13) is infeasible or prohibitive computationally, due to the difficulty of sampling from $\pi(x_1^{n+1}, \theta|y_1^n)$. Instead, VB methods can produce an estimate of $p(Y_{n+1}|y_1^n)$ by approximating, in various ways, the two pieces in equation (13) underlined as (1) and (2). All such methods replace the second underlined term by some approximate posterior for θ, but differ in how they access the first underlined term.

In all the cases of which we are aware, we can separate VB methods for prediction in SSMs into two classes: a class which makes explicit use of a variational approximation to the states, $\widehat{q}_x$ to replace $p(x_1^n|y_1^n, \theta)$; and a class that uses an accurate simulation-based

estimate of $p(x_1^n|y_1^n, \theta)$. Due to space restrictions, we do not give a detailed discussion of how these VB predictives are produced, and instead refer the interested reader to Section A.1 in the Supplementary Appendix.

Any Bayesian method that replaces $\pi(\theta|y_1^n)$ in part (2) by an approximation, e.g., $\widehat{q}_\theta$ in the case of VB, will lead to some inaccuracy, however, as shown by Frazier *et al.* (2019) in the case of ABC, this loss in accuracy is often minimal. Therefore, what really matters in terms of accurate prediction in SSMs using VB is the replacement of (1) in (13). Replacing (1) with an accurate simulation-based estimate is likely to deliver more accurate estimators, at the cost of additional computation. However, it is not necessarily clear that the resulting predictions will perform much better than those approaches based on the approximation $\widehat{q}_x$. In the following section, we demonstrate that even through the inference that results from using $\widehat{q}_x$ instead of (1) can be poor, the resulting predictive performance is often quite reasonable, at least for sample sizes that are not too large.

5 Numerical assessment of VB methods

In this section, we shed further light on the phenomenon of the predictive accuracy of VB methods, and connect the performance of these methods to the inconsistency for θ_0 that can result as the sample size diverges. The results suggest that, in terms of predictive accuracy, there is little difference between methods in small sample sizes or with a small number of out-of-sample observations. However, we document a clear hierarchy across methods as the sample size becomes larger and as the out-of-sample evaluation increases.

5.1 Simulation design

We now compare the inferential and predictive accuracy of the variational methods of Quiroz *et al.* (2018) and Loaiza-Maya *et al.* (2021) against an exact MCMC-based estimate of $\pi(\theta, x_1^n|y_1^n)$, referred to as 'exact Bayes' hereafter, in a simulation exercise. In Section B.1 of the Supplementary Appendix, we provide complete details on the implementation of each of these methods under this particular simulation design. However, we remark here that

Quiroz *et al.* (2018) is an example of a VB method in which the states are approximated via a particular choice of variational family, whilst Loaiza-Maya *et al.* (2021) (as noted in Section 4.1.1) adopt a variational approximation for the posterior of the global parameters only, with the conditional posterior of the states accessed via simulation.

The assumed DGP is specified as an unobserved component model with stochastic volatility (UCSV):

$$\mu_t = \bar{\mu} + \rho_\mu (\mu_{t-1} - \bar{\mu}) + \sigma_\mu \varepsilon_t, \quad h_t = \bar{h} + \rho_h \left(h_{t-1} - \bar{h}\right) + \sigma_h \eta_t, \quad Y_t = \mu_t + \exp(h_t/2) u_t, \quad (12)$$

where $(\varepsilon_t, \eta_t, u_t)' \overset{i.i.d.}{\sim} N(0, I_3)$. The unobserved component term μ_t is a latent variable that captures the persistence in the conditional mean of Y_t, while the stochastic volatility term h_t captures the persistence in the conditional variance. We consider the following three set of values for the true parameters:

$$\begin{aligned}
&\text{DGP 1:} \quad \bar{\mu}_0 = 0;\ \rho_{\mu_0} = 0.8;\ \sigma_{\mu_0} = 0.5;\ \bar{h}_0 = -1.0;\ \rho_{h0} = 0.00;\ \sigma_{h0} = 0.0 \\
&\text{DGP 2:} \quad \bar{\mu}_0 = 0;\ \rho_{\mu_0} = 0.0;\ \sigma_{\mu_0} = 0.5;\ \bar{h}_0 = -1.3;\ \rho_{h0} = 0.95;\ \sigma_{h0} = 0.3 \\
&\text{DGP 3:} \quad \bar{\mu}_0 = 0;\ \rho_{\mu_0} = 0.8;\ \sigma_{\mu_0} = 0.5;\ \bar{h}_0 = -1.3;\ \rho_{h0} = 0.95;\ \sigma_{h0} = 0.3
\end{aligned}$$

The specifications for DGP 1 produce a time series process that has substantial persistence in the conditional mean, and a constant variance; DGP 2 generates a process that has substantial persistence in the conditional variance, and a fixed marginal mean of zero; whilst DGP 3 corresponds to a process that exhibits persistence in both the conditional mean and variance. The true parameter vector in each case is defined as $\theta_0 = \left(\bar{\mu}_0, \rho_{\mu_0}, \sigma_{\mu_0}, \bar{h}_0, \rho_{h0}, \sigma_{h0},\right)'$.

For the predictive assessment we compare exact Bayes with the two variational methods cited above plus the method of Chan and Yu (2020). As discussed in Section B.1 of the Supplementary Appendix, the method of Chan and Yu (2020) exploits a very specific structure in the construction of the variational algorithm, which in this case corresponds to DGP 2 under the parameter restrictions $\rho_{h0} = 1.0$, $\sigma_{\mu_0} = 0.0$, and $\bar{h}_0 = 0.0$. Thus, application of this approach under any of the above true DGPs constitutes misspecified inference; hence, we do not include this technique in the inferential assessment. Due to space constraints, certain tables and figures are included in Section B.2 of the Supplementary Appendix.

5.2 Accuracy of inference on the states

We assess inferential accuracy through lens of state estimation. To this end, we generate a times series of length $T = 11000$ from each of the three true DGP specifications. The full sample is used to produce the exact posterior as well as the two approximate posteriors corresponding to the QNK and LSND methods; hence, we are able to shed some light on the theoretical consistency results provided above. We assess the inferential accuracy of each method (exact and approximate) by calculating the root mean squared error (RMSE) and mean absolute error (MAE) of each sequence of marginal posterior means, for $t = 1, 2, ..., T$, for the unobserved component, μ_t, and the stochastic standard deviation, $\exp(h_t/2)$, relative to the marginal posterior means that results when we condition on the true parameters, denoted respectively by $\mathbb{E}[\mu_t|\theta_0, y_1^T]$ and $\mathbb{E}[\exp(h_t/2)|\theta_0, y_1^T]$, $t = 1, 2, ..., T$. The results are presented in Table 1.

As expected, exact Bayes produces the most accurate point estimates for the two sets of latent variables (both μ_t and $\exp(h_t/2)$), as tallies with the theoretical guarantees of this method. In terms of the VB methods, the LSND results closely match those of exact Bayes as this method does not suffer from the incidental parameter problem. In contrast, the QNK method does not deal directly with this problem and, as a consequence, exhibits - across *all of the designs* recorded in Table 1 - inaccuracy that is between two and ten times greater than that of both exact Bayes and the LSND method. From the results recorded in Table 3 in Supplementary Appendix B.2, we also note that the time taken to estimate the UCSV model, under all three DGPs, is approximately the same for exact Bayes and the LSND method, with the QNK approach taking roughly twice as long as both.

We further highlight the results in Table 1 by plotting, in Figure 1, the marginal posterior means for both μ_t and $\exp(h_t/2)$, for each point in time across a given sample period, for all three methods; with the sequence of 'true' posterior means (that condition on θ_0) included for comparison. For the sake of brevity, we only present results for DGP 2, with the corresponding results for DGPs 1 and 3 placed in Supplementary Appendix B.2. Consistent with the summary results in Table 1, the posterior means for exact Bayes and LSND are both very similar, for each t, and visually very close to the corresponding true time

Table 1: Accuracy in the estimation of the unobserved component and conditional standard deviation. Panel A presents the root mean squared error (RMSE) and mean absolute error (MAE) of the posterior mean estimates of the unobserved component (μ_t). The columns correspond to the three DGP specifications, while the rows correspond to the three predictive methods: exact Bayes, LSND and QNK. Panel B presents the corresponding results for the posterior mean estimates of the conditional standard deviation. The unobserved component error measures are computed relative to $\mathbb{E}[\mu_t|\theta_0, y_1^T]$, while the conditional standard deviation error measures are computed relative to $\mathbb{E}[\exp(h_t/2)|\theta_0, y_1^T]$, where θ_0 denotes the true parameter vector and $T = 11000$.

Panel A: (μ_t)	RMSE				MAE		
	DGP 1	DGP 2	DGP 3		DGP 1	DGP 2	DGP 3
Exact Bayes	0.0463	0.0207	0.0203	Exact Bayes	0.0382	0.0004	0.0155
LSND	0.0495	0.0253	0.0271	LSND	0.0405	0.0006	0.0207
QNK	0.1211	0.2646	0.1098	QNK	0.0980	0.0700	0.0664
Panel B: ($\exp(h_t/2)$)	RMSE				MAE		
	DGP 1	DGP 2	DGP 3		DGP 1	DGP 2	DGP 3
Exact Bayes	0.0497	0.0255	0.0234	Exact Bayes	0.0470	0.0198	0.0180
LSND	0.0520	0.0309	0.0315	LSND	0.0490	0.0242	0.0246
QNK	0.0984	0.2505	0.2231	QNK	0.0983	0.2179	0.1656

posterior means, across the entire sample. In comparison, the QNK method consistently produces point estimates of the states that are very different from the values that condition on the true parameters, as accords with the dependence of the method on a variational approximation for the states, and the consequent loss of Bayesian consistency for θ_0. We note that the additional figures in Supplementary Appendix B.2 demonstrate that, at least visually, the QNK method seems to produce more accurate estimates of μ_t under DGPs 1 and 3 than it does under DGP 2; however, it remains inaccurate in terms of estimating $\exp(h_t/2)$ under these alternative DGPs.

5.3 Predictive accuracy

To assess the predictive accuracy of each method we conduct an expanding window prediction exercise using the same generated data as in the previous subsection. The exercise consists of constructing the Bayesian predictive density for Y_{n+1}, conditional on the sample

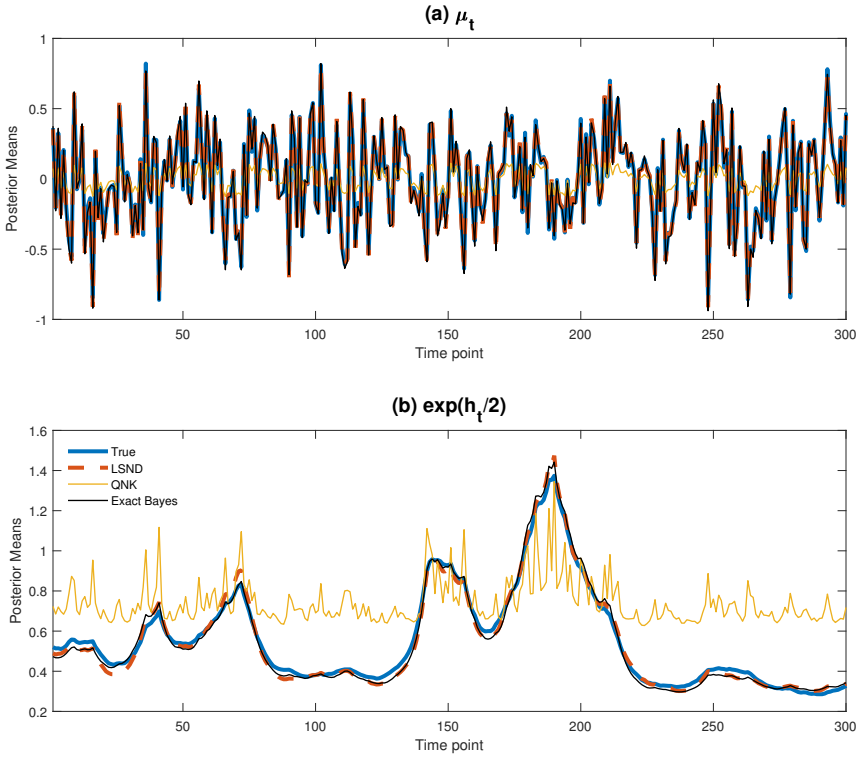

Figure 1: Posterior means of the latent states, under DGP 2, over the first 300 time points. Panel (a) plots the posterior mean for μ_t. The red, gold and black lines plot, respectively, the posterior means based on the LSND, QNK and exact Bayes approaches. The blue line plots the posterior mean that conditions on the true parameters. Panel (b) presents corresponding results for $\exp(h_t/2)$.

y_1^n, for each of the competing approaches and for $n \in \{1000, \ldots, T-1\}$. For each method and each out-of-sample time point we evaluate eight measures of predictive accuracy: the logarithmic score, four censored scores, the continuously ranked probability score, the tail weighted continuously ranked probability score and the interval score. Details of all scoring rules, including appropriate references, are provided in Section B.3 of the Supplementary Appendix. We document results using 100, 1000 and 10000 out-of-sample evaluations respectively, remembering that the CY method is now included in the comparison, but only for the case of DGP 2. For reasons of space, we only present results for the largest number of out-of-sample evaluations (10000) in the main text, in Table 2, while the results for the other evaluation periods are given in Section B.2 of the Supplementary Appendix, in Tables 4 and 5 respectively.

Focussing first on the results in Table 2, based on the very large number of out-of-sample evaluations, we observe an interesting ranking. Across all designs, and according to all measures of accuracy, exact Bayes is the most accurate method. As accords with the inferential results discussed above, the LSND method has a predictive accuracy that often matches, or is extremely similar to, that of exact Bayes, followed, in order, by CY and QNK. A similar ranking holds for the results recorded in Tables 4 and 5 in Section B.2 of the Supplementary Appendix. However, the differences between methods are somewhat less stark over the smaller out-of-sample evaluation periods, which highlights the fact that it is ultimately the consistency properties of the different VB methods (in evidence for the largest evaluation period, given the large size of the expanding estimation windows) that is driving the discrepancies between the predictive accuracy of the competing methods.

Whilst a ranking is in evidence in Table 2, it can be argued that across *certain* DGP and scoring rule combinations, the predictive results across the different methods are still quite similar, both between the exact and (all) VB methods, and between the different VB methods. That is, for certain combinations of DGPs and scoring rules, all methods are seen to perform well (relative to the benchmark of the true predictive), and the more substantial *inferential* discrepancies observed between certain of the methods are not reflected at the predictive level. This finding corroborates the point made earlier, and which has been supported by other findings in the literature, namely that computing a posterior via an approximate method does not *necessarily* reduce predictive accuracy (relative to exact Bayes) by a substantial amount.

However, despite there being certain DGP and scoring rule combinations where the methods perform similarly, this is not true across all DGPs and loss measures, in particular for the larger out-of-sample evaluation period. For example, and with specific reference to Table 2, there is a clear trend that as model complexity increases (i.e. moving from DGP 1 through to DGP 3), variational methods that work harder to correctly approximate the states have greater predictive accuracy. This finding is particularly marked for the log score and the interval score, which directly measure the dispersion of the posterior predictive. In the case of DGP 3, the all-purpose variational method of Quiroz *et al.* (2018) performs

the worst across all the methods under analysis, and most notably for the log score and the interval score. This feature is most likely due to the fact that the posteriors associated with the method of Quiroz *et al.* (2018) have overly thin tails. Consequently, parameter uncertainty is not adequately accounted for when constructing the posterior predictive, which results in a predictive with thin tails, and ultimately translates into poor performance in scores that measure both location and/or dispersion.

Table 2: Predictive performance of competing Bayesian approaches: exact Bayes, LSND and CY and QNK. The column labels indicate the out-of-sample predictive performance measure while the row labels indicate the predictive method. 'True DGP' indicates the productive results that condition on the true parameters. Panels A, B and C correspond to the results for DGP 1, 2 and 3, respectively. The average predictive measures in this table were computed using 10000 out-of-sample evaluations.

Panel A: DGP 1	LS	CS-10%	CS-20%	CS-80%	CS-90%	CRPS	TWCRPS	IS
True DGP	-1.259	-0.308	-0.508	-0.505	-0.297	-0.481	-0.146	-4.001
Exact Bayes	-1.260	-0.308	-0.508	-0.506	-0.297	-0.481	-0.147	-4.012
LSND	-1.261	-0.308	-0.509	-0.507	-0.298	-0.481	-0.147	-4.015
CY	-	-	-	-	-	-	-	-
QNK	-1.262	-0.309	-0.509	-0.507	-0.298	-0.482	-0.147	-4.030
Panel B: DGP 2	LS	CS-10%	CS-20%	CS-80%	CS-90%	CRPS	TWCRPS	IS
True DGP	-1.192	-0.341	-0.551	-0.551	-0.340	-0.454	-0.139	-4.097
Exact Bayes	-1.193	-0.342	-0.551	-0.551	-0.340	-0.454	-0.139	-4.093
LSND	-1.194	-0.342	-0.551	-0.552	-0.341	-0.454	-0.139	-4.101
CY	-1.205	-0.346	-0.557	-0.555	-0.344	-0.456	-0.140	-4.182
QNK	-1.212	-0.350	-0.560	-0.560	-0.349	-0.456	-0.140	-4.316
Panel C: DGP 3	LS	CS-10%	CS-20%	CS-80%	CS-90%	CRPS	TWCRPS	IS
True DGP	-1.268	-0.304	-0.505	-0.521	-0.300	-0.490	-0.150	-4.424
Exact Bayes	-1.268	-0.305	-0.506	-0.520	-0.299	-0.491	-0.150	-4.423
LSND	-1.271	-0.306	-0.507	-0.521	-0.301	-0.491	-0.150	-4.442
CY	-	-	-	-	-	-	-	-
QNK	-1.301	-0.315	-0.521	-0.536	-0.311	-0.497	-0.152	-4.707

6 Discussion

We have systematically documented the behavior of variational methods, in terms of inference and prediction, within the class of state space models (SSMs). Sufficient conditions for (both frequentist and Bayesian) consistency of variational inference (VI) in SSMs have been presented in terms of the so-called Jensen's gap, which measures the discrepancy introduced within VI due to the approximation of the states. Focusing on variational Bayes (VB) methods specifically, we show that only methods that are capable of closing Jensen's gap yield Bayesian consistent inference for the global parameters and, in turn, deliver more accurate inferences for the states.

In the context of empirically relevant SSMs, we find numerical evidence of a clear hierarchy in terms of the accuracy of state inference across different variational methods: methods that can close Jensen's gap produce qualitatively more accurate inferences than those that do not. However, whilst this same hierarchy also holds for VB-based prediction, we find that the extent to which different variational approaches vary in terms for predictive accuracy depends on the data generating process (DGP), the loss in which the different methods are evaluated, and - most importantly - the size of the out-of-sample evaluation period. Indeed, we document that there are certain circumstances, i.e., sample size, DGP and loss combinations, where there is little to separate the various approaches. However, in large samples, methods that attain Bayesian consistent inference on the global parameters produce more accurate predictions.

To keep the length of this paper manageable, we have deliberately analysed and compared only a select few of the variational methods used to conduct inference and prediction in SSMs. Our findings, however, suggest that certain classes of approximations for the state posterior employed in the machine learning literature, e.g., classes based on normalising or autoregressive flows, may be flexible enough to deliver accurate inferences and predictions; we refer to, e.g., Ryder *et al.*, 2018, and the references therein, for a discussion of such methods in SSMs. We leave a comparison between the approaches discussed herein and those commonly used in machine learning for future research.

References

Ait-Sahalia, Y. and Kimmel, R. (2007). Maximum likelihood estimation of stochastic volatility models. *Journal of Financial Economics*, 83(2):413–452.

Aït-Sahalia, Y., Li, C., and Li, C. X. (2021). Closed-form implied volatility surfaces for stochastic volatility models with jumps. *Journal of Econometrics*, 222(1):364–392.

Andersen, T. G. and Srensen, B. E. (1996). GMM estimation of a stochastic volatility model: A Monte Carlo study. *Journal of Business & Economic Statistics*, 14(3):328–352.

Andrieu, C., Doucet, A., and Holenstein, R. (2011). Particle Markov chain Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):269–342. With discussion.

Bates, D. S. (2006). Maximum Likelihood Estimation of Latent Affine Processes. *The Review of Financial Studies*, 19(3):909–965.

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. *https://arxiv.org/abs/1701.02434v2*.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Chan, J. C. and Yu, X. (2020). Fast and accurate variational inference for large Bayesian vars with stochastic volatility. *CAMA Working Paper*.

Chernozhukov, V. and Hong, H. (2003). An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.

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.

Danielsson, J. and Richard, J.-F. (1993). Accelerated Gaussian importance sampler with application to dynamic latent variable models. *Journal of Applied Econometrics*, 8(S1):S153–S173.

Dean, T. A., Singh, S. S., Jasra, A., and Peters, G. W. (2014). Parameter estimation for hidden Markov models with intractable likelihoods. *Scandinavian Journal of Statistics*, 41(4):970–987.

Diks, C., Panchenko, V., and Van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2):215–230.

Douc, R., Moulines, E., Olsson, J., Van Handel, R., et al. (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. *the Annals of Statistics*, 39(1):474–513.

Doucet, A., Pitt, M. K., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313.

Durbin, J. and Koopman, S. J. (2001). *Time Series Analysis by State Space Methods*. OUP.

Fearnhead, P. (2011). Bayesian inference for time series state space models. In Brooks, S., Gelman, A., Jones, G., and Meng, X., editors, *Handbook of Markov Chain Monte Carlo*, chapter 21, pages 513–530. Taylor & Francis.

Flury, T. and Shephard, N. (2011). Bayesian inference based only on a simulated likelihood. *Econometric Theory*, 27:933–956.

Frazier, D. T., Loaiza-Maya, R., Martin, G. M., and Koo, B. (2021). Loss-based variational Bayes prediction. *arXiv preprint arXiv:2104.14054*.

Frazier, D. T., Maneesoonthorn, W., Martin, G. M., and McCabe, B. P. (2019). Approximate Bayesian forecasting. *International Journal of Forecasting*, 35(2):521–539.

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika*, 105(3):593–607.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of time series analysis*, 15(2):183–202.

Gallant, A. R. and Tauchen, G. (1996). Which moments to match? *Econometric Theory*, 12(4):657–681.

Giordani, P., Pitt, M., and Kohn, R. (2011). Bayesian inference for time series state space models. In Geweke, J., Koop, G., and van Dijk, H., editors, *The Oxford Handbook of Bayesian Econometrics*, chapter 3, pages 61–124. OUP.

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

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422.

Gunawan, D., Kohn, R., and Nott, D. (2021). Variational Bayes approximation of factor stochastic volatility models. *International Journal of Forecasting*, 37(4):1355–1375.

Harvey, A., Koopman, S., and Shephard, N. (2004). *State Space and Unobserved Component Models: Theory and Applications*. CUP.

Huber, F., Koop, G., and Onorante, L. (2020). Inducing sparsity and shrinkage in time-varying parameter models. *Journal of Business & Economic Statistics*, pages 1–15.

Jacquier, E., Polson, N. G., and Rossi, P. E. (2002). Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics*, 20(1):69–87.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *The review of economic studies*, 65(3):361–393.

Koop, G. and Korobilis, D. (2020). Bayesian dynamic variable selection in high dimensions. *Available at SSRN 3246472*.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of econometrics*, 95(2):391–413.

Loaiza-Maya, R., Smith, M. S., Nott, D. J., and Danaher, P. J. (2021). Fast and accurate variational inference for models with many latent variables. *Forthcoming. Journal of Econometrics*.

Martin, G. M., McCabe, B. P. M., Frazier, D. T., Maneesoonthorn, W., and Robert, C. P. (2019). Auxiliary likelihood-based approximate Bayesian computation in state space models. *Journal of Computational and Graphical Statistics*, 28(3):508–522.

Miller, J. W. (2021). Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168):1–53.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32.

Ong, V. M.-H., Nott, D. J., and Smith, M. S. (2018). Gaussian variational approximation with a factor covariance structure. *Journal of Computational and Graphical Statistics*, 27(3):465–478.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, pages 1027–1057.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852.

Quiroz, M., Nott, D. J., and Kohn, R. (2018). Gaussian variational approximation for high-dimensional state space models. *arXiv preprint arXiv:1801.07873*.

Ruiz, E. (1994). Quasi-maximum likelihood estimation of stochastic volatility models. *Journal of Econometrics*, 63(1):289–306.

Ryder, T., Golighty, A., McGough, A. S., and Prangle, D. (2018). Black-box autoregressive density estimation for state-space models. *arXiv preprint arXiv:1811.08337*.

Sandmann, G. and Koopman, S. J. (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *Journal of Econometrics*, 87(2):271–301.

Syring, N. and Martin, R. (2020). Gibbs posterior concentration rates under sub-exponential type losses. *arXiv preprint arXiv:2012.04505*.

Tran, M.-N., Nott, D. J., and Kohn, R. (2017). Variational Bayes with intractable likelihood. *Journal of Computational and Graphical Statistics*, 26(4):873–882.

Westling, T. and McCormick, T. (2019). Beyond prediction: A framework for inference with variational approximations in mixture models. *Journal of Computational and Graphical Statistics*, 28(4):778–789.

Yang, Y., Pati, D., Bhattacharya, A., et al. (2020). alpha-variational inference with statistical guarantees. *Annals of Statistics*, 48(2):886–905.

A Further details and discussion on variational methods in SSMs

A.1 Methods for producing variational predictives

Following on from the discussion in Section 4.2 in the main text, in this section we give precise details on how the variational predictives are constructed. Recall that the predictive

pdf for Y_{n+1} can be expressed as

$$p(Y_{n+1}|y_1^n) = \int_\Theta \int_{\mathcal{X}} g_\theta(Y_{n+1}|x_{n+1})\pi(x_1^{n+1}, \theta|y_1^n)\mathrm{d}x_1^{n+1}\mathrm{d}\theta \tag{13}$$
$$= \int_\Theta \int_{\mathcal{X}} \int_{\mathcal{X}} g_\theta(Y_{n+1}|x_{n+1}) \underbrace{p(x_{n+1}|x_n, y_1^n, \theta)p(x_1^n|y_1^n, \theta)}_{(1)} \underbrace{\pi(\theta|y_1^n)}_{(2)} \mathrm{d}x_{n+1}\mathrm{d}x_1^n\mathrm{d}\theta,$$

VB methods can produce an estimate of $p(Y_{n+1}|y_1^n)$ by approximating, in various ways, the two pieces in equation (13) underlined as (1) and (2). VB methods for prediction in SSMs either make explicit use of a variational approximation to the states, $\widehat{q}_x$ to replace $p(x_1^n|y_1^n, \theta)$; or use an accurate simulation-based estimate of $p(x_1^n|y_1^n, \theta)$. We now discuss these two approaches in more detail.

A.1.1 Approximation approaches

The VB methods that approximate $p(Y_{n+1}|y_1^n)$ by constructing an approximation to $p(x_{n+1}|x_n, y_1^n, \theta)$ $\times p(x_1^n|y_1^n, \theta)$ all make use of a variational approximation $\widehat{q}_x$ of $p(x_1^n|y_1^n, \theta)$, in addition to using the structure of the state equation. To illustrate this, it is perhaps easiest to consider the case where we seek to estimate (13) by generating values of Y_{n+1} and using as our estimate of $p(Y_{n+1}|y_1^n)$ the kernel density obtained from the simulations. In this way, we can see that simulation of Y_{n+1} requires simulating the following random variables, in sequence:

$$\theta|y_1^n;\ x_1^n|y_1^n, \theta;\ x_{n+1}|x_n, y_1^n, \theta;\ \text{and } Y_{n+1}|x_{n+1}, \theta.$$

More precisely, consider a fixed value of $\theta^{(j)}$ drawn from some variational approximation of $\pi(\theta|y_1^n)$, call it $\widehat{q}_\theta$. Given the realization $\theta^{(j)}$, we simulate $x_1^n|y_1^n, \theta^{(j)}$ from the VB approximation of the states $\widehat{q}_x$. Next, given $x_n^{(j)} \sim \widehat{q}_x$, we can generate $x_{n+1}^{(j)}$ from $p(x_{n+1}|x_n^{(j)}, y_1^n, \theta^{(j)})$ by generating from the transition density of the states, $x_{n+1}^{(j)} \sim \chi_\theta(x_{n+1}, x_n^{(j)})$, and under the draws $x_n^{(j)}$ and $\theta^{(j)}$. Lastly, $Y_{n+1}^{(j)}$ is generated according to the conditional distribution $Y_{n+1}^{(j)} \sim g_\theta(y_{n+1}|x_{n+1}^{(j)})$. While the above steps are simple to implement, the critical point to realize is that since $x_n^{(j)}$ has not been generated from $p(x_1^n|y_1^n, \theta)$, in general $x_{n+1}^{(j)}$ is not a draw from $p(x_{n+1}|x_n, y_1^n, \theta)p(x_1^n|y_1^n, \theta)$. Hence, the draw $Y_{n+1}^{(j)}$ does not correctly reflect the structure of the assumed model, and $Y_{n+1}^{(j)}$ cannot be viewed as being a draw from the exact predictive density in (13).

Notable uses of the above approach to prediction appear in Quiroz *et al.* (2018), Koop and Korobilis (2020) and Chan and Yu (2020). While similar in form and structure, these three specific approaches are distinct in the sense that the each use different methods to construct $\widehat{q}_x$ (in addition to the differences in the construction of $\widehat{q}_\theta$) and thus to generate $x_{n+1}^{(j)}$.

A.1.2 Simulation approaches

As an alternative, one may estimate $p(Y_{n+1}|y_1^n)$ using exact draws of Y_{n+1}, x_{n+1} and x_1^n, conditional on the draw of θ from some $\widehat{q}_\theta$. For example, if draws from the exact posterior of the states, $p(x_1^n|y_1^n, \theta)$, are readily available via an efficient MCMC algorithm, $p(Y_{n+1}|y_1^n)$ can be estimated via the same set of steps as delineated above, apart from $x_n^{(j)}$ being drawn directly from $p(x_1^n|y_1^n, \theta)$, rather than some $\widehat{q}_x$; see, for example, Loaiza-Maya *et al.* (2021). In this case, $x_{n+1}^{(j)}$ is a draw from $p(x_{n+1}|x_n, y_1^n, \theta)p(x_1^n|y_1^n, \theta)$ and, consequently, the draw $Y_{n+1}^{(j)}$ correctly reflects the model structure. Moreover, due to the Markovian nature of (2), posterior draws of the full vector of states x_1^n are not required, only draws of x_n. As such, any forward (particle) filtering method is all that is required to produce draws of x_n that are conditional on the full vector of observations.

B Computational details and additional results: numerical exercise

B.1 Computational details: methods

This section contains detailed discussions on the computational methods used in the numerical examples in Section 3 of the main paper.

B.1.1 Exact Bayes

Denote the two vectors of latent variables as $\mu_1^n = (\mu_1, \ldots, \mu_n)'$ and $h_1^n = (h_1, \ldots, h_n)'$. The exact posterior density is given as

$$\pi(\theta, \mu_1^n, h_1^n | y_1^n) = \frac{p(y_1^n | \mu_1^n, h_1^n, \theta) p(\mu_1^n, h_1^n | \theta) p(\theta)}{p(y_1^n)}, \tag{14}$$

with prior $p(\theta) = p(\bar{\mu})p(\rho_\mu)p(\sigma_\mu)p(\bar{h})p(\rho_h)p(\sigma_h)$, where $\bar{\mu} \sim N(0, 1000)$, $\rho_\mu \sim \mathrm{U}(0,1)$, ${\sigma_\mu}^2 \sim \mathrm{IG}(1.001, 1.001)$, $\bar{h} \sim N(0, 1000)$, $\rho_h \sim \mathrm{U}(0,1)$ and ${\sigma_h}^2 \sim \mathrm{IG}(1.001, 1.001)$. We draw from (14) using an MCMC algorithm. Specifically, the vector h_1^n is generated using the method proposed in Primiceri (2005), while μ_1^n is generated using the forward-filtering backward-sampling method in Carter and Kohn (1994). Given the choice of priors, the parameters $\bar{\mu}$, σ_μ, $\bar{h}$ and σ_h can be generated directly using Gibbs steps. The parameters ρ_μ and ρ_h are generated using a Metropolis-Hastings step with a Gaussian proposal distribution. The corresponding predictive (expressed using obvious notation),

$$p(Y_{n+1} | y_1^n) = \int_\Theta \int_\mathcal{H} \int_\mathcal{M} g_\theta(Y_{n+1} | \mu_{n+1}, h_{n+1}) p(\mu_{n+1}, h_{n+1} | \theta, \mu_1^n, h_1^n) \pi(\theta, \mu_1^n, h_1^n | y_1^n) \mathrm{d}\mu_1^n \mathrm{d}h_1^n \mathrm{d}\theta, \tag{15}$$

is then estimated (via kernel density methods) using the draws of Y_{n+1} obtained conditional on the draws of θ, μ_1^n and h_1^n.

B.1.2 Quiroz et al. (2018)

Re-cast in terms of our simulation design, Quiroz *et al.* (2018) (QNK hereafter) adopt the variational approximation:

$$q_{\widehat{\lambda}}(\theta, \mu_1^n, h_1^n) = q_{\widehat{\lambda}_1}(\theta) q_{\widehat{\lambda}_2}(x_1^n), \tag{16}$$

where $\widehat{\lambda} = \left(\widehat{\lambda}_1', \widehat{\lambda}_2'\right)'$, $x_t = (\mu_t, h_t)'$ and $x_1^n = \left(x_1', \ldots, x_n'\right)'$. The approximations $q_{\widehat{\lambda}_1}(\theta)$ and $q_{\widehat{\lambda}_2}(x_1^n)$ are optimal elements in the variational classes $\mathcal{Q}_1 = \{q_{\lambda_1}(\theta) : \lambda_1 \in \Lambda_1\}$ and $\mathcal{Q}_2 = \{q_{\lambda_2}(\theta) : \lambda_2 \in \Lambda_2\}$, respectively, where the optimization is performed using a stochastic gradient ascent (SGA) algorithm (Bottou, 2010), and the approximation is based on the same prior as specified above. The elements of the first class are Gaussian densities of the form $q_{\lambda_1}(\theta) = \phi_6\left(\theta; \nu_\theta, BB' + \mathrm{diag}(d^2)\right)$, while the elements of the second class

are of the form $q_{\lambda_2}(x_1^n) = \phi_{2n}\left(x_1^n; \nu_x, CC'\right)$, where C is a three diagonal lower triangular matrix, and the subscript on the symbol for the normal pdf, ϕ, denotes the dimension of the density. (For more details on this approximating class see Ong *et al.*, 2018.) Replacing $\pi(\theta, \mu_1^n, h_1^n|y_1^n)$ in (15) by the approximation in (16), the predictive density is then estimated as described in Section A.1.1.

B.1.3 Loaiza-Maya et al. (2021)

Once again translating their method into our setting, Loaiza-Maya *et al.* (2021) (LSND hereafter), in contrast to Quiroz *et al.* (2018), adopt a variational approximation for $\pi(\theta|y_1^n)$ only, exploiting the exact conditional posterior density of the states, $p\left(\mu_1^n, h_1^n|y_1^n, \theta\right)$. As such, the variational approximation takes the form:

$$q_{\widehat{\lambda}}(\theta, \mu_1^n, h_1^n|y_1^n) = q_{\widehat{\lambda}}(\theta) p\left(\mu_1^n, h_1^n|y_1^n, \theta\right), \tag{17}$$

where $q_{\widehat{\lambda}}(\theta)$ is an optimal element in the variational class $\mathcal{Q} = \{q_\lambda(\theta) : \lambda \in \Lambda\}$, once again found via SGA. For $\mathcal{Q}$ the class of multivariate Gaussian densities with a factor structure is employed, so that $q_\lambda(\theta) = \phi_6\left(\theta; \nu, BB' + \text{diag}(d^2)\right)$, and $\lambda = \left(\nu',\ \text{vec}(B)', d'\right)'$. Replacing $\pi(\theta, \mu_1^n, h_1^n|y_1^n)$ in (15) by the approximation in (17) (once again, with the same underlying prior adopted), the predictive density is then estimated as described in Section A.1.2. Generation from $p\left(\mu_1^n, h_1^n|y_1^n, \theta\right)$ is achieved via an MCMC algorithm that sequentially draws from $p\left(\mu_1^n|y_1^n, \theta, h_1^n\right)$ using the method in Carter and Kohn (1994); and then draws from $p\left(h_1^n|y_1^n, \theta, \mu_1^n\right)$ using the approach in Primiceri (2005).

B.1.4 Chan and Yu (2020)

The final VB method we consider is that of Chan and Yu (2020) (CY hereafter). This approach has been designed specifically for (vector) autoregressive models with stochastic volatility (SV) and not for the UCSV model in (12). The SV component(s) is (are) assumed to have random walk dynamics, which are factored into the construction of the VB approximation for the states. In the case of a scalar random variable (and volatility state) the assumed structure is:

$$h_t = h_{t-1} + \sigma_h \eta_t, \quad Y_t = \exp(h_t/2) u_t.$$

Denoting by h_0 the initial condition of the states, and defining $\theta = (\sigma_h, h_0)'$, CY construct an approximation to the exact posterior $p(\theta, h_1^n|y_1^n)$ as:

$$q_{\widehat{\lambda}}(\theta, h_1^n) = q_{\widehat{\lambda}_1}(\sigma_h^2) q_{\widehat{\lambda}_2}(h_0) q_{\widehat{\lambda}_3}(h_1^n),$$

where $q_{\widehat{\lambda}_1}(\sigma_h^2)$, $q_{\widehat{\lambda}_2}(h_0)$ and $q_{\widehat{\lambda}_3}(h_1^n)$ are optimal elements in the variational classes $\mathcal{Q}_1 = \{q_{\lambda_1}(\sigma_h^2) : \lambda_1 \in \Lambda_1\}$, $\mathcal{Q}_2 = \{q_{\lambda_2}(h_0) : \lambda_2 \in \Lambda_2\}$ and $\mathcal{Q}_3 = \{q_{\lambda_3}(h_1^n) : \lambda_3 \in \Lambda_3\}$, respectively. The elements of each class are defined respectively as $q_{\lambda_1}(\sigma_h^2) = \mathcal{IG}(\sigma_h^2; \nu, S)$, $q_{\lambda_2}(h_0) = \phi_1(h_0; \mu_0, s_0^2)$ and $q_{\lambda_3}(h_1^n) = \phi_n(h_1^n; m, \hat{K}^{-1})$. The variational parameters $\lambda_1 = (\nu, S)^{'}$, $\lambda_2 = (\mu_0, s_0^2)^{'}$ and $\lambda_3 = m$, are calibrated to produce the elements in $\mathcal{Q}_1$, $\mathcal{Q}_2$ and $\mathcal{Q}_3$ that minimise the KL divergence from $p(\theta, h_1^n|y_1^n)$. The authors use a coordinate ascent algorithm (Blei *et al.*, 2017) to perform the optimization, while the value of $\hat{K}^{-1}$ can be optimally computed as a deterministic function of λ_1, λ_2, λ_3 and y_1^n. In our implementation of the CY method, the priors are set to $p(\sigma_h^2) = \mathcal{IG}(\sigma_h^2; 1.001, 1.001)$ and $p(h_0) = \phi_1(h_0; 0, 1000)$, where $\mathcal{IG}$ denotes the inverse gamma distribution. The predictive density is then estimated as described in Section A.1.1, with h_1^n playing the role of x_1^n therein.

B.2 Additional numerical results

B.2.1 State inference

This section contains additional details for the inferential state comparison given in Section 5.2. Firstly, the computation times for the different VB methods used in this section are given in Table 3, and demonstrate that exact Bayes and the method of Loaiza-Maya *et al.* (2021) are comparable in terms of computational cost across the different simulation designs. In contrast, the method of Quiroz *et al.* (2018) takes roughly twice as long to implement.

Figures 2 and 3 plot the posteriors for the unknown states under DGP 1 and 3; see Section 5.2 for details regarding the production and interpretation of these plots. Comparable to the results under DGP 2 in Figure 1, we see that the method of Quiroz *et al.* (2018) performs the worst in terms of state inference for $\exp(h_t/2)$ across both DGPs, while the

Table 3: Estimation time (in seconds) required to estimate the UCSV model on a sample of 11000 time points.

	Estimation times in seconds		
	DGP 1	DGP 2	DGP 3
Exact Bayes	69.1643	66.0900	71.9203
LSND	67.4664	68.5716	71.6248
QNK	125.3298	130.4489	120.8658

method of Loaiza-Maya *et al.* (2021) performs similarly to exact Bayes. For inference on the time-varying mean, μ_t, all methods perform better in general than under DGP 2; in particular, the method of Quiroz *et al.* (2018) appears (visually) to produce much more accurate inferences under DGPs 1 and 3 than in the case of DGP2.

B.2.2 Predictive performance

Herein, we present the results of our predictive analysis for the additional samples sizes referenced in Section 5.3 of the main paper. Table 4 contains results for 100 out-of-sample evaluations and Table 5 results for 1000 out-of-sample evaluations.

B.3 Additional details: scoring rules

In the simulation exercises we have considered five different forms of positively-oriented scoring rules to measure predictive accuracy. To express each of these scoring rules, denote as $P\left(Y_{n+1}|y_1^n\right)$ the predictive distribution associated with the Bayesian predictive density $p(Y_{n+1}|y_1^n)$.

The first scoring rule that we consider is the logarithmic score (LS), which is given by

$$S_{\text{LS}}(P\left(Y_{n+1}|y_1^n\right), y_{n+1}) = \ln p\left(y_{n+1}|y_1^n\right). \tag{18}$$

This score is favourable to predictive distributions that assign high probability mass to the realised value y_{n+1}.

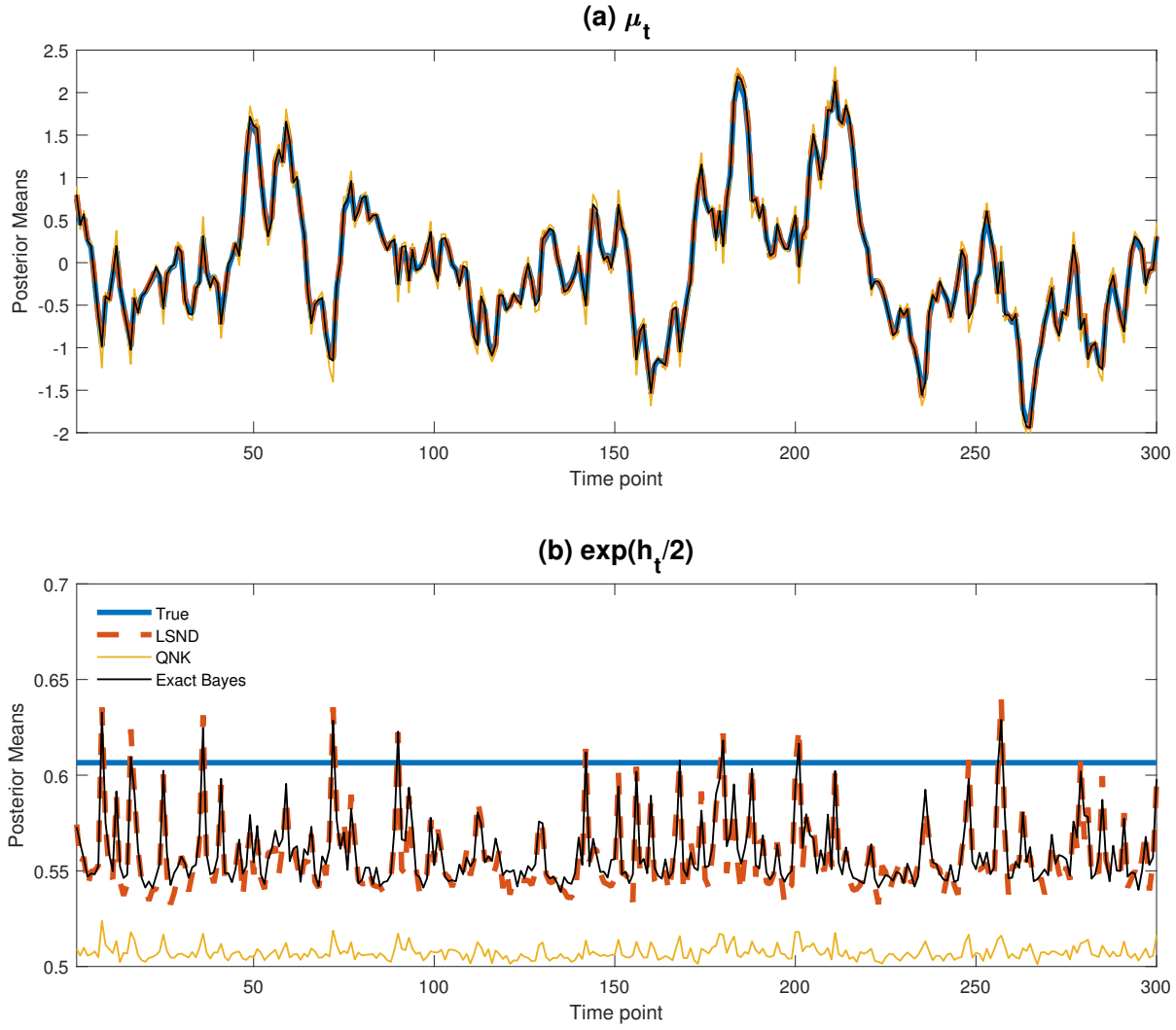

Figure 2: Posterior means of the latent states, under DGP 1, over the first 300 time points. Panel (a) plots the posterior mean for μ_t. The red, gold and black lines plot, respectively, the posterior means based on the LSND, QNK and exact Bayes approaches. The blue line plots the posterior mean that conditions on the true parameters. In the case of the conditional standard deviation, $\exp(h_t/2)$, conditional on θ_0, $\exp(h_t/2)$ is constant, and, thus, so is the posterior mean. Panel (b) presents corresponding results for $\exp(h_t/2)$.

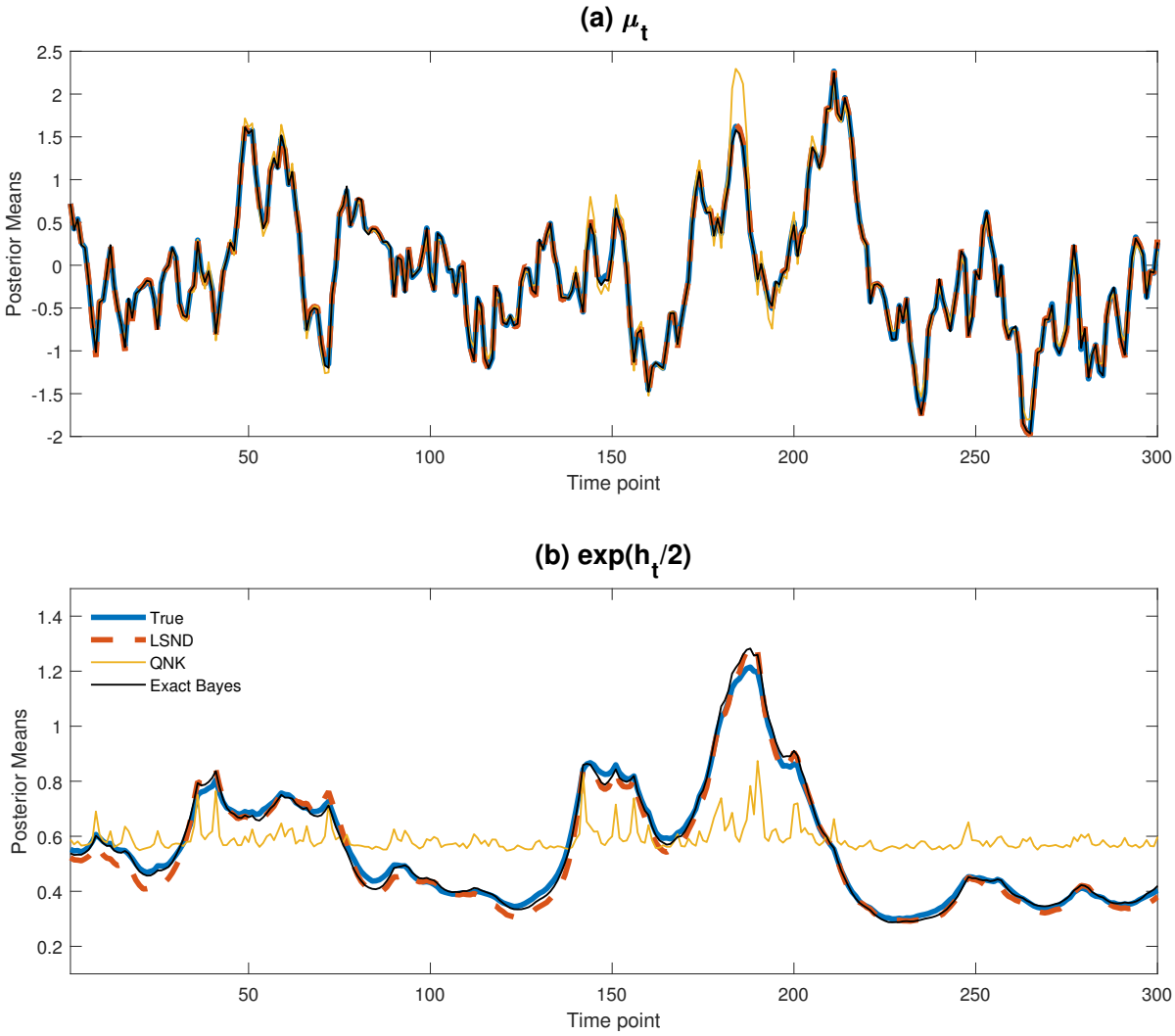

Figure 3: Posterior means of the latent states, under DGP 3, over the first 300 time points. Panel (a) plots the posterior mean for μ_t. The red, gold and black lines plot, respectively, the posterior means based on the LSND, QNK and exact Bayes approaches. The blue line plots the posterior mean that conditions on the true parameters. Panel (b) presents corresponding results for $\exp(h_t/2)$.

Table 4: Predictive performance of competing Bayesian approaches: exact Bayes, LSND and CY and QNK. The column labels indicate the out-of-sample predictive performance measure while the row labels indicate the predictive method. 'True DGP' indicates the productive results that condition on the true parameters. Panels A, B and C correspond to the results for DGP 1, 2 and 3, respectively. The average predictive measures in this table were computed using 100 out-of-sample evaluations.

Panel A: DGP 1	LS	CLS-10%	CLS-20%	CLS-80%	CLS-90%	CRPS	TWCRPS	MSIS
True DGP	-1.206	-0.410	-0.647	-0.348	-0.166	-0.453	-0.134	-3.849
Exact Bayes	-1.210	-0.416	-0.651	-0.349	-0.166	-0.456	-0.136	-3.894
LSND	-1.210	-0.415	-0.648	-0.352	-0.170	-0.455	-0.135	-3.929
CY	-	-	-	-	-	-	-	-
QNK	-1.212	-0.419	-0.649	-0.354	-0.170	-0.454	-0.136	-3.999
Panel B: DGP 2	LS	CLS-10%	CLS-20%	CLS-80%	CLS-90%	CRPS	TWCRPS	MSIS
True DGP	-1.132	-0.295	-0.488	-0.568	-0.332	-0.428	-0.128	-4.417
Exact Bayes	-1.129	-0.300	-0.495	-0.551	-0.315	-0.429	-0.128	-4.339
LSND	-1.144	-0.294	-0.489	-0.570	-0.333	-0.430	-0.128	-4.250
CY	-1.159	-0.314	-0.506	-0.564	-0.321	-0.432	-0.130	-4.583
QNK	-1.151	-0.311	-0.504	-0.562	-0.322	-0.432	-0.129	-4.762
Panel C: DGP 2	LS	CLS-10%	CLS-20%	CLS-80%	CLS-90%	CRPS	TWCRPS	MSIS
True DGP	-1.182	-0.391	-0.569	-0.365	-0.197	-0.452	-0.132	-4.665
Exact Bayes	-1.185	-0.401	-0.576	-0.351	-0.187	-0.451	-0.133	-4.632
LSND	-1.188	-0.399	-0.574	-0.359	-0.197	-0.452	-0.133	-4.634
CY	-	-	-	-	-	-	-	-
QNK	-1.218	-0.415	-0.605	-0.368	-0.201	-0.455	-0.135	-5.119

Table 5: Predictive performance of competing Bayesian approaches: exact Bayes, LSND and CY and QNK. The column labels indicate the out-of-sample predictive performance measure while the row labels indicate the predictive method. 'True DGP' indicates the productive results that condition on the true parameters. Panels A, B and C correspond to the results for DGP 1, 2 and 3, respectively. The average predictive measures in this table were computed using 1000 out-of-sample evaluations.

Panel A: DGP 1	LS	CLS-10%	CLS-20%	CLS-80%	CLS-90%	CRPS	TWCRPS	MSIS
True DGP	-1.243	-0.320	-0.524	-0.495	-0.273	-0.473	-0.143	-3.928
Exact Bayes	-1.246	-0.319	-0.524	-0.497	-0.273	-0.474	-0.143	-3.935
LSND	-1.247	-0.319	-0.522	-0.498	-0.274	-0.474	-0.143	-3.942
CY	-	-	-	-	-	-	-	-
QNK	-1.246	-0.319	-0.523	-0.498	-0.273	-0.474	-0.143	-3.934
Panel B: DGP 2	LS	CLS-10%	CLS-20%	CLS-80%	CLS-90%	CRPS	TWCRPS	MSIS
True DGP	-1.261	-0.379	-0.595	-0.593	-0.403	-0.482	-0.146	-4.477
Exact Bayes	-1.268	-0.382	-0.600	-0.593	-0.405	-0.483	-0.146	-4.504
LSND	-1.273	-0.382	-0.600	-0.598	-0.409	-0.484	-0.146	-4.531
CY	-1.284	-0.386	-0.604	-0.601	-0.410	-0.486	-0.147	-4.623
QNK	-1.274	-0.385	-0.602	-0.596	-0.409	-0.484	-0.146	-4.597
Panel C: DGP 3	LS	CLS-10%	CLS-20%	CLS-80%	CLS-90%	CRPS	TWCRPS	MSIS
True DGP	-1.329	-0.324	-0.548	-0.558	-0.313	-0.517	-0.156	-4.752
Exact Bayes	-1.334	-0.327	-0.551	-0.557	-0.312	-0.518	-0.157	-4.746
LSND	-1.338	-0.328	-0.552	-0.559	-0.314	-0.518	-0.157	-4.773
CY	-	-	-	-	-	-	-	-
QNK	-1.345	-0.332	-0.559	-0.566	-0.322	-0.519	-0.157	-4.987

The second type of scoring rule that we consider is the censored logarithm score (CS) introduced by Diks *et al.* (2011). This rule is defined as

$$S_{\text{CS}}(P\left(Y_{n+1}|y_1^n\right), y_{n+1}) = \ln p\left(y_{n+1}|y_1^n\right) I\left(y_{n+1} \in A\right) + \left[\ln \int_{A^c} p\left(y|y_1^n\right) dy\right] I\left(y_{n+1} \in A^c\right). \tag{19}$$

This score rewards predictive accuracy over the region of interest A (with A^c indicating the complement of this region). Here we report results solely for A defining the lower and upper tail of the predictive distribution, as determined respectively by the 10%, 20%, 80% and 90% quantiles of the empirical distribution of y_t. We label these scores as CS-10%, CS-20%, CS-80% and CS-90% .

The third scoring rule is the continuously ranked probability score (CRPS) proposed by Gneiting and Raftery (2007) and defined as

$$S_{\text{CRPS}}\left[P(Y_{n+1}|y_1^n), y_{n+1}\right] = -\int_{-\infty}^{\infty} \left[P\left(y|y_1^n\right) - I(y \geq y_{n+1})\right]^2 dy. \tag{20}$$

The CRPS is sensitive to distance, rewarding the assignment of high predictive mass near to the realised value of y_{n+1}.

The fourth scoring rule is the left tail weighted CRPS (TWCRPS) proposed in Gneiting and Ranjan (2011), which is defined as

$$S_{\text{TWCRPS}}\left[P(Y_{n+1}|y_1^n), y_{n+1}\right] = -\int_0^1 2\left[I(P^{-1}(\alpha|y_1^n) \geq y_{n+1}) - \alpha\right]\left[P^{-1}(\alpha|y_1^n) - y_{n+1}\right](1-\alpha)^2 d\alpha. \tag{21}$$

This score penalises more heavily longer distances to realised values that are observed in the left tail.

The last score that we consider is the interval score (IS) proposed in Gneiting and Raftery (2007). The IS formula is defined over the $100\left(1-\alpha\right)\%$ prediction interval, and given by

$$\begin{aligned} S_{\text{IS}}\left[P(Y_{n+1}|y_1^n), y_{n+1}\right] = -\Bigg\{ & u_{n+1} - l_{n+1} + \frac{2}{\alpha}\left(l_{n+1} - y_{n+1}\right)\mathbf{1}\{y_{n+1} < l_{n+1}\} \\ & + \frac{2}{\alpha}\left(y_{n+1} - u_{n+1}\right)\mathbf{1}\{y_{n+1} > u_{n+1}\}\Bigg\}, \end{aligned}$$

where l_{n+1} and u_{n+1} denote the $100\left(\frac{\alpha}{2}\right)\%$ and $100\left(1-\frac{\alpha}{2}\right)\%$ predictive quantile, respectively. This score rewards high predictive accuracy of the $100\left(1-\alpha\right)\%$ predictive interval with $0 < \alpha < 1$. In this paper we set $\alpha = 0.05$.

C Technical results

C.1 Proofs of main results

Proof of Lemma 3.1. The proof follows a modification of the standard arguments; see, e.g., Theorem 3.2 in Pakes and Pollard (1989). Fix $\epsilon > 0$. By continuity of $\theta \mapsto H(\theta)$, there exists $\delta > 0$ such that

$$\Pr\left[d(\widehat{\theta}_n, \theta_0) \geq \epsilon\right] \leq \Pr\left[H(\theta_0) - H(\widehat{\theta}_n) \geq \delta\right],$$

where $H(\theta) = \operatorname{plim}_n \ell_n(\theta)$, $\ell_n(\theta) = \frac{1}{n}\log p_\theta(y_1^n)$ and θ_0 satisfies $H(\theta_0) \geq \sup_{\theta\in\Theta} H(\theta)$. The stated result then follows if the RHS is $o(1)$. Since $\hat{\kappa}_n := \Upsilon_n(\widehat{\theta}_n)/n \geq 0$ for all n,

$$H(\theta_0) - H(\widehat{\theta}_n) \leq H(\theta_0) - H(\widehat{\theta}_n) + \hat{\kappa}_n \leq 2\sup_{\theta\in\Theta}|H(\theta) - \ell_n(\theta)| + \ell_n(\theta_0) - \ell_n(\widehat{\theta}_n) + \hat{\kappa}_n$$

By Assumption 3.1, the first term is $o_p(1)$, and we can concentrate on the second term. From the definition of $\widehat{\theta}_n$, and since $0 < p(\theta) < \infty$ for all θ,

$$\left[\ell_n(\theta_0) - \kappa_n + \frac{1}{n}\log p(\theta_0)\right] = [\ell_n(\theta_0) - \kappa_n] + o(1) \leq \left[\ell_n(\widehat{\theta}_n) - \hat{\kappa}_n + \frac{1}{n}\log p(\widehat{\theta}_n)\right] = \left[\ell_n(\widehat{\theta}_n) - \hat{\kappa}_n\right] + o(1).$$

Therefore,

$$\ell_n(\theta_0) - \ell_n(\widehat{\theta}_n) + \hat{\kappa}_n = [\ell_n(\theta_0) - \kappa_n] - [\ell_n(\widehat{\theta}_n) - \hat{\kappa}_n] + \kappa_n \leq o(1) + \kappa_n.$$

Conclude that $H(\theta_0) - H(\widehat{\theta}_n) \leq o_p(1)$ if $\kappa_n = o_p(1)$. □

Proof of Lemma 3.3. The proof follows along the same lines used to prove results for generalized posteriors. See, in particular, Chernozhukov and Hong (2003), Miller (2021), and Syring and Martin (2020).

Define $\Pi_n(\Theta) := \int_\Theta \exp\left\{\widehat{L}_n(\theta)\right\} p(\theta)\mathrm{d}\theta$ and recall that, by hypothesis, for all $n \geq 1$, $\Pi_n(\Theta) < \infty$. Fix $\epsilon > 0$, and let $A_\epsilon := \{\theta : d(\theta, \theta_\star) > \epsilon\}$. For any $\delta > 0$,

$$\begin{aligned}
\widehat{Q}(A_\epsilon|y_1^n) = \frac{\Pi_n(A_\epsilon)}{\Pi_n(\Theta)} &= \frac{\Pi_n(A_\epsilon)\exp\left\{-\widehat{L}(\theta_\star) + n\delta\right\}}{\Pi_n(\Theta)\exp\left\{-\widehat{L}_n(\theta_\star) + n\delta\right\}} \\
&= \frac{\int_{A_\epsilon}\exp\left\{-\widehat{L}(\theta_\star) + n\delta\right\}\exp\left\{\widehat{L}_n(\theta)\right\}p(\theta)\mathrm{d}\theta}{\int_\Theta \exp\left\{-\widehat{L}_n(\theta_\star) + n\delta\right\}\exp\left\{\widehat{L}_n(\theta)\right\}p(\theta)\mathrm{d}\theta} \\
&= \frac{N_n}{D_n}.
\end{aligned}$$

We treat the numerator and denominator separately.

Write the numerator as

$$N_n = \int_{A_\epsilon} \exp\left\{n\left[\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_\star)/n + \delta\right]\right\} p(\theta)\mathrm{d}\theta.$$

Considering $\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_\star)/n$, we have that

$$\begin{aligned}
\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_\star)/n &\leq 2 \sup_{\theta\in\Theta,\lambda\in\Lambda} |\mathcal{L}_n(\theta,\lambda)/n - \mathcal{L}(\theta,\lambda)| + \mathcal{L}[\theta, \widehat{\lambda}_n(\theta)] - \mathcal{L}[\theta_\star, \widehat{\lambda}_n(\theta_\star)] \\
&\leq o_p(1) + \left\{\mathcal{L}[\theta, \widehat{\lambda}_n(\theta)] - \mathcal{L}[\theta, \lambda(\theta)]\right\} - \left\{\mathcal{L}[\theta_\star, \widehat{\lambda}_n(\theta_\star)] - \mathcal{L}[\theta_\star, \lambda(\theta_\star)]\right\} \\
&\quad + \mathcal{L}[\theta, \lambda(\theta)] - \mathcal{L}[\theta_\star, \lambda(\theta_\star)] \\
&\leq o_p(1) + \mathcal{L}[\theta, \lambda(\theta)] - \mathcal{L}[\theta_\star, \lambda(\theta_\star)] \\
&\leq o_p(1) - \delta
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second from Assumption 3.2(2.b), and the third follows from consistency of $\widehat{\lambda}_n(\theta)$, uniformly over θ, Assumption 3.2 (1), and the last follows from the identification condition in Assumption 3.2(2.a). Thus for any $\epsilon > 0$,

$$\liminf_{n\to\infty} \Pr\left[\sup_{\theta: d(\theta,\theta_\star)>\epsilon} \frac{1}{n}\left\{\widehat{L}_n(\theta) - \widehat{L}_n(\theta_\star)\right\} \leq -\delta\right] = 1.$$

Therefore, for any $\theta \in A_\epsilon$,

$$\frac{1}{n}\left\{\widehat{L}_n(\theta) - \widehat{L}_n(\theta_\star)\right\} + \delta \leq 0,$$

with probability converging to one (wpc1), so that for all n large enough

$$\exp\left\{n\left[\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_\star)/n + \delta\right]\right\} \leq 1.$$

Consequently, wpc1,

$$N_n = \int_{A_\epsilon} \exp\left\{n\left[\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_\star)/n + \delta\right]\right\} p(\theta)\mathrm{d}\theta \leq \int_{A_\epsilon} p(\theta)\mathrm{d}\theta \leq 1.$$

To bound the denominator, first define $L(\theta) = \mathcal{L}[\theta, \lambda(\theta)]$ and $G_\delta := \{\theta : L(\theta) - L(\theta_\star) < -\delta/2\}$. For any $\theta \in G_\delta$, by Assumption 3.2(2.b),

$$\left\{\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_\star)/n\right\} + \delta/2 \to L(\theta) - L(\theta_\star) + \delta/2 < 0,$$

wpc1. Thus, for any $\delta > 0$ and any $\theta \in G_\delta$, $\exp\left\{n\left[\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_\star)/n + \delta\right]\right\} \to \infty$ as $n \to \infty$ wpc1. By Fatou's Lemma

$$\begin{aligned}\liminf_{n\to\infty} \exp\left\{-\widehat{L}_n(\theta_\star) + n\delta\right\} \Pi_n(G_\delta) &= \liminf_{n\to\infty} \int_{G_\delta} \exp\left\{n\left[\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_\star)/n + \delta\right]\right\} p(\theta)\mathrm{d}\theta \\ &\geq \liminf_{n\to\infty} \exp\left[n\delta/4\right] \int_{G_\delta} p(\theta)\mathrm{d}\theta.\end{aligned}$$

Since $\int_{G_\delta} p(\theta)\mathrm{d}\theta > 0$ for any $\delta > 0$, by Assumption 3.2(3), the term on the RHS of the inequality diverges as $n \to \infty$. Use the fact that

$$\Pi_n(\Theta) \geq \Pi_n(G_\delta)$$

to deduce that $D_n \to \infty$ as $n \to \infty$ (wpc1). □

Lemma 3.2. The complete data likelihood is proportional to

$$\begin{aligned}p(x_1^n, y_1^n|\theta) =& \{2\pi\sigma_0^2\}^{-n} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{k=1}^{n-1}(x_{k+1} - \rho x_k)^2 - \frac{1}{2\sigma_0^2}\sum_{k=1}^{n}(y_k - \alpha x_k)^2 - \frac{1}{2\sigma_0^2}(x_1)^2\right\} \\ =& \{2\pi\sigma_0^2\}^{-n} \exp\left\{-\frac{1}{2\sigma_0^2}\left[(x_1^n)'\Omega_n(\theta)x_1^n - 2\alpha(y_1^n)'x_1^n + (y_1^n)'y_1^n\right]\right\},\end{aligned}$$

for the matrix

$$\Omega_n(\theta) := \begin{pmatrix} (1+\rho^2+\alpha^2) & -\rho & 0 & \dots & 0 \\ -\rho & (1+\rho^2+\alpha^2) & -\rho & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & -\rho & (1+\alpha^2) \end{pmatrix}.$$

The states can be analytically integrated out, using known results for multivariate normal integrals, to obtain the observed data likelihood $p(y_1^n|\theta)$:

$$p(y_1^n|\theta) = \{2\pi\sigma_0^2\}^{-n}\left[\frac{(2\pi)^n}{|\sigma_0^{-2}\Omega_n(\theta)|}\right]^{1/2} \exp\left\{-\frac{1}{2\sigma_0^2}\left[(y_1^n)'y_1^n - \alpha^2(y_1^n)'\Omega_n(\theta)^{-1}y_1^n\right]\right\},$$

which yields the observable data log-likelihood

$$\log p(y_1^n|\theta) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log(\sigma_0^2) - \frac{1}{2}\log|\Omega(\theta)| + \frac{1}{2\sigma_0^2}\left[\alpha^2(y_1^n)'\Omega_n(\theta)^{-1}y_1^n - (y_1^n)'y_1^n\right].$$

Following Lemma 3.1, consider the infeasible situation where our variational family for θ is

$$\mathcal{Q}_\theta := \{q_\theta : \delta_{\theta_0}(t),\ t \in \Theta\}.$$

Under this choice, consistency follows if $\Upsilon_n(q)/n = \frac{1}{n}\left\{\log p(y_1^n|\theta_0) - \mathcal{L}_n(\theta_0)\right\} = o_p(1)$. In the remainder, we drop the dependence of $q_x(x_1^n|\theta)$ on θ_0 and simply denote $q_x(x_1^n) = q_x(x_1^n|\theta_0)$.

Under the choice of $\mathcal{Q}_x$,

$$\begin{aligned}\mathcal{L}_n(\theta_0) = \int_{\mathcal{X}} q_x(x_1^n)\log\frac{p(x_1^n, y_1^n|\theta)}{q_x(x_1^n)}\mathrm{d}x_1^n =& -n\log 2\pi - n\log\sigma_0^2 \\ &- \frac{1}{2\sigma_0^2}\sum_{k=2}\int (x_k - \rho_0 x_{k-1})^2 q_x(x_k, x_{k-1})\mathrm{d}x_k\mathrm{d}x_{k-1} \\ &- \frac{1}{2\sigma_0^2}\sum_{k=1}\int (y_k - \alpha_0 x_k)^2 q_x(x_{k+1}, x_k)\mathrm{d}x_{k+1}\mathrm{d}x_k \\ &- \frac{1}{2\sigma_0^2}\int (x_1)^2\mathcal{N}(x_1; 0, \sigma_0^2)\mathrm{d}x_1 \\ &- \int q_x(x_1^n)\log\prod_{k=2} q(x_k|x_{k-1})\mathrm{d}x_1^n,\end{aligned}$$

where each of the above individual pieces can be solved explicitly:

$$\begin{aligned}&\int q_x(x_k, x_{k-1})\log q(x_k|x_{k-1})\mathrm{d}x_k\mathrm{d}x_{k-1} = -\frac{1}{2}\log 2\pi - \frac{1}{2} - \frac{1}{2}\log\sigma_0^2(1-\rho_0^2) \\ &\int (x_k - \rho_0 x_{k-1})^2 q_x(x_k, x_{k-1})\mathrm{d}x_k\mathrm{d}x_{k-1} = \sigma_0^2(1-\rho_0^2) \\ &\int (x_1)^2\mathcal{N}(x_1; 0, \sigma_0^2)\mathrm{d}x_1 = \sigma_0^2 \\ &\int (y_k - \alpha x_k)^2 q_x(x_k)\mathrm{d}x_k = y_k^2 + \alpha^2\sigma_0^2,\end{aligned}$$

to obtain

$$\mathcal{L}_n(\theta_0) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_0^2 + \frac{n}{2} + \frac{n}{2}\log(1-\rho_0^2) - \frac{1}{2\sigma_0^2}\{n\sigma_0^2(1-\rho_0^2)\} - \frac{1}{2\sigma_0^2}(y_1^n)'y_1^n - \frac{n}{2}\alpha_0^2$$

Similarly, we have that

$$\log p(y_1^n|\theta_0) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_0^2 - \frac{1}{2}\log|\Omega_n(\theta_0)| + \frac{\alpha_0^2}{2\sigma_0^2}(y_1^n)'\Omega_n(\theta_0)^{-1}y_1^n - \frac{1}{2\sigma_0^2}(y_1^n)'y_1^n$$

and Jensen's Gap is

$$\Upsilon_n(q) = -\frac{1}{2}\log|\Omega(\theta_0)| + \frac{1}{2\sigma_0^2}[\alpha_0^2(y_1^n)'\Omega(\theta_0)^{-1}y_1^n + n\alpha_0^2\sigma_0^2] - \frac{n}{2} - \frac{n}{2}\log(1-\rho_0^2) + \frac{1}{2}\{n(1-\rho_0^2)\}$$

To determine whether $\Upsilon_n(q)/n = o_p(1)$, we must first consider the behavior of the first and second terms in $\Upsilon_n(q)$. For the first term, we note that $|\Omega_n(\theta_0)|$ is a deterministic function of θ_0 and, it can be shown that, see Lemma C.1 for details, if $(1+\alpha_0^2+\rho_0^2)^2-4\rho_0^2 \neq 0$, then, for $a = (1 + \alpha_0^2 + \rho_0^2)$ and $d := \sqrt{a_0^2 - 4\rho_0^2}$,

$$|\Omega_n(\theta_0)| = \frac{1}{d}\left(\left(\frac{a+d}{2}\right)^{n+1} - \left(\frac{a-d}{2}\right)^{n+1}\right),$$

and we can define

$$C_1(\theta_0) := \operatorname*{plim}_{n\to\infty} \log|\Omega_n(\theta_0)|/n.$$

Conversely, if $(1 + \alpha_0^2 + \rho_0^2) - 4\alpha_0^2 = 0$, then

$$|\Omega_n(\theta_0)| = (n+1)(a/2)^n,$$

and we can define $C_1(\theta_0)$ similarly in this case.

Now, consider the second term in $\Upsilon_n(q)$. From the structure of the model, for $\sigma_x^2 = \sigma_0^2/(1-\rho_0^2)$,

$$y_1^n \sim \mathcal{N}(0, M),\ M := \sigma_x^2[(\sigma_0^2/\sigma_x^2)I + V],\ V^{-1} := \begin{pmatrix} (1+\rho_0^2) & -\rho_0 & 0 & \dots & 0 \\ -\rho_0 & (1+\rho_0^2) & -\rho_0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & -\rho_0 & (1+\rho_0^2) \end{pmatrix},$$

so that we can conclude that, for any $n \geq 2$,

$$\mathbb{E}[(y_1^n)'\Omega_n^{-1}(\theta_0)y_1^n] = \mathrm{Tr}\left[\Omega_n(\theta_0)^{-1}M\right],\ \mathrm{Var}\left[(y_1^n)'\Omega_n(\theta_0)^{-1}y_1^n\right] = 2\mathrm{Tr}\left[\Omega_n(\theta_0)^{-1}M\Omega_n(\theta_0)^{-1}M\right].$$

Define $Z_n = (y_1^n)'\Omega_n(\theta_0)^{-1}y_1^n$ and apply Markov's inequality to Z_n to obtain, for any $\epsilon > 0$,

$$\Pr\left(|Z_n - \mathbb{E}[Z_n]| > n\epsilon\right) \leq \mathrm{Var}[Z_n]/(n^2\epsilon^2) = \mathrm{Tr}\left[\Omega_n(\theta_0)^{-1}M\Omega_n(\theta_0)^{-1}M\right]/(n^2\epsilon^2) \tag{22}$$

In addition, for D_1 denoting the first diagonal element of $\Omega_n(\theta_0)^{-1}M\Omega_n(\theta_0)^{-1}M$, and D_n the n-th,

$$\mathrm{Var}[Z_n] = \mathrm{Tr}\left[\Omega_n(\theta_0)^{-1}M\Omega_n(\theta_0)^{-1}M\right] \leq n \cdot \sup_{n\geq 1}\{|D_1|, \dots, |D_n|\}. \tag{23}$$

Define the sequence $c_n := \sup_{n\geq 1}\{|D_1|,\ldots,|D_n|\}$. For any $0 \leq \rho_0^2 < 1$ and $0 \leq |\alpha_0| < M < \infty$, the sequence c_n is non-random and bounded for each n, hence we have that $c_n/n \to 0$. From the boundedness of c_n, apply equations (22) and (23) to conclude that, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{\mathrm{Var}[Z_n]}{(n\epsilon)^2} \leq \lim_{n\to\infty} \frac{\sup_{n\geq 1} |\mathrm{diag}\{\Omega_n(\theta_0)^{-1} M \Omega_n(\theta_0)^{-1} M\}|}{n\epsilon^2} = \lim_{n\to\infty} \frac{c_n}{n\epsilon^2} \to 0.$$

The above argument and equation (22) allow us to conclude that $C_2(\theta_0) := \frac{\alpha_0^2}{2\sigma_0^2} \lim_{n\to\infty} \mathrm{Tr}[\Omega_n(\theta_0)M]/n$ exists and that

$$\operatorname*{plim}_{n\to\infty} \frac{\alpha_0^2}{2\sigma_0^2 n} (y_1^n)' \Omega_n(\theta_0)^{-1} y_1^n = C_2(\theta_0).$$

We are now ready to specialize the above to the two cases of interest.

Case 1: If $\rho_0 = 0$, then $|\Omega_n(\theta_0)| = (1+\alpha_0^2)^n$, and $\log|\Omega_n(\theta_0)| = n\log(1+\alpha_0^2)$. In addition, $M = 2\sigma_0^2 I$ and $\Omega_n(\theta_0)^{-1} = \frac{1}{(1+\alpha_0^2)} I_n$, with I_n the n-dimensional identity matrix, so that $\mathrm{Tr}(\Omega_n(\theta_0)^{-1}M) = n\frac{2\sigma_0^2}{1+\alpha_0^2}$. Therefore, we have that

$$C_1(\theta_0) = \frac{1}{2}\log(1+\alpha_0^2) \text{ and } C_2(\theta_0) = \alpha_0^2/(1+\alpha_0^2).$$

Since $n^{-1}\log p_{\theta_0}(y_1^n) \to_p H(\theta_0) \geq 0$, which minimizes entropy, and since $\Upsilon_n(q) \geq 0$, consequently, VI for α_0 will be consistent iff

$$\operatorname*{plim}_{n\to\infty} \Upsilon_n(q)/n = 0 = -\frac{1}{2}\log(1+\alpha_0^2) + \frac{\alpha_0^2}{1+\alpha_0^2} + \alpha_0^2.$$

Over $0 \leq |\alpha_0| < M$, M finite, the above equation has the unique solution $\alpha_0 = 0$.

Case 2: $\alpha_0 = 0$**.** Similar to the above, since $H(\theta_0)$ is entropy minimizing, and since $\Upsilon_n(q)/n \geq 0$, it must be that $\Upsilon_n(q)/n = o_p(1)$ if VI is to be consistent. However, if $\alpha_0 = 0$, we have that

$$\Upsilon_n(q) = -\frac{1}{2}\log|\Omega_n(\theta_0)| - \frac{n}{2} - \frac{n}{2}\log(1-\rho_0^2) + \frac{n}{2}(1-\rho_0^2).$$

Apply Lemma C.1 in the supplementary material to obtain $|\Omega_n(\theta_0)|$ with $a = (1+\rho_0^2)$, and $b = c = -\rho_0$. In particular, use the fact, for $0 \leq |\rho_0| < 1$, $d = \sqrt{a^2 - 4bc} = 1 - \rho_0^2$, and note that $a + d = 2$ and $a - d = 2\rho_0^2$, which allows us to specialize the general result in Lemma

C.1 as

$$|\Omega_n(\theta_0)| = \frac{1}{(1-\rho_0^2)}\left\{\left[1-\left(\rho_0^2\right)^n\right] - \rho_0^2\left[1-\left(\rho_0^2\right)^{n-1}\right]\right\} = \frac{1}{1-\rho_0^2}\left\{1-\rho^2-\rho_0^{2n}+\rho^2(\rho^{2(n-1)})\right\}$$
$$= 1$$

Conclude that VI consistent is iff

$$\operatorname*{plim}_{n\to\infty} \Upsilon_n(q)/n = 0 = -\frac{1}{2}\log(1+\rho_0^2) - \frac{1}{2} - \frac{1}{2}\log(1-\rho_0^2) + \frac{1}{2}(1-\rho_0^2).$$

The only solution to the above equation is $\rho_0 = 0$.

□

Proof of Lemma 3.4. Recall the complete data likelihood from the proof of Lemma 3.2:

$$p(x_1^n, y_1^n|\theta) = \{2\pi\sigma_0^2\}^{-n}\exp\left\{-\frac{1}{2\sigma_0^2}\sum_{k=1}^{n-1}(x_{k+1}-\rho_0 x_k)^2 - \frac{1}{2\sigma_0^2}\sum_{k=1}^{n}(y_k-\alpha x_k)^2 - \frac{1}{2\sigma_0^2}(x_1)^2\right\}$$
$$= \{2\pi\sigma_0^2\}^{-n}\exp\left\{-\frac{1}{2\sigma_0^2}\left[(x_1^n)'\Omega_n(\theta)x_1^n - 2\alpha(y_1^n)'x_1^n + (y_1^n)'y_1^n\right]\right\}.$$

In this case, we calculate

$$\mathcal{L}_n(\theta,\lambda) = \int_{\mathcal{X}} q_\lambda(x_1^n)\log\frac{p(x_1^n, y_1^n|\theta)}{q_\lambda(x_1^n)}\mathrm{d}x_1^n$$
$$= \int_{\mathcal{X}} q_\lambda(x_1^n)\log p(x_1^n, y_1^n|\theta)\mathrm{d}x_1^n - \int_{\mathcal{X}} q_\lambda(x_1^n)\log q_\lambda(x_1^n)\mathrm{d}x_1^n.$$

Writing the two terms as $\mathcal{L}_{1,n}(\theta,\lambda)$ and $\mathcal{L}_{2,n}(\theta,\lambda)$, let us focus on the first term. This can be rewritten as

$$\mathcal{L}_{1,n}(\theta,\lambda) = -n\log 2\pi - \frac{1}{2}\mathbb{E}_{x_1^n}[(x_1^n)'\Omega_n(\theta)x_1^n] - \alpha(y_1^n)'\mathbb{E}_{x_1^n}[x_1^n] - \frac{1}{2}(y_1^n)'y_1^n$$
$$= -n\log 2\pi - \frac{1}{2}(y_1^n)'y_1^n - \frac{1}{2}\mathrm{Tr}[\Omega_n(\theta)\nu(\lambda)\Phi_n(\lambda)]$$
$$= -n\log 2\pi - \frac{1}{2}(y_1^n)'y_1^n - \frac{\nu(\lambda)}{2}\mathrm{Tr}[\Omega_n(\theta)\Phi_n(\lambda)],$$

where the second equation comes from the fact that $\mathbb{E}_{x_1^n}[x_1^n] = 0$, under $q_\lambda(x_1^n)$, and properties of quadratic forms, and the third follows from linearity of $textTr(\cdot)$. Tedious algebraic calculations show that

$$\mathrm{Tr}[\Omega_n(\theta)\Phi_n(\lambda)] = (n-1)(1+\alpha^2+\rho^2-\rho\lambda) + (1+\alpha^2-\rho\lambda)$$

and so we obtain

$$\mathcal{L}_{1,n}(\theta,\lambda) = -n\log 2\pi - \frac{1}{2}(y_1^n)'y_1^n - \frac{(n-1)\nu(\lambda)}{2}(1+\alpha^2+\rho^2-\rho\lambda) - \frac{\nu(\lambda)}{2}(1+\alpha^2-\rho\lambda)$$

The second term can be written as

$$\begin{aligned}\mathcal{L}_{2,n}(\theta,\lambda) &= \int_{\mathcal{X}} q_\lambda(x_1^n)\log q_\lambda(x_1^n)\mathrm{d}x_1^n = \mathbb{E}_{x_1^n}[\log q_\lambda(x_1^n)] = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\nu(\lambda)\Phi_n(\lambda)| - n/2 \\ &= -\frac{n}{2}\log 2\pi - \frac{n}{2} - \frac{1}{2}\log\frac{\sigma_0^2}{(1-\lambda_\rho^2)} \\ &= -\frac{n}{2}\log 2\pi - \frac{n}{2} - \frac{n}{2}\log\sigma_0^2 + \frac{1}{2}\log(1-\lambda_\rho^2)\end{aligned}$$

where we have used the fact that for the matrix $\Phi_n(\lambda)$, $|\Phi_n(\lambda)| = (1-\lambda^2)^{n-1}$, so that we have

$$|\nu(\lambda)\Phi_n(\lambda)| = \nu(\lambda)^n|\Phi_n(\lambda)| = \frac{\lambda_\sigma^n}{(1-\lambda_\rho^2)^n}(1-\lambda^2)^{n-1}.$$

Dividing these terms by n, and taking $n\to\infty$ yields the following limit criterion

$$\mathcal{L}(\theta,\lambda) = -\log 2\pi - \frac{1}{2}\mathbb{E}_{\theta_0}\left[(y_1^n)'(y_1^n)\right] - \frac{\nu(\lambda)}{2}(1+\alpha^2+\rho^2-\rho\lambda).$$

Differentiating $\mathcal{L}(\theta,\lambda)$ with respect to λ and solving yields two solutions:

$$\lambda(\theta) = \begin{cases} \frac{\alpha^2+\rho^2-\sqrt{(\alpha^2+\rho^2+\rho+1)(\alpha^2+\rho^2-\rho+1)}+1}{\rho} \\ \frac{\alpha^2+\rho^2+\sqrt{(\alpha^2+\rho^2+\rho+1)(\alpha^2+\rho^2-\rho+1)}+1}{\rho} \end{cases} \tag{24}$$

It can be shown that the first solution is the maximum, while the second is the minimum. Using the function $\lambda(\theta)$, the concentrated objective function is

$$\mathcal{L}[\theta,\lambda(\theta)] = \frac{\sqrt{(\alpha^2+\rho^2+\rho+1)\ (\alpha^2+\rho^2-\rho+1)}}{2\left(\frac{\left(\alpha^2+\rho^2-\sqrt{(\alpha^2+\rho^2+\rho+1)(\alpha^2+\rho^2-\rho+1)}+1\right)^2}{\rho^2}-1\right)}. \tag{25}$$

On the compact space $[\underline{\alpha},\overline{\alpha}]\times[\underline{\rho},\overline{\rho}]$, the function $\mathcal{L}[\theta,\lambda(\theta)]$ is continuous and bounded. Hence, by the extreme value theorem $\mathcal{L}[\theta,\lambda(\theta)]$ achieves its maximum at some point in $[\underline{\alpha},\overline{\alpha}]\times[\underline{\rho},\overline{\rho}]$.

To prove that $\theta^\star = (0,\underline{\rho})'$, we can consider the following two cases:

1. $\alpha^\star \geq 0$, and $\rho^\star = \underline{\rho}$;

2. $\rho^\star \geq \underline{\rho}$, and $\alpha^\star = 0$;

Case 1: Take $\rho^\star = \underline{\rho} > 0$, but $\underline{\rho}$ close to zero. Under this choice, we can approximate $\mathcal{L}[\theta, \lambda(\theta)]$ as

$$\mathcal{L}[\theta, \lambda(\theta)]|_{\rho=\rho^\star} \approx \frac{\sqrt{(\alpha^2+1)\ (\alpha^2+1)}}{2\left(\frac{\left(1+\alpha^2-\sqrt{(\alpha^2+1)\,(\alpha^2+1)}\right)^2}{\underline{\rho}^2} - 1\right)} = -\frac{1}{2}(1+\alpha^2). \tag{26}$$

The RHS of the above is a decreasing function of α, so that its maximum over $[0, \overline{\alpha}]$ is attained at $\alpha = 0$.

Case 2: Taking $\alpha^\star = 0$, we have

$$\mathcal{L}[\theta, \lambda(\theta)]|_{\alpha=0} = \frac{\sqrt{(\rho^2+\rho+1)\ (\rho^2-\rho+1)}}{\frac{2\left(\rho^2-\sqrt{(\rho^2+\rho+1)\,(\rho^2-\rho+1)}+1\right)^2}{\rho^2} - 2} = \frac{1}{2}\frac{\sqrt{\rho^4+\rho^2+1}}{\left[\frac{(1+\rho^2-\sqrt{\rho^4+\rho^2+1})^2}{\rho^2} - 1\right]}. \tag{27}$$

For $0 < \underline{\rho} \leq \rho \leq \overline{\rho} < 1$, the denominator of the RHS is always larger than -1, and always less than $(2-\sqrt{3})^2 - 1 \approx -.9282$. It can verified (e.g., numerically) that the above function is monotonically decreasing over $[\underline{\rho}, \overline{\rho}]$, so that its maximum is attained at $\rho = \underline{\rho}$.

Hence, the maximum of $\mathcal{L}[\theta, \lambda(\theta)]$ over $[0, \overline{\alpha}] \times [\underline{\rho}, \overline{\rho}]$ is given by $\theta^\star = (0, \underline{\rho})'$.

□

Proof of Corollary 3.1. The result is a direct consequence of Lemma 3.4 and known results. For two multivariate normal distributions (with the same dimension $d \geq 1$) $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$, the KL divergence is

$$\mathrm{KL}\left[N(\mu_1, \Sigma_1)||N(\mu_2, \Sigma_2)\right] = \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - d + \mathrm{tr}\left\{\Sigma_2^{-1}\Sigma_1\right\} + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right] \tag{28}$$

For known θ, the posterior of the latent states can be obtained via the Kalman filter, and has the known form:

$$\pi(x_n|y_1^n, \theta_0) = N\{x_n; x_{n|n}(\theta_0), P_{n|n}(\theta_0)\},$$

where $x_{n|n}$ and $P_{n|n}$ are known functions that are obtained from the Kalman filter recursions. Let $\lambda_\star = \lambda(\theta_\star)$, and note that since $q_{\lambda_\star}(x_1^n)$ is multivariate Gaussian with mean

0 and variance $\Phi_n(\lambda_\star)$, we immediately obtain that the marginal density of x_n under the variational family is $q_{\lambda_\star}(x_n) = N\{x_n; 0, \sigma_0^2/(1-\lambda_\star^2)\}$. Applying equation (28) then yields

$$\begin{aligned}
\mathrm{KL}\left[\pi(x_n|y_1^n, \theta_0)||q_{\lambda_\star}(x_n)\right] &= \mathrm{KL}\left[N\{x_{n|n}(\theta_0), P_{n|n}(\theta_0)\}||N\{0, \sigma_0^2/(1-\lambda_\star^2)\}\right] \\
&= \frac{1}{2}\left[\log\frac{\sigma_0^2/(1-\lambda_\star^2)}{P_{n|n}(\theta_0)} - 1 + \frac{P_{n|n}(\theta_0)}{\sigma_0^2/(1-\lambda_\star^2)} + \frac{x_{n|n}^2(\theta_0)}{\sigma_0^2/(1-\lambda_\star^2)}\right] \\
&= \frac{1}{2}\left[-\log([1-\lambda_\star^2]P_{n|n}(\theta_0)) - 1 + [1-\lambda_\star^2]\left[P_{n|n}(\theta_0) + x_{n|n}^2\right]\right] \\
&= \frac{1}{2}\left[\log\frac{\exp\{[1-\lambda_\star^2]P_{n|n}(\theta_0)\}}{[1-\lambda_\star^2]P_{n|n}(\theta_0)} - 1 + [1-\lambda_\star^2]x_{n|n}^2(\theta_0)\right]
\end{aligned}$$

where the third inequality follows from the fact that $\sigma_0^2 = 1$, and the last from re-arranging terms.

For any $y \geq C > 0$, differentiating the function $\exp(y)/y$ and solving for its zero yields

$$\frac{\exp(y)}{y}[1-(1/y)] = 0 \iff [1-(1/y)] = 0$$

which yields the unique solution $y = 1$ on $y \geq C > 0$. A second round of differentiation shows that this function is positive at $y = 1$. Hence, $\exp(y)/y$ attains a unique minimum at $y = 1$, and we have $\exp(y)/y \geq \exp(1)$ for all $y \geq C > 0$. Consequently, $\log\{\exp(y)/y\} \geq 1$ and we have shown that the first term in the KL divergence is positive when $P_{n|n}(\theta_0) > 0$. Since $[1-\lambda_\star^2]x_{n|n}^2(\theta_0) \geq 0$, the stated result follows.

□

C.2 Additional lemmas

Lemma C.1. *Let*

$$\Omega_n := \begin{pmatrix} a & c & 0 & \dots & 0 \\ b & a & c & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & b & 1 \end{pmatrix}, \; a > 0, \; a^2 - 4bc \neq 0.$$

Then, for $d = \sqrt{a^2 - 4bc}$,

$$|\Omega_n| = \frac{1}{d}\left[\left(\frac{a+d}{2}\right)^n - \left(\frac{a-d}{2}\right)^n\right] - bc\frac{1}{d}\left[\left(\frac{a+d}{2}\right)^{n-1} - \left(\frac{a-d}{2}\right)^{n-1}\right].$$

Proof. The determinant of tridiagonal matrices satisfy the following recurrence relationship: for $f_k = |\Omega_k|$, with Ω_k denoting the $k \times k$ matrix, $1 < k \leq n$,

$$f_n = a_n f_{n-1} - c_{n-1} b_{n-1} f_{n-2},$$

where $f_0 = 0$ and $f_1 = 1$, and c_k, b_k refer to the elements above and below, respectively, the diagonal term a_n. In this case, this relationship implies that $f_n = |\Omega_n|$ satisfies

$$f_n = f_{n-1} - cbf_{n-2}.$$

However, note that, for an $1 \leq k < n$, f_k is actually a $k \times k$ dimensional Toeplitz matrix. Applying a Laplace expansion to f_k twice yields the linear homogenous recurrence equation

$$f_k = af_{k-1} - bcf_{k-2},$$

which has characteristic polynomial $p(x) = x^2 - ax + bc$ that admits two solutions

$$x = \frac{a \pm \sqrt{a^2 - 4bc}}{2}$$

Under the condition that $a^2 - 4bc \neq 0$, the roots are distinct and we have that

$$f_k = c_1 \left(\frac{a + \sqrt{a^2 - 4bc}}{2} \right)^k + c_2 \left(\frac{a - \sqrt{a^2 - 4bc}}{2} \right)^k,$$

for some c_1 and c_2 that satisfy the initial conditions of the recurrent relation. In particular, we have that $f_1 = a$, and $f_2 = a^2 - bc$, so that

$$a^2 - bc = a(f_1) - bc(f_0) = a^2 - bc(f_0),$$

which implies that $d_0 = 1$. Consequently, $c_1 + c_2 = 1$. Letting $d = \sqrt{a^2 - 4bc}$, we see that the case of $k = 1$ implies

$$\begin{aligned}
2a = k_1(a + d) + k_2(a - d) = a + (c_1 - c_2)d &\implies c_1 = c_2 + a/d \\
&\implies 1 = 2c_2 + a/d \\
&\implies c_2 = \frac{d - a}{2d} = -\frac{1}{d}\frac{(a - d)}{2} \\
&\implies c_1 = \frac{d - a + 2a}{2d} = \frac{1}{d}\frac{a + d}{2}
\end{aligned}$$

Therefore, we can conclude that

$$f_k = \frac{1}{d}\left[\left(\frac{a+d}{2}\right)^{k+1} - \left(\frac{a-d}{2}\right)^{k+1}\right],$$

and we then have closed form expressions for the determinants f_{n-1} and f_{n-2}. Plugging in these definitions

$$\begin{aligned} f_n &= f_{n-1} - bcf_{n-2} \\ &= \frac{1}{d}\left[\left(\frac{a+d}{2}\right)^{n} - \left(\frac{a-d}{2}\right)^{n}\right] - bc\frac{1}{d}\left[\left(\frac{a+d}{2}\right)^{n-1} - \left(\frac{a-d}{2}\right)^{n-1}\right]. \end{aligned}$$

□