A Note on the Accuracy of Variational Bayes in State Space Models: Inference and Prediction

David T. Frazier*, Rubén Loaiza-Maya and Gael M. Martin

Department of Econometrics and Business Statistics, Monash University
and Australian Centre of Excellence in Mathematics and Statistics

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 marked over longer out-of-sample periods. 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.

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 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

*Corresponding author: david.frazier@monash.edu.
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 Sørensen, 1996, Gallant and Tauchen, 1996, Sandmann and Koopman, 1998, Bates, 2006, Ait-Sahalia and Kimmel, 2007, and Ait-Sahalia et al., 2020, 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, 2018; Chan and Yu, 2020; Loaiza-Maya et al., 2021).

In this short 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 issue identified elsewhere in the literature as the ‘incidental parameter problem’ (Neyman and Scott, 1948; Lancaster, 2000). 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. We believe this to be a novel insight into the role of VB in SSMs, and one that can lead to best practice, if heeded. 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.

Throughout the remainder, we make use of the following notational conventions. Generic $p, g$ are used to denote densities, and $\pi$ 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 \(z^n\). For a sequence \(a_n\), the terms \(O_p(a_n)\), \(o_p(a_n)\) and \(\rightarrow_p\) have their usual meaning. Similarly, we let \(\text{plim}_n X_n = c\) denote \(X_n \rightarrow_p c\). 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 numerical results, 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 \((X, \mathcal{F}_X, \mu)\), and \(\{Y_t\}\) is a process taking values in a measure space \((Y, \mathcal{F}_Y, \chi)\), such that, conditional on \(\{X_t\}\), the sequence \(\{Y_t\}\) is independent. The model is formulated through the following state and measurement equations: for a vector of unknown random parameters \(\theta\) taking values in the probability space \((\Theta, \mathcal{F}_\theta, P_\theta)\), where \(P_\theta\) admits the density function \(p_\theta\),

\[
X_t = \phi(X_{t-1}; \theta) + \Sigma(\theta)^{1/2}\epsilon_t, \quad (1)
\]

\[
Y_t = r(X_t, \eta_t; \theta), \quad (2)
\]

where \(\phi(\cdot), \Sigma(\cdot),\) and \(r(\cdot)\) are known functions and \(\{\epsilon_t\}\) and \(\{\eta_t\}\) are independent sequences of i.i.d. random variables independent of the initial value \(X_0\), which is distributed according to the initial measure \(\nu\). The system in (1) and (2) gives rise to the following conditional and transition densities:

\[
Y_t|X_t, \theta \sim g_\theta(y_t|x_t), \quad X_{t+1}|X_t, \theta \sim \chi_\theta(x_{t+1}, x_t),
\]

where \(\chi_\theta(\cdot, \cdot)\) denotes the transition kernel wrt the measure \(\mu\). For simplicity, throughout the remainder we disregard terms dependence on the initial measure \(\nu\) and the invariant measure \(\mu\), when no confusion will result.

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

\[
p_\theta(y^n_1, x^n_1) = \nu(x_1)g_\theta(y_1|x_1)\prod_{p=2}^n \chi_\theta(x_p, x_{p-1})g_\theta(y_p|x_p).
\]

The (average) observed data log-likelihood is thus

\[
\ell_n(\theta) := \frac{1}{n} \log p_\theta(y^n_1) = \frac{1}{n} \log \int p_\theta(y^n_1, x^n_1)dx^n_1, \quad (3)
\]

and the maximum likelihood estimator (MLE) of \(\theta\) is \(\hat{\theta}_n^{\text{MLE}} := \arg\max_{\theta \in \Theta} \ell_n(\theta)\). 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 (1) and (2) define a linear Gaussian state space model (LGSSM). Similarly, for \(p(\theta)\)
denoting the prior density, the exact (marginal) posterior for \( \theta \), defined as

\[
\pi(\theta|y^n_1) = \int \pi(\theta, x^n_1|y^n_1)dx^n_1, \quad \text{where} \quad \pi(\theta, x^n_1|y^n_1) \propto p(y^n_1|x^n_1, \theta)p(x^n_1|\theta)p(\theta),
\]  

(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 \( \theta \) 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 \( \theta \), 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 \( \pi(x^n_1, \theta|y^n_1) \) 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 \( \pi(x^n_1, \theta|y^n_1) \) by minimizing the KL divergence between a family of densities \( Q \), with generic element \( q(x^n_1, \theta) \), and \( \pi \):

\[
KL(q||\pi) = \int q(x^n_1, \theta) \log \frac{q(x^n_1, \theta)}{\pi(x^n_1, \theta|y^n_1)} dx^n_1 d\theta. \tag{5}
\]

Optimizing the KL divergence directly is not feasible since it depends on the unknown \( \pi(x^n_1, \theta|y^n_1) \); the very quantity we are trying to approximate. However, minimizing the KL divergence between \( q \) and \( \pi \) is equivalent to maximizing the so-called variational evidence lower bound (ELBO):

\[
\text{ELBO}(q||\pi) := \int q(x^n_1, \theta) \log \frac{p(y^n_1|x^n_1, \theta)p(x^n_1|\theta)p(\theta)}{q(x^n_1, \theta)} dx^n_1 d\theta, \tag{6}
\]

which we can access. Hence, for a given class \( Q \), we may define the variational approximation as

\[
\hat{q} := \arg\max_{q \in Q} \text{ELBO}(q||\pi).
\]
The standard approach to obtaining \( \hat{q} \) is to consider a class of product distributions

\[
Q = \{ q : q(x^n_1, \theta) = q_\theta(\theta)q_x(x^n_1|\theta) \},
\]

with \( Q \) often restricted to be mean-field, i.e., \( \theta \) independent of \( \theta_j \), \( i \neq j \), and \( x^n_i \) independent of \( \theta \).

Regardless of the variational family adopted, \( \text{KL}(q||\pi) \), and hence \( \text{ELBO}(q||\pi) \), involve both \( \theta \) and \( x^n_i \). The product form of \( Q \) allows us to write:

\[
\text{ELBO}(q||\pi) = \int_\Theta \int_{\mathcal{X}} q_\theta(\theta)q_x(x^n_1|\theta) \log \frac{p(y^n_1|x^n_1, \theta)p(x^n_1|\theta)p(\theta)}{q_\theta(\theta)q_x(x^n_1|\theta)} dx^n_1 d\theta
\]

\[
= \int_\Theta \int_{\mathcal{X}} q_\theta(\theta)q_x(x^n_1|\theta) \log \left[ \frac{p(y^n_1|x^n_1, \theta)p(x^n_1|\theta)}{q_x(x^n_1|\theta)} \right] dx^n_1 d\theta - \text{KL}[q_\theta(\theta)||p(\theta)],
\]

where the last line follows from Fubini’s theorem and the fact that \( q_x(x^n_1|\theta) \), by assumption, is a proper density function, for all \( \theta \). Further, defining

\[
\mathcal{L}_n(\theta) := \int_{\mathcal{X}} q_x(x^n_1|\theta) \log \frac{p(y^n_1|x^n_1, \theta)p(x^n_1|\theta)}{q_x(x^n_1|\theta)} dx^n_1,
\]

by Jensen’s inequality

\[
\log p_\theta(y^n_1) = \log \int_{\mathcal{X}} p_\theta(y^n_1, x^n_1) dx^n_1 = \log \int_{\mathcal{X}} q_x(x^n_1|\theta) \left\{ \frac{p(y^n_1|x^n_1, \theta)p(x^n_1|\theta)}{q_x(x^n_1|\theta)} \right\} dx^n_1 \geq \mathcal{L}_n(\theta).
\]

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

\[
\Upsilon_n(q) := \int_\Theta \{ \log p_\theta(y^n_1) - \mathcal{L}_n(\theta) \} q_\theta(\theta) d\theta,
\]

the \( \text{ELBO}(q||\pi) \) can then be expressed as

\[
\text{ELBO}(q||\pi) = \int_\Theta \log p_\theta(y^n_1)q_\theta(\theta)d\theta - \Upsilon_n(q) - \text{KL}[q_\theta(\theta)||p(\theta)].
\]

This representation decomposes \( \text{ELBO}(q||\pi) \) into three components, two of which only depend on the variational approximation of the global parameters \( \theta \), and a third component, \( \Upsilon_n(q) \), that Yang et al. (2020) refer to as the average (wrt \( q_\theta(\theta) \)) “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 $\theta$. In this case, we can think of the variational family as

$$Q := \{ q : q(\theta, x^n_t) = \delta_0 \times q_x(x^n_t|\theta) \},$$

where $\delta_0$ is the Dirac delta function at $\theta$, and we can then write

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

where we abuse notation and represent functions with arguments $\delta_0$ only by the parameter value $\theta \in \Theta$, and also make use of the short-hand notation $q_x$ for $q_x(x^n_t|\theta)$. Define the variational point estimator as

$$(\hat{\theta}_n, \hat{q}_x) := \arg\max_{\theta \in \Theta, q} \frac{1}{n} \text{ELBO}(\theta \times q_x||\pi).$$

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

**Assumption 3.1.** (i) The parameter space $\Theta$ 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 $\hat{\theta}_n$ (for $\theta_0$) is guaranteed if the variational family for the states is ‘good enough’.

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

The above result demonstrates that for the variational point estimator $\hat{\theta}_n$ 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, $\theta_0$, 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 $\theta_0$.

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

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

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 $\theta = (\rho, \alpha)'$ are unknown while $\sigma_0$ is known.
We make use of the autoregressive nature of the state process to approximate the posterior for \( \pi(x^n_1|\theta, y^n_1) \) via the variational family:

\[
Q_x := \left\{ q_x : q_x(x^n_1|\rho, \sigma_0) = \mathcal{N}(x_1; 0, \sigma_0^2/(1 - \rho^2)) \prod_{k=2}^{n} q(x_k|x_{k-1}, \rho, \sigma_0) \right\}, \quad q(x_k|x_{k-1}, \rho, \sigma_0) := \mathcal{N}(x_k; \rho x_{k-1}, \sigma_0^2).
\]

When evaluated at \( \theta_0 = (\rho_0, \sigma_0)' \), \( Q_x \) 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 \( 0 \leq |\rho_0| < 1 \), \( 0 \leq |\alpha_0| < M \). (i) Consider that \( \rho_0 = 0 \) and known, then the variational estimator \( \hat{\alpha} \) is consistent if and only if \( \alpha_0 = 0 \). (ii) Consider that \( \alpha_0 = 0 \) and known, then the variational estimator for \( \hat{\rho} \) is consistent if and only if \( \rho_0 = 0 \).

### 3.3 Lack of Bayes consistency of the variational posterior

While the above results pertain to variational point estimators of \( \theta_0 \), 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,

\[
q_x(x^n_1|\theta) := q_\lambda(x^n_1|\theta),
\]

where \( \lambda \in \Lambda \) denotes the vector of so-called ‘variational parameters’ that characterize the elements in \( Q \). With reference to (7), making the dependence of \( q_\lambda(x^n_1|\theta) \) on the variational parameter \( \lambda \) explicit leads to the criterion \( L_n(\theta, \lambda) \), where \( q_x(x^n_1|\theta) \) in (7) is replaced by \( q_\lambda(x^n_1|\theta) \). Optimizing over \( \lambda \) for fixed \( \theta \) yields the profiled criterion,

\[
\hat{L}_n(\theta) := L_n[\theta, \hat{\lambda}_n(\theta)] \equiv \sup_{\lambda \in \Lambda} L_n(\theta, \lambda),
\]

and the ‘idealized’ variational posterior for \( \theta \),

\[
\hat{q}(\theta|y^n_1) \propto \exp \left\{ \hat{L}_n(\theta) \right\} p(\theta).
\]

Note 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 \( \theta_0 \) is that \( \theta_0 \) is the maximum of a well-defined limit counterpart to \( \hat{L}_n(\theta) \).

**Assumption 3.2.** (i) There exists a map \( \theta \mapsto \lambda(\theta) \in \Lambda \) such that \( \sup_{\theta \in \Theta} \| \hat{\lambda}_n(\theta) - \lambda(\theta) \| = o_p(1) \). (ii) There exist a deterministic function \( L : \Theta \times \Lambda \rightarrow \mathbb{R} \) and a \( \theta_* \in \Theta \) such that the following are satisfied: (a) for all \( \epsilon > 0 \) there exists some \( \delta > 0 \) such that \( \inf_{\theta \in \Theta, \lambda \in \Lambda} [L(\theta, \lambda) - L(\theta_*, \lambda(\theta_*))] \leq -\delta \); (b) \( \sup_{\theta \in \Theta, \lambda \in \Lambda} |L_n(\theta, \lambda)/n - L(\theta, \lambda)| = o_p(1) \). (iii) For any \( \epsilon > 0 \), \( \int_{\Theta} 1 \{ \theta : L(\theta, \lambda) - L(\theta_*, \lambda(\theta_*)) < \epsilon \} p(\theta) d\theta > 0 \). (iv) For all \( n \) large, \( \int_{\Theta} \exp \left\{ \hat{L}_n(\theta) \right\} p(\theta) d\theta < \infty \).

**Lemma 3.3.** Under Assumption 3.2, for any \( \epsilon > 0 \), \( \hat{Q} (\{ \theta \in \Theta : d(\theta, \theta_*) > \epsilon \}|y^n_1) = o_p(1) \).
Assumption 3.2(2.b) implies that $L_n[\theta, \lambda]/n$ converges to $L[\theta, \lambda]$ (uniformly in $\theta$ and $\lambda$); while part (2.a) is an identification condition and states that $\mathcal{L}[\theta, \lambda]$ is maximized at some $\theta_*$, which may differ from $\theta_0$. This identification condition makes clear that if $\theta_* \neq \theta_0$, then $\mathcal{L}[\theta_*, \lambda(\theta_*)] > \mathcal{L}[\theta_0, \lambda(\theta_0)]$ and the idealized posterior for $\theta$ will not concentrate onto $\theta_0$. 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 $\theta_0$.

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

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

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 $\theta_0$ 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.

## 4 Existing Variational Approaches: Implications for Inference

As the previous discussion illustrates, the need to approximate the posterior of $x^n_1$ introduces a discrepancy between the exact posterior, obtained via standard methods, and that which results from VB methods. In this way, we can view the latent states $x^n_1$ 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. As the theoretical results demonstrate, the introduction of, and requirement to perform inference on, the incidental parameters means that the (ideal) VB posterior may not concentrate onto the true parameter value that has generated the data, $\theta_0$. It is then clear that VB methods must somehow solve the *incidental parameter problem* if they wish to obtain reliable inference for $\theta_0$.

That being said, the incidental parameter problem has not stopped researchers from using VB methods to conduct inference in SSMs. The general conclusions elucidated above apply, in principle, to all such methods. However, it is useful to explore specific categories of VB methods in greater detail, and comment on their ability to deliver consistent inference for $\theta_0$. We begin with reference to a variational approach to point estimation of $\theta_0$, followed by brief outlines of two classes of approach that target the posterior for $\theta$ via variational methods. In Section 5, we continue the exposition of VB methods (only), but explicitly within the context of prediction. The implications of Bayesian consistency (or lack thereof) for predictive accuracy are discussed therein, and with numerical examples used to document the differences between various approaches.

### 4.1 Profiling out the states

The point estimation approach proposed by Westling and McCormick (2019) splits the optimization procedure into two parts: one for $x^n_1$ conditional on $\theta$; and one for $\theta$ conditional on $x^n_1$. Such an approach allows
one to view variational estimators as (profiled) M-estimators. In particular, this profiled strategy posits a class of variational posteriors for $x_1^n$, $q_\theta(x_1^n)$, which depend on parameters $\lambda \in \Lambda$, and proposes to estimate $\theta$ by maximizing the profiled criterion $L_n[\theta, \hat{\lambda}_n(\theta)]$ in (10). As detailed in Section 3, the optimizer of the limit criterion $\mathcal{L}[\theta, \lambda]$ will not in general coincide with the optimizer of $H(\theta)$, i.e. $\theta_0$. Hence, consistency of such a variational estimator for $\theta_0$ is not guaranteed. This is indeed highlighted by example in Westling and McCormick (2019), where the authors show that inconsistency can occur, even in the case of independent observations, if delicate care is not taken with the choice of variational class for $x_1^n$.

### 4.2 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 $(\theta, x_1^n)$. Such an approach can be motivated by the fact that if we take $q_\theta(x_1^n|\theta) = \pi(x_1^n|y_1^n, \theta)$, (i.e. take the variational approximation for $x_1^n$ to be equivalent to the exact posterior for $x_1^n$ conditional on $\theta$), then we can rewrite $\text{KL}(q||\pi)$ as

$$\text{KL}(q||\pi) = \int \int_{\Theta} q_\theta(\theta) \pi(x_1^n|y_1^n, \theta) \log \frac{q_\theta(\theta) \pi(x_1^n|y_1^n, \theta)}{\pi(x_1^n|y_1^n, \theta) \pi(\theta|y_1^n)} d\theta dx_1^n,$$

with the final line exploiting the fact that $\pi(x_1^n|y_1^n, \theta)$ integrates to one for all $\theta$. 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 $(\theta, x_1^n)$ into a variational problem for $\theta$ 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 $q_\theta$ 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 $\hat{p}_\theta(y_1^n)$. We follow Tran et al. (2017) and write $\hat{p}_\theta(y_1^n)$ as $\hat{p}(y_1^n|\theta, z)$ 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 $z = \log \hat{p}(y_1^n|\theta, z) - \log p_\theta(y_1^n)$. For $g(z|\theta)$ denoting the density of $z|\theta$, Tran et al. (2017) consider VB for the
augmented posterior

\[ p(\theta, z|y_i^n) = \tilde{p}(y_i^n|\theta, z)g(z|\theta)p(\theta)/p(y_i^n) = p_\theta(y_i^n) \exp(z)g(z|\theta)p(\theta)/p(y_i^n) = \pi(\theta|y_i^n) \exp(z)g(z|\theta), \]

which, marginal of \( z \), has the correct target posterior \( \pi(\theta|y_i^n) \) due to the unbiasedness of the estimator \( \tilde{p}(y_i^n|\theta, z) \). 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 \( q(\theta, z) = q_\theta(\theta)g(z|\theta) \) and \( \pi(\theta, z|y_i^n) \) that minimizes the KL divergence between \( q(\theta, z) = q_\theta(\theta)g(z|\theta) \) and \( \pi(\theta, z|y_i^n) \):

\[
\text{KL}[q(\theta, z)||\pi(\theta, z|y_i^n)] = \int_\Theta \int_z q_\theta(\theta)g(z|\theta) \log \frac{q_\theta(\theta)g(z|\theta)}{\pi(\theta|y_i^n) \exp(z)g(z|\theta)} dz d\theta
= \int_\Theta \int_z q_\theta(\theta)g(z|\theta) \log \frac{q_\theta(\theta)}{p_\theta(y_i^n) \exp(z) p(\theta)} dz d\theta + \log p(y_i^n)
= -\int_\Theta q_\theta(\theta) \log p_\theta(y_i^n) d\theta + \text{KL}(q_\theta||p_\theta) + \Upsilon_n[q(\theta, z)] + \log p(y_i^n),
\]

where in this case

\[
\Upsilon_n[q(\theta, z)] = \int_\Theta \int_z q_\theta(\theta)g(z|\theta) \{ \log p_\theta(y_i^n) - \log \tilde{p}(y_i^n|\theta, z) \} dz d\theta.
\]

For fixed \( \theta \), \( \mathbb{E}_z[\tilde{p}(y_i^n|\theta, z)] = p_\theta(y_i^n) \), but in general \( \log \tilde{p}(y_i^n|\theta, z) \) is a biased estimator of \( \log p_\theta(y_i^n) \), 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 \( q_\theta(x_i^n|\theta) \), VBIL can achieve consistent inference on \( \theta_0 \) by choosing an appropriate number of particles \( N \) in the production of \( \tilde{p}(y_i^n|\theta, z) \).

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 \( \text{Var}[z|\theta] = \gamma(\theta)^2/N \), where \( \gamma(\theta)^2 \) is bounded uniformly over \( \Theta \); 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 \( \text{Var}[z|\theta] = \sigma^2 > 0 \), \( 0 < \sigma < \infty \). Note that, under this choice for \( N \), for any \( \varepsilon > 0 \)

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

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 \( \theta \) alone. Consequently, existing results on the posterior concentration of VB methods for \( \theta \) alone can be used to deduce posterior concentration of the VBIL posterior for \( \theta \).

### 4.3 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 approximation 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 (2018) 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 \( \pi(x^n_1|y^n_1, \theta) \) by approximating the relationship between \( X_t \) and \( X_{t-1} \), which may in truth be non-linear in \( \theta \), by the random walk model \( X_t = X_{t-1} + \epsilon_t \), with \( \epsilon_t \sim i.i.d. N(0, \sigma^2_0) \), 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 \( q(x^n_1|\theta) = q_0(\theta)q_2(x^n_1) \), where \( q_2(x^n_1) \propto \prod_{k \geq 1} \exp(-\{x_k - \hat{x}_{k|k}\}^2 (1 - K_k)P_{k|k-1}/2) \) and where the terms \( K_k, P_{k|k-1}, \hat{x}_{k|k} \) are explicitly calculated using the Kalman recursion: \( \hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k (y^t_k - \hat{x}_{k|k-1}) \), and where \( K_k \) is the Kalman gain, \( P_{k|k-1} \) is the predicted variance of the state, and in the application of Koop and Korobilis (2018), \( y^t_k = \log(y^2_t) \).

While the solution proposed by the VBKF is likely to lead to better inference on the states, especially when \( x^n_1 \) behaves like a random walk, ultimately we are still ‘conducting inference’ on \( x^n_1 \), and thus we still encounter the incidental parameter problem as a consequence. Indeed, taking as the variational family for \( x^n_1 \) the Kalman filter approximation yields, at time \( k \geq 1 \), a conditionally normal density with mean \( \hat{x}_{k|k} \) and variance \( (1 - K_k)P_{k|k-1} \). 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 \( \theta \) 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 \( q_2(x^n_1) \). As with the VBKF approach, the assumed nature of the state process used by Chan and Yu (2020) to estimate \( q_2(x^n_1) \) implies that, in general, it is unlikely that Bayesian consistency can be achieved. Since this method plays a role in the numerical prediction exercises below, we forgo further discussion of the specific details of the approach until Section 6.

5 VB-based Prediction: Implications for Predictive Accuracy

5.1 Overview

As highlighted by the above analysis and, indeed, as also acknowledged by other authors (e.g. Koop and Korobilis, 2018; Gunawan et al., 2020), 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. However, as highlighted by Quiroz et al. (2018) and Frazier et al. (2021), amongst others, VB approximations can perform admirably in predictive settings, 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 \( \theta_0 \) that has generated the data (as shown in Lemma 3.3), so long as the value onto
which it is concentrating is not too far away, it may be that predictions produced by VB approaches will also perform well in practice.

In this section, we shed further light on the phenomena of the predictive accuracy of VB methods, and connect the performance of these methods to the inconsistency for $\theta_0$ that can result as the sample size diverges. The results suggest that, while there is little difference between variational methods in predictive settings with either a small sample size or a small number of out-of-sample observations, there is a clear hierarchy in terms of predictive accuracy across methods as the sample size becomes larger and as the out-of-sample evaluation increases.

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

$$p(Y_{n+1}|y_1^n) = \int_{\Theta} \int_X g_\theta(Y_{n+1}|x_{n+1}) \pi(x_{n+1}^n, \theta|y_1^n) dx_{n+1} d\theta$$

$$= \int_{\Theta} \int_X \int_X \int_{X'} g_\theta(Y_{n+1}|x_{n+1}) p(x_{n+1}|x_n, y_1^n, \theta) p(x_1^n|y_1^n, \theta) \pi(\theta|y_1^n) dx_{n+1} dx_1^n d\theta,$$

where the last line follows from the Markovianity of the state transition equation (see equation (1)). In many large SSMs, using MCMC methods to estimate (11) is infeasible or prohibitive computationally, due to the difficulty of sampling from $\pi(x_{n+1}^n, \theta|y_1^n)$. 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 (11) underlined as (1) and (2). All such methods replace the second underlined term by some approximate posterior for $\theta$, 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, $\tilde{q}_x$ to replace $p(x_1^n|y_1^n, \theta)$; and a class that uses an accurate simulation-based estimate of $p(x_1^n|y_1^n, \theta)$. We discuss these two strategies in greater detail in the following section.

5.2 Methods of producing the variational predictive

5.2.1 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 $\tilde{q}_x$ 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 (11) 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 $q_\theta$. Given the realization $\theta^{(j)}$, we simulate $x_1^n|y_1^n, \theta^{(j)}$ from the VB approximation of the states $\tilde{q}_x$. Next, given $x_n^{(j)} \sim \tilde{q}_x$, 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 $x_{n+1}$ is not a draw from $p(x_{n+1} | x_{n}, y_{n}^{n}, \theta) p(x_{n}^{n} | y_{1}^{n}, \theta)$. Hence, the draw $Y_{n+1}^{(j)}$ does not correctly reflect the structure of the assumed model, and $Y_{n+1}$ cannot be viewed as being a draw from the exact predictive density in (11).

Notable uses of the above approach to prediction appear in Quiroz et al. (2018), Koop and Korobilis (2018) 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 $\hat{q}_{x}$ (in addition to the differences in the construction of $\hat{q}_{\theta}$) and thus to generate $x_{n+1}^{(j)}$.

5.2.2 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 $\theta$ from some $\hat{q}_{\theta}$. 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 $\hat{q}_{x}$; see, for example, Loaiza-Maya et al. (2021). In this case, $x_{n+1}^{(j)}$ is a draw from $p(x_{n+1} | x_{n}, y_{n}^{n}, \theta) p(x_{n}^{n} | y_{n}^{n}, \theta)$ and, consequently, the draw $Y_{n+1}^{(j)}$ correctly reflects the model structure. Moreover, due to the Markovian nature of (1), 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.

6 Numerical Assessment of VB methods

6.1 Simulation design

We now undertake a simulation exercise to compare the inferential and predictive accuracy of several competing variational methods. Beginning with the inferential assessment, we compare the 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. The assumed DGP is specified as an unobserved component model with stochastic volatility (UCSV):

$$\begin{align*}
\mu_{t} &= \mu + \rho_{\mu} (\mu_{t-1} - \bar{\mu}) + \sigma_{\mu} \varepsilon_{t}, \quad h_{t} = \bar{h} + \rho_{h} (h_{t-1} - \bar{h}) + \sigma_{h} \eta_{t}, \quad Y_{t} = \mu_{t} + \exp(h_{t}/2) u_{t},
\end{align*}$$

(12)

where $(\varepsilon_{t}, \eta_{t}, u_{t}) \overset{i.i.d.}{\sim} N(0, I_{3})$. The unobserved component term $\mu_{t}$ 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:

DGP 1: $\bar{\mu}_{0} = 0; \rho_{\mu_{0}} = 0.8; \sigma_{\mu_{0}} = 0.5; \bar{h}_{0} = -1.0; \rho_{h_{0}} = 0.00; \sigma_{h_{0}} = 0.0$

DGP 2: $\bar{\mu}_{0} = 0; \rho_{\mu_{0}} = 0.0; \sigma_{\mu_{0}} = 0.0; \bar{h}_{0} = -1.3; \rho_{h_{0}} = 0.95; \sigma_{h_{0}} = 0.3$

DGP 3: $\bar{\mu}_{0} = 0; \rho_{\mu_{0}} = 0.8; \sigma_{\mu_{0}} = 0.5; \bar{h}_{0} = -1.3; \rho_{h_{0}} = 0.95; \sigma_{h_{0}} = 0.3$
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 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 \( \theta_0 = (\bar{\mu}_0, \rho_{\mu_0}, \sigma_{\mu_0}, \bar{h}_0, \rho_{h_0}, \sigma_{h_0})' \).

For the predictive assessment we compare exact Bayes with the two variational methods cited above plus the method of Chan and Yu (2020). As noted earlier, this latter method exploits a very specific structure in the construction of the variational algorithm, in this case a model that corresponds to DGP 2, but with \( \rho_{h_0} = 1.0 \) and \( \bar{h}_0 = 0 \), i.e. to a model with a fixed mean and a random walk process for the variance. Thus, application of this approach under any of the above true DGPs constitutes misspecified inference; hence why we do not include this technique in the inferential assessment. However, and consistent with the opening comment made in Section 5.1, any inferential inaccuracy that this misspecification might induce may not have severe consequences for predictive accuracy, and we do include this method in the predictive assessment under DGP 2.

### 6.2 Competing methods

#### Exact Bayes

Denote the two vectors of latent variables as \( \mu^n_1 = (\mu_1, \ldots, \mu_n)' \) and \( h^n_1 = (h_1, \ldots, h_n)' \). The exact posterior density is given as

\[
\pi(\theta, \mu^n_1, h^n_1|y^n_1) = \frac{p(y^n_1|\mu^n_1, h^n_1, \theta)p(\mu^n_1, h^n_1|\theta)p(\theta)}{p(y^n_1)},
\]

with prior \( p(\theta) = p(\bar{\mu})p(\rho_\mu)p(\bar{h})p(\rho_h)p(\sigma_h) \), where \( \bar{\mu} \sim N(0, 1000), \rho_\mu \sim U(0, 1), \sigma_\mu^2 \sim IG(1.001, 1.001), \bar{h} \sim N(0, 1000), \rho_h \sim U(0, 1) \) and \( \sigma_h^2 \sim IG(1.001, 1.001) \). We draw from (13) using an MCMC algorithm. Specifically, the vector \( h^n_1 \) is generated using the method proposed in Primiceri (2005), while \( \mu^n_1 \) is generated using the forward-filtering backward-sampling method in Carter and Kohn (1994). Given the choice of priors, the parameters \( \bar{\mu}, \sigma_\mu, \bar{h} \) and \( \sigma_h \) can be generated directly using Gibbs steps. The parameters \( \rho_\mu \) and \( \rho_h \) are generated using a Metropolis-Hastings step with a Gaussian proposal distribution. The corresponding predictive (expressed using obvious notation),

\[
p(Y_{n+1}|y^n_1) = \int \int \int g_\theta(Y_{n+1}|\mu_{n+1}, h_{n+1})p(\mu_{n+1}, h_{n+1}|\theta, \mu^n_1, h^n_1)\pi(\theta, \mu^n_1, h^n_1|y^n_1)d\mu_{n+1}dh_{n+1}d\theta,
\]

is then estimated (via kernel density methods) using the draws of \( Y_{n+1} \) obtained conditional on the draws of \( \theta, \mu^n_1 \) and \( h^n_1 \).

**Quiroz et al. (2018)**

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

\[
qu_\lambda(\theta, \mu^n_1, h^n_1) = qu_\lambda(\theta)qu_\lambda(x^n_1),
\]

(15)
where $\hat{\lambda} = (\hat{\lambda}_1', \hat{\lambda}_2')'$, $x_t = (\mu_t, h_t)'$ and $x_1^n = (x_1', \ldots, x_n')'$. The approximations $q_{\lambda_1}(\theta)$ and $q_{\lambda_2}(x_1^n)$ are optimal elements in the variational classes $Q_1 = \{q_{\lambda_1}(\theta) : \lambda_1 \in \Lambda_1\}$ and $Q_2 = \{q_{\lambda_2}(\theta) : \lambda_2 \in \Lambda_2\}$, 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 $q_{\lambda_1}(\theta) = \phi_0(\theta; \nu_0, BB' + \text{diag}(d^2))$, while the elements of the second class are of the form $q_{\lambda_2}(x_1^n) = \phi_{2n}(x_1^n; \nu_x, CC')$, where $C$ is a three diagonal lower triangular matrix, and the subscript on the symbol for the normal pdf, $\phi$, denotes the dimension of the density. (For more details on this approximating class see Ong et al., 2018.) Replacing $\pi(\theta, \mu^n_1, h^n_1|y^n_1)$ in (14) by the approximation in (15), the predictive density is then estimated as described in Section 5.2.1.

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 $\pi(\theta|y^n_1)$ only, exploiting the exact conditional posterior density of the states, $p(\mu_1^n, h_1^n|y_1^n, \theta)$. As such, the variational approximation takes the form:

$$q_{\lambda}(\theta, \mu_1^n, h_1^n|y_1^n) = q_{\lambda}(\theta)p(\mu_1^n, h_1^n|y_1^n, \theta),$$  

(16)

where $q_{\lambda}(\theta)$ is an optimal element in the variational class $Q = \{q_{\lambda}(\theta) : \lambda \in \Lambda\}$, once again found via SGA. For $Q$ the class of multivariate Gaussian densities with a factor structure is employed, so that

$q_{\lambda}(\theta) = \phi_0(\theta; \nu, BB' + \text{diag}(d^2))$, and $\lambda = (\nu', \text{vec}(B)', d')'$. Replacing $\pi(\theta, \mu_1^n, h_1^n|y_1^n)$ in (14) by the approximation in (16) (once again, with the same underlying prior adopted), the predictive density is then estimated as described in Section 5.2.2. Generation from $p(\mu_1^n, h_1^n|y_1^n, \theta)$ is achieved via an MCMC algorithm that sequentially draws from $p(\mu_1^n|y_1^n, \theta, h_1^n)$ using the method in Carter and Kohn (1994); and then draws from $p(h_1^n|y_1^n, \theta, \mu_1^n)$ using the approach in Primiceri (2005).

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:

$$q_{\lambda}(\theta, h_1^n) = q_{\lambda_1}(\sigma_h^2)q_{\lambda_2}(h_0)q_{\lambda_3}(h_1^n),$$

15
where \( q_{\lambda_1}(\sigma_n^2) \), \( q_{\lambda_2}(h_0) \) and \( q_{\lambda_3}(h^n_1) \) are optimal elements in the variational classes \( Q_1 = \{ q_{\lambda_1}(\sigma_n^2) : \lambda_1 \in \Lambda_1 \} \), \( Q_2 = \{ q_{\lambda_2}(h_0) : \lambda_2 \in \Lambda_2 \} \) and \( Q_3 = \{ q_{\lambda_3}(h^n_1) : \lambda_3 \in \Lambda_3 \} \), respectively. The elements of each class are defined respectively as \( q_{\lambda_1}(\sigma_n^2) = IG(\sigma_n^2; \nu, S) \), \( q_{\lambda_2}(h_0) = \phi_0(h_0; \mu_0, s_0^2) \) and \( q_{\lambda_3}(h^n_1) = \phi_n(h^n_1; m, \hat{K}^{-1}) \). 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 \( Q_1, Q_2 \) and \( Q_3 \) that minimise the KL divergence from \( p(\theta, h^n_1|y^n_1) \). 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 \( \lambda_1, \lambda_2, \lambda_3 \) and \( y^n_1 \). In our implementation of the CY method, the priors are set to \( p(\sigma_n^2) = IG(\sigma_n^2; 1.001, 1.001) \) and \( p(h_0) = \phi_0(h_0; 0, 1000) \), where \( IG \) denotes the inverse gamma distribution. The predictive density is then estimated as described in Section 5.2.1, with \( h^n_1 \) playing the role of \( x^n_1 \) therein.

### 6.3 Inferential accuracy

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 \( t = 1, 2, \ldots, T \), for the unobserved component, \( \mu_t \), and the stochastic standard deviation, \( \exp(1/2h_t) \), relative to the marginal posterior means that results when we condition on the true parameters, denoted respectively by \( E[\mu_t|\theta_0, y^n_1] \) and \( E[\exp(1/2h_t)|\theta_0, y^n_1] \), \( t = 1, 2, \ldots, T \). Table 1 presents the results.

As expected, exact Bayes produces the most accurate point estimates for the two sets of latent variables (both \( \mu_t \) and \( \exp(h_t/2) \)), as tallies with the theoretical consistency of this method, which ensures that the posterior concentrates onto \( \theta_0 \). That is, Bayesian consistency (for \( \theta_0 \)) of the exact posterior (given the required regularity) implies that the average of the RMSE and MAE between the exact posterior mean of \( \mu_t \) and \( E[\mu_t|\theta_0, y^n_1] \) will converge to zero as the sample size diverges; and similarly for the posterior mean of \( \exp(1/2h_t) \).

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, as highlighted in Section 4.2. Instead, a variational approximation of \( \pi(\theta|y^n_1) \) only is invoked and, with sufficient regularity, Bayesian consistent results are anticipated. In contrast, the QNK method does not deal directly with this problem and, as a consequence, exhibits - in 10 of the 12 designs recorded in Table 1 - inaccuracy that is between two and eight times greater than that of both exact Bayes and the LSND method. As we will see in the following section, however, this inferential inaccuracy does not necessarily translate into the same degree of predictive inaccuracy.

### 6.4 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 \( y^n_1 \), for each of the competing approaches and for
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 ($\mu_t$). 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 $E[\mu_t|\theta_0, y_{1:n}]$, while the conditional standard deviation error measures are computed relative to $E[\exp(h_t/2)|\theta_0, y_{1:n}]$, where $\theta_0$ denotes the true parameter vector.

Panel A: Unobserved component ($\mu_t$)

|               | RMSE | MAE  |
|---------------|------|------|
|               | DGP 1 | DGP 2 | DGP 3 | DGP 1 | DGP 2 | DGP 3 |
| Exact Bayes   | 0.0463 | 0.0865 | 0.0203 | 0.0382 | 0.0075 | 0.0155 |
| LSND          | 0.0495 | 0.0817 | 0.0271 | 0.0405 | 0.0067 | 0.0207 |
| QNK           | 0.1211 | 0.0353 | 0.1098 | 0.0980 | 0.0012 | 0.0664 |

Panel B: Conditional standard deviation ($\exp(h_t/2)$)

|               | RMSE | MAE  |
|---------------|------|------|
|               | DGP 1 | DGP 2 | DGP 3 | DGP 1 | DGP 2 | DGP 3 |
| Exact Bayes   | 0.0497 | 0.0540 | 0.0234 | 0.0470 | 0.0492 | 0.0180 |
| LSND          | 0.0520 | 0.0519 | 0.0315 | 0.0490 | 0.0468 | 0.0246 |
| QNK           | 0.0984 | 0.1336 | 0.2231 | 0.0983 | 0.0925 | 0.1656 |

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 A.3 of the Supplementary Appendix. The results using 10000 out-of-sample evaluations are presented in Table 2, remembering that the CY method is now included in the comparison, but only for the case of DGP 2. Whilst not reported in the main text due to space constraints, results for 100 and 1000 out-of-sample evaluations are given in Section A.4 of the Supplementary Appendix.

From Table 2 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. From the results recorded in Section A.4 of the Supplementary Appendix a similar ranking holds for the smaller out-of-sample evaluation periods. However, the differences 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 only for the larger sample size) that is driving the discrepancies between the predictive accuracy of the competing methods.

Whilst a ranking is in clear evidence in Table 2, it can be argued that across certain DGP and scoring rule combinations, the predictive results across the different methods are somewhat 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 fairly well, 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 (rel-
Table 2: Predictive performance of the competing Bayesian approaches: exact Bayes, LSND, QNK and CY. The column labels indicate the out-of-sample predictive performance measure while the row labels indicate the predictive method. Panels A to C correspond to the results for DGP 1 to 3, respectively. The average predictive measures in this table were computed using 10000 out-of-sample evaluations.

|                | LS   | CS-10% | CS-20% | CS-80% | CS-90% | CRPS   | TWCRPS | IS    |
|----------------|------|--------|--------|--------|--------|--------|--------|-------|
| **Panel A: DGP 1** |      |        |        |        |        |        |        |       |
| 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             | -1.262 | -0.309 | -0.509 | -0.507 | -0.298 | -0.482 | -0.147 | -4.030 |
| QNK            | -1.262 | -0.309 | -0.509 | -0.507 | -0.298 | -0.482 | -0.147 | -4.030 |
| **Panel B: DGP 2** |      |        |        |        |        |        |        |       |
| True DGP       | -0.862 | -0.309 | -0.499 | -0.497 | -0.305 | -0.343 | -0.105 | -3.268 |
| Exact Bayes    | -0.865 | -0.309 | -0.499 | -0.498 | -0.306 | -0.343 | -0.105 | -3.273 |
| LSND           | -0.866 | -0.310 | -0.500 | -0.498 | -0.306 | -0.343 | -0.105 | -3.280 |
| CY             | -0.873 | -0.312 | -0.503 | -0.500 | -0.308 | -0.344 | -0.105 | -3.344 |
| QNK            | -0.899 | -0.322 | -0.514 | -0.511 | -0.317 | -0.346 | -0.106 | -3.489 |
| **Panel C: DGP 3** |      |        |        |        |        |        |        |       |
| 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           | -1.261 | -0.306 | -0.507 | -0.521 | -0.301 | -0.491 | -0.150 | -4.442 |
| CY             | -1.301 | -0.315 | -0.521 | -0.536 | -0.311 | -0.497 | -0.152 | -4.707 |

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. For example, there is a clear trend that as model complexity increases, 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 these cases, the all-purpose variational method of Quiroz et al. (2018) performs the worst across all the DGPs under analysis. 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.

7 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 methods that are capable of closing Jensen’s gap can be nearly as accurate as exact inference. However, if the VB methods cannot close Jensen’s gap, the resulting inference is inferior to that obtained by exact inference methods.

Contrary to what has been reported in the literature so far, we find a clear hierarchy in terms of predictive accuracy across different variational methods: methods that can close Jensen’s gap produce qualitatively better predictions than those that do not. However, we find that the extent of the discrepancy between different types of variational approaches depends on the data generating process (DGP), the loss in which the different methods are evaluated, and on 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 certain variational methods are substantially less accurate than exact methods, and other circumstances where there is little difference between the various approaches.

The broad findings of this paper are two fold. Firstly, if inferential accuracy is the overall goal, then the closer the variational method is to correctly approximating the states, the higher the gains. However, if prediction is the main goal, then there is greater flexibility in terms of the resulting approximations one chooses. In general, while the inferential accuracy hierarchy discussed above also holds for prediction, the differences become less stark: for instance, the structured approximation of Chan and Yu (2020) performs only slightly worse than the more accurate inferential method of Loaiza-Maya et al. (2021). Whilst the so-called ‘all-purpose’ approximation of Quiroz et al. (2018), generally performs the worst, in terms of both inference and prediction, it is worth pointing out that it still produces reasonably accurate predictions, and can be applied in a plethora of situations where other approaches cannot be implemented either due to computational cost or due to particular features of the model under analysis.

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.

Ait-Sahalia, Y., Li, C., and Li, C. (2020). Maximum likelihood estimation of stochastic volatility models. Forthcoming, Journal of Econometrics.

Andersen, T. G. and Sørensen, 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. jrssb, 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.

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., Singh, S., Jasra, A., and Peters, G. (2014). Parameter inference for hidden Markov models with intractable likelihoods. Scand. J. Statist. (to appear).

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. (2020). Variational approximation of factor stochastic volatility models. arXiv preprint arXiv:2010.06738.

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. (2018). Variational Bayes inference in high-dimensional time-varying parameter models. Forthcoming, Journal of Econometrics.

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. (2019). Asymptotic normality, concentration, and coverage of generalized posteriors. arXiv preprint arXiv:1907.09611.

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., Golightly, 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.
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. 2, 9, 10

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. 8, 9

Yang, Y., Pati, D., Bhattacharya, A., et al. (2020). alpha-variational inference with statistical guarantees. *Annals of Statistics*, 48(2):886–905. 5

**A Appendix**

This appendix contains proofs of all theoretical results, certain definitions, and additional numerical results referenced in the main text.

**A.1 Proofs of Main Results**

*Proof of Lemma 3.1.* The proof follows a modification of the standard argument; 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

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

where $H(\theta) = \plim_n \ell_n(\theta)$, $\ell_n(\theta) = \frac{1}{n} \log p_0(y_1^n)$ and $\theta_0$ satisfies $H(\theta_0) \geq \sup_{\theta \in \Theta} H(\theta)$. The stated result then follows if the RHS is $o(1)$.

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

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

Therefore,

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

Conclude that $H(\theta_0) - H(\hat{\theta}_n) \leq o_p(1)$ 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 (2019), and Syring and Martin (2020).
Define \( \Pi_n(\Theta) := \int_\Theta \exp \left\{ \widehat{L}_n(\theta) \right\} p(\theta) d\theta \) 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 \( \delta > 0 \),

\[
\tilde{Q}(A_\epsilon | y^n_n) = \frac{\Pi_n(A_\epsilon)}{\Pi_n(\Theta)} = \frac{\frac{\Pi_n(A_\epsilon)}{\Pi_n(\Theta)} \exp \left\{ -\widehat{L}_n(\theta_\star) + n\delta \right\}}{\frac{\Pi_n(A_\epsilon)}{\Pi_n(\Theta)} \exp \left\{ -\widehat{L}_n(\theta_\star) + n\delta \right\}}
\]

\[
= \frac{\int_{A_\epsilon} \exp \left\{ -\widehat{L}_n(\theta_\star) + n\delta \right\} \exp \left\{ \widehat{L}_n(\theta) \right\} p(\theta) d\theta}{\int_{\Theta} \exp \left\{ -\widehat{L}_n(\theta_\star) + n\delta \right\} \exp \left\{ \widehat{L}_n(\theta) \right\} p(\theta) d\theta}
\]

\[
= \frac{N_n}{D_n}.
\]

We treat the numerator and denominator separately.

Write the numerator as

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

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

\[
\frac{\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_\star)/n}{n} \leq 2 \sup_{\theta \in \Theta, \lambda \in \Lambda} |\mathcal{L}_n(\theta, \lambda)/n - \mathcal{L}(\theta, \lambda)| + \mathcal{L}[\theta, \widehat{\lambda}_n(\theta)] - \mathcal{L}[\theta_\star, \widehat{\lambda}_n(\theta_\star)]
\]

\[
\leq o_p(1) + \left\{ \mathcal{L}[\theta, \widehat{\lambda}_n(\theta)] - \mathcal{L}[\theta, \lambda(\theta)] \right\} - \left\{ \mathcal{L}[\theta_\star, \widehat{\lambda}_n(\theta_\star)] - \mathcal{L}[\theta_\star, \lambda(\theta_\star)] \right\}
\]

\[
+ \mathcal{L}[\theta, \lambda(\theta)] - \mathcal{L}[\theta_\star, \lambda(\theta_\star)]
\]

\[
\leq o_p(1) + \mathcal{L}[\theta, \lambda(\theta)] - \mathcal{L}[\theta_\star, \lambda(\theta_\star)]
\]

\[
\leq o_p(1) - \delta
\]

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 \( \theta \), 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

\[
\exp \left\{ n \left[ \frac{\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_\star)/n + \delta}{n} \right] \right\} \leq 1.
\]
Consequently, wpc1,

\[ N_n = \int_{A_e} \exp \left\{ n \left[ \hat{L}_n(\theta)/n - \hat{L}_n(\theta_*)/n + \delta \right] \right\} p(\theta) d\theta \leq \Pi(A_e) \leq 1. \]

To handle the denominator, first define \( L(\theta) = \mathcal{L}[\theta, \lambda(\theta)] \) and \( G_\delta := \{ \theta : L(\theta) - L(\theta_*) < -\delta/2 \} \) and note that, for any \( \theta \in G_\delta \), by Assumption 3.2(2.b),

\[ \left\{ \hat{L}_n(\theta)/n - \hat{L}_n(\theta_*)/n \right\} + \delta/2 \to L(\theta) - L(\theta_*) + \delta/2 < 0, \]

wpc1. Thus, for any \( \delta > 0 \) and any \( \theta \in G_\delta \), \( \exp \left\{ n \left[ \hat{L}_n(\theta)/n - \hat{L}_n(\theta_*)/n + \delta \right] \right\} \to \infty \) as \( n \to \infty \) wpc1. From Fatous lemma

\[
\liminf_{n \to \infty} \exp \left\{ -\hat{L}_n(\theta_*) + n\delta \right\} \Pi_n(G_\delta) = \liminf_{n \to \infty} \int_{G_\delta} \exp \left\{ n \left[ \hat{L}_n(\theta)/n - \hat{L}_n(\theta_*)/n + \delta \right] \right\} p(\theta) d\theta \\
\geq \liminf_{n \to \infty} \exp [n\delta/4] \Pi(G_\delta).
\]

Since \( \Pi(G_\delta) > 0 \) 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 \( D_n \to \infty \) as \( n \to \infty \) (wpc1).

\[ \square \]

**Lemma 3.2.** The complete data likelihood is proportional to

\[
p(x^n_1, y^n_1|\theta) = \{2\pi\sigma^2_0\}^{-n} \exp \left\{ -\frac{1}{2\sigma^2_0} \sum_{k=1}^{n-1} (x_{k+1} - \rho_0 x_k)^2 - \frac{1}{2\sigma^2_0} \sum_{k=1}^{n} (y_k - \alpha x_k)^2 - \frac{1}{2\sigma^2_0} (x_1)^2 \right\} \\
= \{2\pi\sigma^2_0\}^{-n} \exp \left\{ -\frac{1}{2\sigma^2_0} [(x^n_1)'\Omega_n(\theta)x^n_1 - 2\alpha(y^n_1)'x^n_1 + (y^n_1)'y^n_1] \right\},
\]

for the matrix

\[
\Omega_n(\theta) := \begin{pmatrix}
(1 + \rho^2 + \alpha^2) & -\rho & 0 & \ldots & 0 \\
-\rho & (1 + \rho^2 + \alpha^2) & -\rho & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & (1 + \alpha^2)
\end{pmatrix}.
\]

The states can be analytically integrated out, using known results for multivariate normal integrals, to obtain the observed data likelihood \( p(y^n_1|\theta) \):

\[
p(y^n_1|\theta) = \{2\pi\sigma^2_0\}^{-n} \left[ \frac{(2\pi)^n}{\sigma^2_0 \Omega_n(\theta)} \right]^{1/2} \exp \left\{ -\frac{1}{2\sigma^2_0} [(y^n_1)'y^n_1 - \alpha^2(y^n_1)'\Omega_n(\theta)^{-1}y^n_1] \right\},
\]

\[ 24 \]
which yields the observable data log-likelihood

$$\log p(y^n_1|\theta) = - \frac{n}{2} \log 2\pi - \frac{n}{2} \log(\sigma_0^2) - \frac{1}{2} \log |\Omega(\theta)| + \frac{1}{2\sigma_0^2} \left[ \alpha^2(y^n_1)^T \Omega_n(\theta)^{-1} y^n_1 - (y^n_1)^T y^n_1 \right].$$

Following Lemma 3.1, consider the infeasible situation where our variational family for $\theta$ is

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

Under this choice, consistency follows if $T_n(q)/n = \frac{1}{n} \{ \log p(y^n_1|\theta_0) - \mathcal{L}_n(\theta_0) \} = o_p(1)$. In the remainder, we drop the dependence of $q_x(x^n_1|\theta)$ on $\theta_0$ and simply denote $q_x(x^n_1) = q_x(x^n_1|\theta_0)$.

Under the choice of $Q_\theta$, 

$$\mathcal{L}_n(\theta_0) = \int_{\mathcal{X}} q_x(x^n_1) \log \frac{p(x^n_1, y^n_1|\theta)}{q_x(x^n_1)} \, dx^n_1 = -n \log 2\pi - n \log \sigma_0^2$$ 

$$\quad - \frac{1}{2\sigma_0^2} \sum_{k=2}^n \int (x_k - \rho_0 x_{k-1})^2 q_x(x_k, x_{k-1}) \, dx_k \, dx_{k-1}$$ 

$$\quad - \frac{1}{2\sigma_0^2} \sum_{k=1}^{n-1} \int (y_k - \alpha_0 x_k)^2 q_x(x_{k+1}, x_k) \, dx_{k+1} \, dx_k$$ 

$$\quad - \frac{1}{2\sigma_0^2} \int (x_1)^2 \mathcal{N}(x_1; 0, \sigma_0^2) \, dx_1$$ 

$$\quad - \int q_x(x^n_1) \log \prod_{k=2}^n q(x_k | x_{k-1}) \, dx^n_1,$$

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

$$\int q_x(x_k, x_{k-1}) \log q(x_k | x_{k-1}) \, dx_k \, dx_{k-1} = -\frac{1}{2} \log 2\pi - \frac{1}{2} - \frac{1}{2} \log \sigma_0^2(1 - \rho_0^2)$$

$$\int (x_k - \rho_0 x_{k-1})^2 q_x(x_k, x_{k-1}) \, dx_k \, dx_{k-1} = \sigma_0^2(1 - \rho_0^2)$$

$$\int (x_1)^2 \mathcal{N}(x_1; 0, \sigma_0^2) \, dx_1 = \sigma_0^2$$

$$\int (y_k - \alpha_0 x_k)^2 \, dx_k = y_k^2 + \alpha_0^2 \sigma_0^2,$$

to obtain

$$\mathcal{L}_n(\theta_0) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma_0^2 + \frac{n}{2} \log (1 - \rho_0^2) - \frac{1}{2\sigma_0^2} \left\{ n \sigma_0^2 (1 - \rho_0^2) \right\} - \frac{1}{2\sigma_0^2} (y^n_1)^T y^n_1 - \frac{n}{2} \frac{\alpha_0^2}{2\sigma_0^2} (y^n_1)^T \Omega_n(\theta_0)^{-1} y^n_1.$$

Similarly, we have that

$$\log p(y^n_1|\theta_0) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma_0^2 - \frac{1}{2} \log |\Omega_n(\theta_0)| + \frac{\alpha_0^2}{2\sigma_0^2} (y^n_1)^T \Omega_n(\theta_0)^{-1} y^n_1 - \frac{1}{2\sigma_0^2} (y^n_1)^T y^n_1.$$
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 \( \theta_0 \) and, it can be shown that, \(^1\) 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 \( d := \sqrt{a^2 - 4\rho_0^2} \),

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

and we can define
\[ C_1(\theta_0) := \lim_{n \to \infty} \log |\Omega_n(\theta_0)|/n. \]

Conversely, if \( (1 + \alpha_0^2 + \rho_0^2)^2 - 4\alpha_0^2 = 0 \), 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 \( \sigma_x^2 = \sigma_0^2/(1 - \rho_0^2) \),

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

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

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

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 \( \epsilon > 0 \),

\[ \Pr(|Z_n - \mathbb{E}[Z_n]| > n\epsilon) \leq \text{Var}[Z_n]/(n^2\epsilon^2) = \text{Tr} \left[ \Omega_n(\theta_0)^{-1}M\Omega_n(\theta_0)^{-1}M \right]/(n^2\epsilon^2) \quad (17) \]

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,

\[ \text{Var}[Z_n] = \text{Tr} \left[ \Omega_n(\theta_0)^{-1}M\Omega_n(\theta_0)^{-1}M \right] \leq n \cdot \sup_{n \geq 1} \{|D_1|, \ldots, |D_n|\}. \quad (18) \]

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

\(^1\)See Lemma A.1 for details.
of $c_n$, apply equations (17) and (18) to conclude that, for any $\epsilon > 0$,

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

The above argument and equation (17) allow us to conclude that $C_2(\theta_0) := \frac{\alpha_0^2}{2\sigma_0^2} \lim_{n \to \infty} \operatorname{Tr}[\Omega_n(\theta_0)M]/n$ exists and that

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

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 $\operatorname{Tr}(\Omega_n(\theta_0)^{-1} M) = n \frac{2\sigma_0^2}{1 + \alpha_0^2}$. Therefore, we have that

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

Since $n^{-1} \log p_{\theta_0}(y_i^n) \to_p H(\theta_0) \geq 0$, which minimizes entropy, and since $\Upsilon_n(q) \geq 0$, consequently, VI for $\alpha_0$ will be consistent iff

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

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 A.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 A.1 as

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

Conclude that VI is iff

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

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

27
A.2 Additional Results

Lemma A.1. Let

\[ \Omega_n := \begin{pmatrix}
    a & c & 0 & \ldots & 0 \\
    b & a & c & \ldots & 0 \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & 0 & \ldots & b \\
\end{pmatrix}, \quad a > 0, \ a^2 - 4bc \neq 0. \]

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 \( \Omega_k \) denoting the \( k \times k \) matrix, \( 1 < k < n \),

\[ f_n = a_nf_{n-1} - c_{n-1}b_{n-1}f_{n-2}, \]

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

\[ f_k = af_{k-1} - bc f_{k-2}, \]

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
\[ f_1 = a, \text{ and } f_2 = a^2 - bc, \text{ so that} \]
\[ a^2 - bc = a(f_1) - bc(f_0) = a^2 - bc(f_0), \]
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
\[
2a = k_1(a + d) + k_2(a - d) = a + (c_1 - c_2)d \\
\implies c_1 = c_2 + a/d \\
\implies 1 = 2c_2 + a/d \\
\implies c_2 = \frac{d - a}{2d} = -\frac{1}{d} \left(\frac{a - d}{2}\right) \\
\implies c_1 = \frac{a - d + 2a}{2d} = \frac{1}{d} \left(\frac{a + d}{2}\right)
\]
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
\[
f_n = f_{n-1} - bc f_{n-2}
\]
\[
= \frac{1}{d} \left[ \left(\frac{a + d}{2}\right)^n - \left(\frac{a - d}{2}\right)^n \right] - bc \frac{1}{d} \left[ \left(\frac{a + d}{2}\right)^{n-1} - \left(\frac{a - d}{2}\right)^{n-1} \right].
\]

\[ \blacksquare \]

### A.3 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 \( P(Y_{n+1}|y^n_1) \) the predictive distribution associated with the Bayesian predictive density \( p(Y_{n+1}|y^n_1) \).

The first scoring rule that we consider is the logarithmic score (LS), which is given by
\[
S_{LS}(P(Y_{n+1}|y^n_1), y_{n+1}) = \ln p(y_{n+1}|y^n_1). \tag{19}
\]
This score is favourable to predictive distributions that assign high probability mass to the realised value \( y_{n+1} \).

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
\[
S_{CS}(P(Y_{n+1}|y^n_1), y_{n+1}) = \ln p(y_{n+1}|y^n_1) I(y_{n+1} \in A) + \left[ \ln \int_{A^c} p(y|y^n_1) dy \right] I(y_{n+1} \in A^c). \tag{20}
\]
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

$$S_{\text{CRPS}}[P(Y_{n+1}|y^n_1), y_{n+1}] = -\int_{-\infty}^{\infty} [P(y|y^n_1) - I(y \geq y_{n+1})]^2 dy. \quad (21)$$

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

$$S_{\text{TWCRPS}}[P(Y_{n+1}|y^n_1), y_{n+1}] = -\int_0^1 2 \left[ I(P^{-1}(\alpha|y^n_1) \geq y_{n+1}) - \alpha \right] \left[ P^{-1}(\alpha|y^n_1) - y_{n+1} \right] (1 - \alpha)^2 d\alpha. \quad (22)$$

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 $100 (1 - \alpha)\%$ prediction interval, and given by

$$S_{\text{IS}}[P(Y_{n+1}|y^n_1), y_{n+1}] = -\left\{ u_{n+1} - l_{n+1} + \frac{2}{\alpha} (l_{n+1} - y_{n+1}) 1\{y_{n+1} < l_{n+1}\} + \frac{2}{\alpha} (y_{n+1} - u_{n+1}) 1\{y_{n+1} > u_{n+1}\} \right\},$$

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 $100 (1 - \alpha)\%$ predictive interval with $0 < \alpha < 1$. In this paper we set $\alpha = 0.05$.

**A.4 Additional Results**

This section contains additional numerical results for the expanding window predictive exercise in Section 6. For each method and each out-of-sample time point we compute the same eight measures of predictive accuracy, but consider out-of-sample evaluation periods of 100 (Table 3) and 1000 (Table 4) periods respectively.
Table 3: Predictive performance for the competing Bayesian approaches: exact Bayes, LSND, QNK and CY. The column labels indicate the out-of-sample predictive performance measure while the row labels indicate the predictive method. Panels A to C correspond to the results for DGP 1 to 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 A: DGP 2 | LS | CLS-10% | CLS-20% | CLS-80% | CLS-90% | CRPS | TWCRPS | MSIS |
|---------------|----|---------|---------|---------|---------|------|--------|------|
| True DGP      | -0.790 | -0.160 | -0.351 | -0.552 | -0.313 | -0.326 | -0.093 | -3.743 |
| Exact Bayes   | -0.789 | -0.158 | -0.352 | -0.547 | -0.304 | -0.328 | -0.094 | -3.687 |
| LSND          | -0.788 | -0.157 | -0.347 | -0.554 | -0.315 | -0.327 | -0.094 | -3.702 |
| CY            | -0.807 | -0.166 | -0.360 | -0.547 | -0.302 | -0.329 | -0.096 | -3.659 |
| QNK           | -0.802 | -0.172 | -0.361 | -0.555 | -0.313 | -0.328 | -0.094 | -3.920 |

| Panel A: DGP 3 | 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 4: Predictive performance for the competing Bayesian approaches: exact Bayes, LSND, QNK and CY. The column labels indicate the out-of-sample predictive performance measure while the row labels indicate the predictive method. Panels A to C correspond to the results for DGP 1 to 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 A: DGP 2

|         | LS  | CLS-10% | CLS-20% | CLS-80% | CLS-90% | CRPS  | TWCRPS | MSIS  |
|---------|-----|---------|---------|---------|---------|-------|--------|-------|
| True DGP| -1.010 | -0.370 | -0.568 | -0.557 | -0.347 | -0.383 | -0.117 | -3.674 |
| Exact Bayes | -1.015 | -0.373 | -0.571 | -0.557 | -0.347 | -0.384 | -0.117 | -3.699 |
| LSND    | -1.018 | -0.375 | -0.573 | -0.558 | -0.348 | -0.384 | -0.117 | -3.691 |
| CY      | -1.026 | -0.376 | -0.576 | -0.561 | -0.350 | -0.385 | -0.118 | -3.796 |
| QNK     | -1.031 | -0.378 | -0.576 | -0.565 | -0.354 | -0.386 | -0.118 | -3.775 |

### Panel A: 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 |

32