Variational Bayes in State Space Models: Inferential and Predictive Accuracy

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

Department of Econometrics and Business Statistics, Monash University, Clayton, Australia

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 adequately approximate the states yielding superior accuracy over methods that do not. We also document numerically that over small out-of-sample evaluation periods the inferential discrepancies between the various methods often yield only small discrepancies in predictive accuracy. Nevertheless, in certain settings, and over a longer out-of-sample period, these predictive discrepancies can become meaningful. 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. Supplementary materials for this article are available online.

1. Introduction
A common class of models used for time series modeling 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, Koopman, and Shephard 2004; Giordani, Pitt, and Kohn 2011, for extensive reviews).

The key feature of SSMs is their dependence on hidden, or latent, “local” variables, also called states, which govern the dependence of the observed data, in conjunction with a vector of unknown “global” parameters. This feature leads to inferential challenges with, for example, the likelihood function for the global parameters being analytically unavailable, except in special cases. Although 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; Aït-Sahalia and Kimmel 2007; Aït-Sahalia et al. 2021, amongst others), it is arguable that Bayesian Markov chain Monte Carlo (MCMC) methods have become the most common tool for analyzing general SSMs, with such techniques expanded in more recent times to accommodate particle filtering, via pseudo-marginal variants such as particle MCMC (PMCMC) (Andrieu, Doucet, and Holenstein 2011; Flury and Shephard 2011). See Giordani, Pitt, and Kohn (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, and would be viewed as the gold standard, they do suffer from certain well-known limitations. First: they require that either the (complete) likelihood function is available in closed form or that an unbiased estimator of it is available. Second: they do not necessarily scale well to high-dimensional problems; for example, to models with multiple observed and/or state processes (Rue, Martino, and Chopin 2009; Braun and McAuliffe 2010; Betancourt 2018; Johndrow et al. 2019) or, indeed, to problems characterized by “big” or “tall,” datasets alone (Bardenet, Doucet, and Holmes 2017). That is, whilst the construction of an MCMC algorithm may well be feasible in high-dimensional problems, convergent draws may not be possible to produce in a reasonable computing time.

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), and the need to simulate repeated sets of data from the assumed DGP impacts on computational speed when the sample size is large.

Variational Bayes (VB) methods (see Blei, Kucukelbir, and McAuliffe 2017; Zhang et al. 2018, for reviews) 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 SSMs, by exploiting modern optimization-based techniques instead of relying on posterior simulation. In short, in cases where the computational cost of MCMC methods (in particular) has been deemed to be prohibitive, VB has been used to
produce inferential and predictive results within an acceptable time frame (Tran, Nott, and Kohn 2017; Quiroz, Nott, and Kohn 2022; Koop and Korobilis 2020; Chan and Yu 2022; Loaiza-Mayo et al. 2022). The price paid for computational speed is that posterior inference and prediction is intrinsically approximate; that is, the gold standard of exact Bayesian analysis is sacrificed. While this may be seen as a necessary compromise when conducting Bayesian analysis in large and challenging SSMs, there is still room to investigate the nature and extent of the sacrifice made; and that is the goal of this article.

In summary, we make three contributions to the literature on the application of VB to SSMs. The first contribution is to highlight the fundamental issue that lies at the heart of the use of VB in an SSM setting, linking this to an existing issue identified in the literature as the “incidental parameter problem” (Neyman and Scott 1948; Lancaster 2000; Westling and McCormick 2019). In brief, without due care, the application of VB to the local parameters in an SSM leads to a lack of Bayesian consistency for the global parameters. Moreover, in a class of common SSMs, we demonstrate analytically the impact of this inconsistency on the resulting state inference, and show that even in idealized settings inconsistent inference for the global parameters can lead to highly inaccurate inferences for the local parameters (i.e., the states). The second contribution is to examine existing variational methods currently in use within SSMs, and to link their prospects for consistency to the manner in which they do, or do not, circumvent the incidental parameter problem. Third, we undertake a numerical comparison of several competing variational methods, in terms of both inferential and predictive accuracy. The key findings are that: (a) correct management of the local variables leads to inferential accuracy that closely matches that of exact (MCMC-based) Bayes; (b) inadequate treatment of the local variables leads, in contrast, to noticeably less accurate inference; (c) VB-based predictions demonstrate robustness to inferential inaccuracy, but only at small sample sizes. For large sample sizes, 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 comparable out-of-sample accuracy to exact Bayes in some settings.

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

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 $\to_p$ have their usual meaning. Similarly, we let $\lim_{n} X_n = c$ denote $X_n \to_p c$. We let $d(\cdot, \cdot)$ denote a metric on $\Theta \subseteq \mathbb{R}^{d_\Theta}$. The proofs of all theoretical results, certain definitions, plus additional tables and figures, are included in the supplementary materials.

## 2. State Space Models: Exact Inference

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

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

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

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

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

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

The (average) observed data log-likelihood is thus

$$\ell_n(\theta) := \frac{1}{n} \log p_\theta(y^n) = \frac{1}{n} \log \int p_\theta(x^n_1|x^n_1)d\pi_\theta,$$  \hspace{1cm} (3)

and the maximum likelihood estimator (MLE) of $\theta$ is $\hat{\theta}^{\text{MLE}} = \arg\max_{\theta \in \Theta} \ell_n(\theta)$. As is standard knowledge, $\ell_n(\theta)$ is able in closed form only for particular choices of $g_\theta(y_t|x_t)$ and $\chi_\theta(x_{t+1}, x_t)$: the canonical example being when (2) and (1) define a linear Gaussian state space model (LGSSM). Similarly, for $p(\theta)$ denoting the prior density, the exact (marginal) posterior for $\theta$, defined as

$$\pi(\theta|y^n) = \int \pi(\theta, x^n_1|y^n)d\pi_\theta,$$

$$\pi(\theta, x^n_1|y^n) \propto p(y^n_1|x^n_1, \theta)p(x^n_1|\theta)p(\theta),$$  \hspace{1cm} (4)

is available (e.g., via straightforward MCMC methods) only in limited cases, the LGSSM being one such case. In more complex settings and—in particular—settings where either $\theta$ or $(\{X_t, Y_t\})$, or both, are high-dimensional, accessing (4) can be difficult. In particular, standard MCMC methods can result in a poor exploration of the joint space for the unknowns, due to either slow mixing of the MCMC chains, and/or to a computationally prohibitive run-time for likelihood evaluation, with the end result being potentially unreliable inferences, for a finite computing budget (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 article, 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 behavior 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 \( \pi(x_n^\theta, \Theta | y_n^T) \) 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 \( \pi(x_n^\theta, \Theta | y_n^T) \) by minimizing the KL divergence between a family of densities \( Q \), with generic element \( q(x_n^\theta, \Theta) \), and \( \pi \):

\[
\text{KL}(Q||\pi) = \int q(x_n^\theta, \Theta) \log \frac{q(x_n^\theta, \Theta)}{\pi(x_n^\theta, \Theta | y_n^T)} dx_n^\theta d\Theta.
\]

Optimizing the KL divergence directly is not feasible since it depends on the unknown \( \pi(x_n^\theta, \Theta | y_n^T) \); 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):

\[
\text{ELBO}(Q||\pi) := \int q(x_n^\theta, \Theta) \log \frac{p(y_n^T | x_n^\theta, \Theta)p(x_n^\theta | \Theta)p(\Theta)}{q(x_n^\theta, \Theta)} dx_n^\theta d\Theta,
\]

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 \( Q = \{ q : q(x_n^\theta, \Theta) = q_\theta(\Theta)q_x(x_n^\theta|\Theta) \} \), with \( Q \) often restricted to be mean-field, that is, \( \theta \) independent of \( \theta_j \), \( i \neq j \), and \( x_n^\theta \) independent of \( \theta \).

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

\[
\text{ELBO}(Q||\pi) = \int_\Theta \int_{\mathcal{X}} q_\theta(\Theta)q_x(x_n^\theta|\Theta) \log \frac{p(y_n^T | x_n^\theta, \Theta)p(x_n^\theta | \Theta)p(\Theta)}{q_\theta(\Theta)q_x(x_n^\theta | \Theta)} dx_n^\theta d\Theta
\]

\[
= \int_\Theta \int_{\mathcal{X}} q_\theta(\Theta)q_x(x_n^\theta|\Theta) \left[ \log \frac{p(y_n^T | x_n^\theta, \Theta)p(x_n^\theta | \Theta)}{q_x(x_n^\theta | \Theta)} \right] dx_n^\theta d\Theta
\]

\[
- \text{KL}(q_\theta(\Theta)||p(\Theta)),
\]

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

\[
\mathcal{L}_n(\Theta, q_x) := \int_{\mathcal{X}} q_x(x_n^\theta|\Theta) \log \frac{p(y_n^T | x_n^\theta, \Theta)p(x_n^\theta | \Theta)}{q_x(x_n^\theta | \Theta)} dx_n^\theta,
\]

by Jensen’s inequality

\[
\log p_\theta(y_n^T) = \log \int_{\mathcal{X}} p_\theta(y_n^T, x_n^\theta) dx_n^\theta
\]

\[
= \log \int_{\mathcal{X}} q_x(x_n^\theta|\Theta) \left\{ \frac{p(y_n^T | x_n^\theta, \Theta)p(x_n^\theta | \Theta)}{q_x(x_n^\theta | \Theta)} \right\} dx_n^\theta
\]

\[
\geq \mathcal{L}_n(\Theta, q_x),
\]

where we use the shorthand \( q_x \) to denote \( q_x(x_n^\theta | \Theta) \). Thus, \( \mathcal{L}_n(\Theta, q_x) \) can be viewed as an approximation (from below) to the observed data log-likelihood. Defining

\[
\Upsilon_n(q) := \int_\Theta \left\{ \log p_\theta(y_n^T) - \mathcal{L}_n(\Theta, q_x) \right\} q_\theta(\Theta)d\Theta,
\]

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

\[
\text{ELBO}(Q||\pi) = \int_\Theta \log p_\theta(y_n^T)q_\theta(\Theta)d\Theta
\]

\[
- \Upsilon_n(q) - \text{KL}(p_\theta(\Theta)||q_\theta(\Theta)).
\]

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, Pati, and Bhattacharya (2020) refer to as the average (with respect to \( q_\theta(\Theta) \)) “Jensen’s gap,” which encapsulates the error introduced by approximating the latent states using a given variational class. While the first and last term in the decomposition can easily be controlled by choosing an appropriate class for \( q_\theta(\Theta) \), it is the average Jensen’s gap that ultimately determines the behavior of the variational approximation for the posterior.

#### 3.2. Consistency of Variational Point Estimators

The decomposition in (9) has specific implications for variational inference in SSMs, which can be readily seen if we consider the case where we only employ a variational approximation for the states, while inference for \( \theta \) is conducted via point estimation. That is, in this section, we follow Westling and McCormick (2019) and consider the theoretical behavior of frequentist (point) inferences for the global parameters, \( \Theta \), when conducting variational inference for the states, \( x_n^\theta \).

In this case, we do not wish to approximate the posterior for \( \theta \), but only obtain point inferences, which still necessitates approximating the posterior for the states.2 Following Wang and Blei (2018), we can then define the “variational log-likelihood” as

\[
\mathcal{M}_n(\Theta, q_x) := \ell_n(\Theta | q_x) / n,
\]

where \( \ell_n(\Theta, q_x) / n \) incorporates the criterion’s dependence on Jensen’s gap, and the “variational (point) estimators” are

\[
(\hat{\theta}_n, \hat{q}_x) := \arg \max_{\Theta \in \Theta, Q} \mathcal{M}_n(\Theta, q_x).
\]

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

---

1We note here that the analysis of Westling and McCormick (2019) is restricted to cross-sectional settings with independent and identically distributed states, while we explicitly consider temporally dependent data and latent states.

2One interpretation of this approach is that we are employing the variational family \( Q := \{ q(\theta, x_n^\theta) = \delta_{\theta} \times q_x(x_n^\theta|\Theta) \} \), where \( \delta_{\theta} \) is the Dirac delta function at \( \theta \), to conduct joint inferences.
Assumption 3.1. (i) The parameter space $\Theta$ is compact. (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 article 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 unnecessary 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 $\kappa_n := \frac{1}{n} \inf_{q_{x} \in \mathcal{Q}_x} \Upsilon_n(\theta_0, q_{x})$, and note that $\kappa_n \geq 0$. If Assumption 3.1 is satisfied, and if $\kappa_n = o_p(1)$, then $\hat{\theta}_n \rightarrow_p \theta_0$.

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, that is, $n$.

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.

Lemma 3.1 implies that more complex variational families can produce more accurate inferences in at least two situations: if the additional complexity allows $\mathcal{Q}$ to encompass $\pi(x^n_t|y^n_t, \theta)$, so that $\hat{q}_x = \pi(x^n_t|y^n_t, \theta)$, and $\Upsilon_n(\theta, \hat{q}_x) = 0$; if $\mathcal{Q}$ is such that the rate of growth for the error induced by approximation is of order smaller than $n$. If these conditions are not met, additional flexibility for $\mathcal{Q}$ is unlikely to produce more accurate inference on the global parameters.

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, $Y_t = \alpha X_t + \sigma_0 \epsilon_t$, with state equation $X_{t+1} = \rho X_t + \sigma_0 \epsilon_t$, $X_t \sim N(0, \sigma^2_0)$, and with $\{\epsilon_t\}$ and $\{\epsilon_t\}$ independent sequences of iid standard normal random variables. We observe a sequence $\{Y_t\}$ from the above model, but the states $\{X_t\}$ are unobserved. Furthermore, consider that $\theta = (\rho, \alpha)'$ is unknown while $\sigma_0$ is known.

We make use of the autoregressive nature of the state process to approximate the posterior for $\pi(x^n_t|\theta, y^n_t)$ via the variational family: $\mathcal{Q}_x := \{\lambda \in [0, 1] : q_x(x^n_t|\sigma_0) = N(x^n_t|0, \nu(\lambda) \Phi_n(\lambda))\}$.

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

Lemma 3.2. Let $\sigma_0 > 0$, $0 \leq |\rho_0| < 1$ and $0 \leq |\alpha_0| < M$. Assume the parameter space $\Theta$ is compact. (i) If $\rho_0 = 0$ and known, then the variational point estimator $\hat{\alpha}$ is consistent if and only if $\alpha_0 = 0$. (ii) If $\alpha_0 = 0$ and known, then the variational point estimator for $\hat{\rho}$ is consistent if and only if $\rho_0 = 0$.

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

3.3. Lack of Bayes Consistency of the Variational Posterior

While the above analysis pertains to variational point estimators of $\theta_0$, a similar result obtains in terms of a variational posterior. To state this result, consider the canonical case where we approximate the state posterior using the class of variational approximations $q_{x}(x^n_t) := q_{x}(x^n_t)$, to Equation (7), the explicit dependence of the class of variational approximations $q_{x}(x^n_t)$ on $\lambda$. The explicit dependence of $q_{x}(x^n_t)$ on $\lambda$ yields the criterion $\mathcal{L}_n(\theta, \lambda)$, with $q_{x}(x^n_t|\theta)$ in (7) replaced by $q_{x}(x^n_t)$.

To understand the implications of approximating the state posterior using $q_{x}(x^n_t)$, we follow Wang and Blei (2018, 2019) and consider the behavior of the so-called “ideal variational posterior,” or “VB ideal” for short, which is, itself, an approximation that sits between the exact posterior and the genuine variational posterior. The VB ideal is a generalized posterior, in the sense of Bissiri, Holmes, and Walker (2016), that results from first optimizing over $\lambda$ for fixed $\theta$, to produce the profiled criterion

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

then exponentiating $\tilde{L}_n(\theta)$ and normalizing this criterion to produce a generalized posterior for $\theta$ as

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

Unlike with the frequentist optimization problem, the idealized VB posterior incorporates a component of Jensen’s gap directly into the definition of that posterior.$^5$

$^3$A condition of this kind is known to be satisfied in certain models with iid observations that are driven by latent variables, such as mixture models and some random effects models; see Wang and Blei (2016) for a discussion. However, it is unlikely that a similar condition is satisfied in situations with temporally dependent states. We conjecture that this might be the case in multivariate models where the observables are driven by the same set of unobservable factors, but further exploration is required to determine the validity of this conjecture.

$^4$In this same linear Gaussian model, a similar result was derived by Turner and Sahani (2011) for inference performed using variational expectation maximization with a mean-field variational family.

$^5$Wang and Blei (2018) demonstrate that, asymptotically, the difference between the variational ideal and the genuine variational posterior for
Given that \( \hat{q}(\theta | y^n) \) is constructed from a profiled criterion, the ideal variational posterior is a Bayesian version of the frequentist profiled variational inference approach described in Westling and McCormick (2019). In their analysis, the authors view variational point estimators of the global parameters \( \theta \) as \( M \)-estimators based on the profiled variational criterion function in (10). They then explore conditions and examples under which the variational point estimator, based on maximizing \( \tilde{L}_n(\theta) \), is, or is not, consistent for \( \theta_0 \).

While Westling and McCormick (2019) focus on consistency of variational point estimators, we study concentration of the posterior distribution \( \hat{q}(\theta | y^n) \). The following result shows that, under regularity conditions similar to those maintained in Westling and McCormick (2019), \( \hat{q}(\theta | y^n) \) is Bayes consistent for some value that may or may not coincide with \( \theta_0 \).

**Assumption 3.2.** (i) There exists a map \( \theta \mapsto \lambda(\theta) \in \Lambda \) such that \( \sup_{\theta \in \Theta} \| L_n(\theta) - L_0(\theta) \| = o_p(1) \). (ii) There exist a deterministic function \( L : \Theta \times \Lambda \mapsto \mathbb{R} \) and a \( \theta_* \in \Theta \) such that the following are satisfied: (a) for all \( \epsilon > 0 \) there exists some \( \delta > 0 \) such that \( \inf_{\theta \in \Theta} \{ L(\theta, \lambda(\theta)) - L(\theta_0, \lambda_0(\theta)) \} \leq -\delta \); (b) \( \sup_{\theta \in \Theta, \lambda \in \Lambda} \| L_n(\theta, \lambda) - L(\theta, \lambda(\theta)) \| = o_p(1) \). (iii) For any \( \epsilon > 0 \), \( \int_0^\infty \mathcal{I}(\theta : L(\theta, \lambda) - L(\theta_0, \lambda_0(\theta)) < \epsilon) p(\theta) d\theta > 0 \). (iv) For all \( n \) large, \( \int_0^\infty \mathcal{I}(\theta : d(\theta_0, \theta) > \epsilon) |y^n|^{2n} = o_p(1) \).

**Lemma 3.3.** Under Assumption 3.2, for any \( \epsilon > 0 \), \( \hat{Q}(\theta | y^n) = \hat{Q}(\theta : d(\theta, \theta_*) > \epsilon) |y^n|^{2n} = o_p(1) \).

**Corollary 1.** For \( \theta_n = \int_0^\infty \theta d\hat{Q}(\theta) \), under Assumption 3.2, \( \| \theta_n - \theta_0 \| \leq \| \theta - \theta_* \| + o_p(1) \).

When \( \theta_0 \neq \theta_* \), Lemma 3.3 demonstrates that the idealized posterior for \( \theta \) will not concentrate onto \( \theta_0 \). In this case, Corollary 3.1 demonstrates that the point inferences for \( \theta_0 \) are biased, and do not coincide with those obtained from exact methods.

These results can be interpreted explicitly in terms of Jensen’s gap as defined in (8) by recalling that under Assumption 3.1, \( \ell_n(\theta_0) \rightarrow_{p} H(\theta_0) \), and by considering the limit of the (scaled) Jensen’s gap evaluated at \( \theta_0 \),

\[
\lim_{n \to \infty} \frac{1}{n} \gamma_n \left( \theta_0, q_{\theta_0}(\theta_0) \right) = H(\theta_0) - L(\theta_0, \lambda(\theta_0)) \\
\geq H(\theta_0) - L(\theta_0, \lambda_0(\theta_0)) + \delta,
\]

for some \( \delta \geq 0 \). If Assumption 3.2(ii.a) is satisfied at \( \theta_0 \neq \theta_0 \), then \( \delta > 0 \), and \( \kappa_n := \frac{\gamma_n(\theta_0, q_{\theta_0}(\theta_0))}{n} \rightarrow \infty \). \(^{7}\)

Taken together, Lemmas 3.1 and 3.3, and Corollary 3.1, 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. The point is further exemplified in the follow example, where we explore the Bayesian consistency of the ideal variational posterior in the same linear Gaussian SSM.

**Example 3.2** (Linear Gaussian model revisited). Returning to the linear Gaussian SSM in Example 3.1, let us again consider the case where \( \theta = (\alpha, \rho)^T \) is unknown, while \( \sigma_0 \) is known, and consider variational inference for \( \theta \) using the variational family as in the form of Example 3.1, let us again consider the case where \( \theta = (\alpha, \rho)^T \) is unknown, while \( \sigma_0 \) is known, and consider variational inference for \( \theta \) using the variational family as in the form of Example 3.1. Moreover, the marginal state posterior \( \pi(x_n | y^n, \theta) \) is also known in closed form, for a given value of \( \theta \), and any \( n \geq 2 \). Given this, we can analytically evaluate the KL divergence between \( q_{\theta}(x_n) \) and \( \pi(x_n | y^n, \theta) \), at any \( n \), to characterize the accuracy of the resulting state approximation.

**Corollary 2.** Under \( \hat{Q}_x \), for any \( n \geq 2 \), we have \( \text{KL}[\pi(x_n | y^n, \theta_0) | q_{\theta}(x_n)] > 0 \).

\(^{6}\)Similar to the regularity conditions in Assumption 3.1, low-level sufficient conditions that imply Assumption 3.2 can be obtained using the result of Douc et al. (2011), and are known to be valid for many different classes of SSMs, including multivariate linear Gaussian models and certain stochastic volatility models; see Douc et al. (2011) for details. Given that this is not our main focus, we eschew these technical details and maintain the high-level conditions in Assumption 3.2.

\(^{7}\)It is worth pointing out that a smaller numerical value of Jensen’s gap is not always better. That is, given two different approximation methods that deliver two different nonzero values of Jensen’s gap, it is not always the case that the method that produces the smaller gap delivers point inferences that are more accurate; see Rainforth et al. (2018) for a discussion.
The above result demonstrates that for any $n \geq 2$, the optimal variational state density is a biased approximation of the exact state density. Thus, even if $\theta_0$ were known, and we only wished to conduct inference on $x_{1}^{n}$, the resulting variational approximation of the state density would ultimately deliver a poor approximation.

While it is well-known that VB delivers poor uncertainty quantification, it is generally believed that VB delivers reasonable estimators for the unknown locations. *Lemmas 3.1, and 3.3, and Corollary 3.1* demonstrate, in general terms, that variational methods produce inaccurate point inferences for the global parameters in SSMs, due to the necessary approximation of the state posterior; while the specific results in the running example (Lemma 3.4 and Corollary 3.2) demonstrate that this general result is present even in the simplest of SSMs, the linear Gaussian model.

### 4. Implications

The above results suggest that VB methods can produce inaccurate parameter and state inference in SSMs, both in terms of posterior location and uncertainty quantification. While the poor uncertainty quantification of VB has been long known (as noted above), the results regarding the inaccuracy of the VB posterior locations are the only general results of which we are aware in the context of SSMs. In this section, we discuss further the implications of these results for inference on the global parameters, plus their implications for predictive accuracy.

#### 4.1. Inference on Global Parameters

When conducting VB in SSMs, the need to approximate the posterior of $x_{1}^{n}$ introduces a discrepancy between the exact posterior, and that which results from the VB approach. In this way, we can view the latent states $x_{1}^{n}$ as incidental or nuisance parameters (see Lancaster 2000, for a review), which are needed to make feasible the overall optimization problem, but which, in and of themselves, are not the object of interest. A similar point is made by Westling and McCormick (2019) in the case of independent states, and frequentist variational inference, where the authors demonstrate that inconsistency can occur, even in the case of independent observations, if delicate care is not taken with the choice of variational class for $x_{1}^{n}$.

However, the incidental parameter problem has not stopped researchers from using VB methods to conduct inference on $\theta$ in SSMs. While the general conclusions elucidated above apply, in principle, to all such methods, we next discuss two specific categories of VB methods in greater detail, and comment on their ability to deliver consistent inference for $\theta_0$.

#### 4.1.1. Integration Approaches

A possible VB approach is to first “integrate out” the latent states so that there is no need to perform joint inference on $(\theta, x_{1}^{n})$. 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 $KL(q||\pi)$ as

$$KL(q||\pi) = \int \int_{X} q_{\theta}(\theta) \pi(x_{1}^{n}|y_{1}^{n}, \theta)$$

$$\times \log \frac{q_{\theta}(\theta) \pi(x_{1}^{n}|y_{1}^{n}, \theta)}{\pi(x_{1}^{n}|y_{1}^{n}, \theta)} \pi(y_{1}^{n}|\theta) \pi(x_{1}^{n}|\theta) \pi(x_{1}^{n}|\theta)$$

$$= KL[q_{\theta}||\pi(\theta|y_{1}^{n})].$$

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. (2022), and is applicable in any case where draws from $\pi(x_{1}^{n}|y_{1}^{n}, \theta)$ 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. (2022) results in the above simplification, the real key to their approach is that it can be used to unbiasedly estimate the gradient of 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, Shephard, and Chib 1998; Jacquier, Polson, and Rossi 2002; Primiceri 2005; Huber, Koop, and Onorante 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 $p_{0}(y_{1}^{n})$ using a particle filter, which we denote by $\hat{p}_{0}(y_{1}^{n})$. We follow Tran, Nott, and Kohn (2017) and write $\hat{p}_{0}(y_{1}^{n})$ 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 $z = \log \hat{p}(y_{1}^{n}|\theta) - \log p_{0}(y_{1}^{n})$. For $g(z|\theta)$ denoting the density of $z|\theta$, Tran, Nott, and Kohn (2017) consider VB for the augmented posterior

$$\pi(\theta, z|y_{1}^{n}) = \hat{p}(y_{1}^{n}|\theta, z)g(z|\theta)p(\theta)/p(y_{1}^{n}) = p_{0}(y_{1}^{n}) \exp \{z \hat{g}(z|\theta)p(\theta)/p(y_{1}^{n})\}$$

$$= \pi(\theta|y_{1}^{n}) \exp \{z \hat{g}(z|\theta),$$

which, marginal of $z$, has the correct target posterior $\pi(\theta|y_{1}^{n})$ due to the unbiasedness of the estimator $\hat{p}(y_{1}^{n}|\theta, z)$. The authors refer to the resulting method as variational Bayes with an intractable likelihood function (VBIL). The VBIL posteriors can be obtained by considering a variational approximation to
\pi(\theta, z | \eta_0)\) that minimizes the KL divergence between \(q(\theta, z) = q_0(\theta)g(z | \theta)\) and \(\pi(\theta, z | \eta_0)\):

\[
KL(q(\theta, z) || \pi(\theta, z | \eta_0)) = \int_\theta \int_z q_\theta(z | \theta) \log \frac{q_\theta(z | \theta)}{\pi(\theta | \eta_0)} \exp(z g(z | \theta)) \, dz \, d\theta = \int_\theta \int_z q_\theta(z | \theta) \log \frac{q_\theta(z | \theta)}{p_\theta(y_0 | \theta)} \exp(z p(\theta)) \, dz \, d\theta + \log p(y_0) = -\int_\theta q_\theta(\theta) \log p_\theta(y_0 | \theta) \, d\theta + KL(q(\theta, z) || \pi(\theta, z | \eta_0))
\]

where in this case \(\mathcal{T}_n[q(\theta, z)] = \int_\theta \int_z q_\theta(z | \theta) \log \frac{p_\theta(y_0 | \theta)}{q_\theta(z | \theta)} \, dz \, d\theta\). For fixed \(\theta\), \(\mathbb{E}_z [p(y_0 | \theta, z)] = p_\theta(y_0 | \theta)\), but in general \(\log p_\theta(y_0 | \theta, z)\) is a biased estimator of \(p_\theta(y_0 | \theta)\), from which it follows that \(\mathcal{T}_n[q(\theta, z)] \geq 0\). However, in contrast to the general approximation of the states discussed in Section 3, which intimately relies on the choice of the approximating density \(q_\theta(\eta_0 | \theta)\), VBIL can achieve consistent inference on \(\theta_0\) by choosing an appropriate number of particles \(N\) in the production of \(p(\eta_0 | \theta, z)\).

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, for example, Assumption 1 in Doucet et al. (2015) and Assumption 1 in Tran, Nott, and Kohn (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 \(\epsilon > 0\)

\[
\lim_{n \to \infty} \mathbb{P}_n\{\mathcal{T}_n[q(\theta_0, z)] / n > \epsilon\} = \lim_{n \to \infty} \mathbb{P}_n\{-q(\theta_0) \mathbb{E}[z | \theta_0] > ne\} = \lim_{n \to \infty} \mathbb{P}_n\{q(\theta_0)\sigma^2 / 2 > ne\} = 0,
\]

assuming \(q(\theta_0), \sigma^2 < \infty\).

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 \(\theta\).

### 4.1.2. Structured Approximations of the States

Yet another approach for dealing with variational inference in the presence of states is to consider a structured approximation that allows for a dynamic updating of the 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 (2020) propose the use of the Kalman filter within VB (VBKF) as a means of approximating the posterior density of the states using Kalman recursions. In particular, the authors approximate the posterior \(\pi(x_t | y_{t-1}, \theta)\) by approximating the relationship between \(X_t\) and \(X_{t-1}\), which may in truth be nonlinear in \(\theta\), by the random walk model \(X_t = X_{t-1} + \epsilon_t\), with \(\epsilon_t \sim \text{iid} N(0, \sigma^2)\), and then use Kalman filtering to update the states in conjunction with a linear approximation to the measurement equation. Using this formulation, the variational approximation is of the form \(q(x_t | y_{t-1}, \theta) = q_0(x_t | y_{t-1}, \theta)\), where \(q_0(x_t | y_{t-1}, \theta) \propto \prod_{k=1}^T \exp\left(-\frac{1}{2} (x_k - \tilde{x}_k)^2 (1 - K_k) P_{k-1/2}\right)\) and where the terms \(K_k, P_{k-1/2}\) are explicitly calculated using the Kalman recursion: \(\tilde{x}_k = \tilde{x}_{k-1} + K_k (y_t - \tilde{x}_{k-1})\), and where \(K_k, P_{k-1/2}\) is the Kalman gain, \(P_{k-1/2}\) is the predicted variance of the state, and in the application of Koop and Korobilis (2020), \(y_t^2 = \log(y_t^2)\).

While the solution proposed by the VBKF is likely to lead to more accurate inferences on the states than standard VB, especially when \(x_t\) behaves like a random walk, ultimately we are still “conducting inference” on \(x_t\), and thus we still encounter the incidental parameter problem as a consequence. Indeed, taking as the variational family for \(x_t\) the Kalman filter approximation yields, at time \(k \geq 1\), a conditionally normal density with mean \(\tilde{x}_k\) and variance \((1 - K_k) P_{k-1/2}\). 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 (2022) 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 \(q_\theta(x_t)\). As with the VBKF approach, the assumed nature of the state process used by Chan and Yu (2022) to estimate \(q_\theta(x_t)\) implies that, in general, it is unlikely that Bayesian consistency can be achieved. Due to space restrictions, further discussion on the specifics of this approach are relegated to Sections A.1.1 and B.1.4 of the supplementary materials.

### 4.2. VB-based Prediction

VB provides, at best, an approximation to the posterior and, as a result, may well yield less accurate inferences than those produced by the exact posterior (see, e.g., Koop and Korobilis 2020; Gunawan, Kohn, and Nott 2021). However, VB can perform admirably in predictive settings, see, for example, Quiroz, Nott, and Kohn (2022) and Frazier et al. (2021), amongst others, in the sense of replicating the out-of-sample accuracy achieved by exact predictives, when such comparators are available. (See Frazier et al. 2019 for a comparable finding in the context of predictions based on ABC.) Therefore, even though the VB posterior may not necessarily converge to the true value \(\theta_0\), so long as the value onto which it is concentrating is not too far away from \(\theta_0\), it may be that VB-based predictions perform well in practice.

Recall the conditional density of \(Y_{n+1}\) given \(x_{n+1}\) and \(\theta\) is

\[
p(Y_{n+1} | x_{n+1}, \theta) = \int_\theta \int_x g_\theta(Y_{n+1} | x_{n+1}, \pi(x_{n+1}, \theta) \, dx \, d\theta
\]

where \(\pi(x_{n+1}, \theta)\) is the predictive density for \(x_{n+1}\) which can be expressed as

\[
p(Y_{n+1} | x_{n+1}, \theta) = \int_\theta \int_x g_\theta(Y_{n+1} | x_{n+1}, \pi(x_{n+1}, \theta) \, dx \, d\theta
\]
where the last line follows from the Markovianity of the state transition equation (see Equation (2)). In many large SSMs, using MCMC methods to estimate (11) is infeasible or prohibitive computationally, due to the difficulty of sampling from \( \pi(\theta) \). Instead, VB methods can produce an estimate of \( p(Y_{n+1} | y_n^T) \) by approximating, in various ways, the two pieces in (11) underlined as (1) and (2). All such methods replace the second underlined term by some approximate posterior for \( x_t^T \). However, it is not necessarily clear that the resulting predictions will necessarily perform much better than those approaches based on the approximation \( \hat{q}_t \). In the following section, we explore this question.

5. Numerical Assessment of VB Methods

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

5.1. Simulation Design

We now compare the inferential and predictive accuracy of the Gaussian variational approximation suggested by Quiroz, Nott, and Kohn (2022) for use in high-dimensional state space models, against the approach suggested by Loaiza-Maya et al. (2022) and against an exact MCMC-based estimate of \( \pi(\theta, x_t^T | y_n^T) \), referred to as “exact Bayes” hereafter, in a simulation exercise. In Section B.1 of supplementary materials, we provide complete details on the implementation of each of these methods under this particular simulation design. However, we remark here that the Gaussian approximation suggested in Quiroz, Nott, and Kohn (2022) is an example of a VB method in which the states are approximated via a particular choice of variational family, whilst Loaiza-Maya et al. (2022) (as noted in Section 4.1.1) adopt a variational approximation for the posterior of the global parameters only, with the conditional posterior of the states accessed via simulation.

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

\[
\begin{align*}
\mu_t &= \bar{\mu} + \rho_{\mu} (\mu_{t-1} - \bar{\mu}) + \sigma_{\mu} \epsilon_t, \\
h_t &= h_0 + \rho_h (h_{t-1} - h_0) + \sigma_h \eta_t, \\
Y_t &= \mu_t + \exp(h_t/2) u_t, \\
\end{align*}
\]

where \( (\epsilon_t, \eta_t, u_t) \overset{iid}{\sim} N(0, I_3) \). 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 sets of values for the true parameters:

DGP 1: \( \mu_0 = 0; \rho_{\mu} = 0.8; \sigma_{\mu} = 0.5; h_0 = -1.0; \rho_h = 0.00; \sigma_h = 0.0 \)

DGP 2: \( \mu_0 = 0; \rho_{\mu} = 0.0; \sigma_{\mu} = 0.5; h_0 = -1.3; \rho_h = 0.95; \sigma_h = 0.3 \)

DGP 3: \( \mu_0 = 0; \rho_{\mu} = 0.8; \sigma_{\mu} = 0.5; h_0 = -1.3; \rho_h = 0.95; \sigma_h = 0.3 \).

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 marginal mean of zero; while 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 \( \theta_0 = (\mu_0, \rho_{\mu}, \sigma_{\mu}, h_0, \rho_h, \sigma_h)' \).

For the predictive assessment we compare exact Bayes with the two variational methods cited above plus the method of Chan and Yu (2022). As discussed in Section B.1 of the supplementary materials the method of Chan and Yu (2022) exploits a very specific structure in the construction of the variational algorithm, which in this case corresponds to DGP 2 under the parameter restrictions \( \rho_h = 1.0, \sigma_{\mu} = 0.0 \) and \( h_0 = 0.0 \). Thus, application of this approach under any of the above true DGPs constitutes misspecified inference; hence, we do not include this technique in the inferential assessment. Due to space constraints, certain tables and figures are included in Section B.2 of the supplementary materials.

5.2. Accuracy of Inference on the States

We assess inferential accuracy through lens of state estimation. To this end, we generate a times series of length \( T = 11,000 \) 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 based, respectively, on the Gaussian
Table 1. Accuracy in the estimation of the unobserved component and conditional standard deviation.

|                | RMSE       | MAE        |
|----------------|------------|------------|
|                | DGP 1      | DGP 2      | DGP 3      | DGP 1      | DGP 2      | DGP 3      |
| Exact Bayes    | 0.0463     | 0.0207     | 0.0203     | 0.0382     | 0.0004     | 0.0155     |
| LSND           | 0.0495     | 0.0253     | 0.0271     | 0.0405     | 0.0006     | 0.0207     |
| QNK            | 0.1211     | 0.2646     | 0.1098     | 0.0990     | 0.0700     | 0.0664     |

Panel A: \(\mu_t\)

Panel B: \(\exp(h_t/2)\)

NOTE: 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 inferential 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 \(E[\mu_t(\theta_0)|y_T]\), while the conditional standard deviation error measures are computed relative to \(E[\exp(h_t/2)(\theta_0)|y_T]\), where \(\theta_0\) denotes the true parameter vector and \(T = 11,000\).

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

The table shows the accuracy in the estimation of the unobserved component and the conditional standard deviation. The columns represent different DGPs, and the rows represent different methods: Exact Bayes, LSND, and QNK. The RMSE and MAE values are given for each method and DGP.

The results highlight the accuracy of the estimations for different methods and DGPs. For instance, the Exact Bayes method consistently produces the most accurate estimates, especially under DGP 1 and 2. The QNK method, on the other hand, shows some inaccuracy under certain conditions, especially under DGP 3.

The figures visually represent the posterior means of the latent states and their conditional standard deviation. The graphs exhibit how the posterior means evolve over time, showing the model's ability to track changes in the latent variables. The plots compare different methods and demonstrate their relative performance.

This analysis underscores the importance of selecting the right method based on the specific characteristics of the data and the model. The QNK method, while generally accurate, may not always match the precision of exact Bayes, especially under conditions where it relies on variational approximations.
Table 2. Predictive performance of competing Bayesian approaches: exact Bayes, LSND and CY and QNK.

| Panel A: DGP 1 | LS | CS-10% | CS-20% | CS-80% | CS-90% | CRPS | TWCPRS | IS |
|---------------|----|--------|--------|--------|--------|------|--------|----|
| True DGP      | −1.259 | −0.308 | −0.508 | −0.505 | −0.297 | −0.481 | −0.414  | −0.011 |
| Exact Bayes   | −1.260 | −0.308 | −0.508 | −0.506 | −0.297 | −0.481 | −0.414  | −0.011 |
| LSND          | −1.261 | −0.308 | −0.509 | −0.507 | −0.298 | −0.481 | −0.414  | −0.011 |
| CY            | −1.262 | −0.309 | −0.509 | −0.507 | −0.298 | −0.482 | −0.417  | −0.013 |
| QNK           | −1.262 | −0.309 | −0.509 | −0.507 | −0.298 | −0.482 | −0.417  | −0.013 |

| Panel B: DGP 2 | LS | CS-10% | CS-20% | CS-80% | CS-90% | CRPS | TWCPRS | IS |
|---------------|----|--------|--------|--------|--------|------|--------|----|
| True DGP      | −1.192 | −0.341 | −0.551 | −0.551 | −0.340 | −0.454 | −0.313  | −0.097 |
| Exact Bayes   | −1.193 | −0.342 | −0.551 | −0.551 | −0.340 | −0.454 | −0.313  | −0.097 |
| LSND          | −1.194 | −0.342 | −0.551 | −0.551 | −0.341 | −0.454 | −0.314  | −0.101 |
| CY            | −1.205 | −0.346 | −0.557 | −0.555 | −0.344 | −0.456 | −0.410  | −0.182 |
| QNK           | −1.212 | −0.350 | −0.560 | −0.560 | −0.349 | −0.456 | −0.410  | −0.182 |

| Panel C: DGP 3 | LS | CS-10% | CS-20% | CS-80% | CS-90% | CRPS | TWCPRS | IS |
|---------------|----|--------|--------|--------|--------|------|--------|----|
| True DGP      | −1.268 | −0.304 | −0.505 | −0.521 | −0.300 | −0.490 | −0.150  | −0.424 |
| Exact Bayes   | −1.268 | −0.305 | −0.506 | −0.520 | −0.299 | −0.491 | −0.150  | −0.423 |
| LSND          | −1.271 | −0.306 | −0.507 | −0.521 | −0.301 | −0.491 | −0.150  | −0.442 |
| CY            | −1.301 | −0.315 | −0.521 | −0.536 | −0.311 | −0.497 | −0.152  | −0.707 |
| QNK           | −1.301 | −0.315 | −0.521 | −0.536 | −0.311 | −0.497 | −0.152  | −0.707 |

NOTE: The column labels indicate the out-of-sample predictive performance measure while the row labels indicate the predictive method. “True DGP” indicates the predictive results that condition on the true parameters. Panels A, B, and C correspond to the results for DGP 1, 2, and 3, respectively. The average predictive measures in this table were computed using 10,000 out-of-sample evaluations.

5.3. Predictive Accuracy

To assess the predictive accuracy of each method we conduct an expanding window prediction exercise using the same generated data as in the previous section. The exercise consists of constructing the Bayesian predictive density for $Y_{n+1}$, conditional on the sample $y_{1:n}$, for each of the competing approaches and for $n \in \{1000, \ldots , T−1\}$. For each method and each out-of-sample time point we evaluate eight measures of predictive accuracy: the logarithmic score, four censored scores, the continuously ranked probability score, the tail weighted continuously ranked probability score and the interval score. Details of all scoring rules, including appropriate references, are provided in Section B.3 of the supplementary materials. We document results using 100, 1000 and 10,000 out-of-sample evaluations, respectively, remembering that the Chan and Yu (2022) method (the CY method hereafter) method is now included in the comparison, but only for the case of DGP 2. For reasons of space, we only present results for the largest number of out-of-sample evaluations (10000) in the main text, in Table 2, while the results for the other evaluation periods are given in Section B.2 of the supplementary materials.

Focusing first on the results in Table 2, based on the very large number of out-of-sample evaluations, we observe an interesting ranking. Across all designs, and according to all measures of accuracy, exact Bayes is the most accurate method. As accords with the inferential results discussed above, the LSND method has a predictive accuracy that often matches, or is extremely similar to, that of exact Bayes, followed, in order, by CY and QNK. A similar ranking holds for the results recorded in Tables B.2 and B.3 in Section B.2 of the supplementary materials. However, the differences between methods are somewhat less stark over the smaller out-of-sample evaluation periods, which highlights the fact that it is ultimately the consistency properties of the different VB methods (in evidence for the largest evaluation period, given the large size of the expanding estimation windows) that is driving the discrepancies between the predictive accuracy of the competing methods.

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

However, despite there being certain DGP and scoring rule combinations where the methods perform similarly, this is not true across all DGPs and loss measures, in particular for the larger out-of-sample evaluation period. For example, and with specific reference to Table 2, there is a clear trend that as model complexity increases (i.e., moving from DGP 1 through to DGP 3), variational methods that work harder to correctly approximate the states have greater predictive accuracy. This finding is particularly marked for the log score and the interval score, which directly measure the dispersion of the posterior predictive. In the case of DGP 3, the all-purpose variational method of Quiroz, Nott, and Kohn (2022) performs the worst across all the methods under analysis, and most notably for the log score and the interval score. This feature is most likely due to the fact that the posteriors associated with the method of Quiroz, Nott, and Kohn (2022) have overly thin tails. Consequently, parameter uncertainty is not adequately accounted for when constructing the predictive, which results in a predictive with thin tails, and ultimately translates into poor performance in scores that measure both location and/or dispersion.
6. Discussion

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

In the context of empirically relevant SSMs, we present convincing numerical evidence of a clear hierarchy in terms of the accuracy of state inference across different variational methods: methods that can close Jensen's gap produce qualitatively more accurate inferences than those that do not. However, in terms of predictive accuracy, our findings are more complex: the predictive accuracy of different VB methods depends on the interplay between the underlying DGP and the sample sizes encountered in the analysis. In large out-of-sample settings, there is a one-to-one relationship between methods that close Jensen's gap, that is, methods that produce accurate inferences, and those that produce accurate predictions. In small samples however, our numerical results suggest that less sophisticated VB methods, which do not close Jensen's gap and produce inaccurate inferences as a consequence, can still produce reasonably accurate VB-based predictions. Indeed, we document that there are certain circumstances, that is, sample size, DGP and loss combinations, where there is little to separate the various approaches.

To keep the length of this article 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, for example, classes based on normalizing or autoregressive flows, may be flexible enough to deliver accurate inferences and predictions; we refer to, for example, 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.

Supplementary Materials

This supplementary appendix contains discussion on the variational methods compared in the paper, additional numerical results, and proofs of all technical results presented in the main text.

Acknowledgments

The authors thank two anonymous referees and the Associate Editor for very helpful comments.

Funding

David Frazier and Gael Martin were also supported by the Australian Research Council’s Discovery project grant DP200101414.

References

Aït-Sahalia, Y., and Kimmel, R. (2007), “Maximum Likelihood Estimation of Stochastic Volatility Models,” Journal of Financial Economics, 83, 413–452. [793]

Aït-Sahalia, Y., Li, C., and Li, C. X. (2021), “Closed-Form Implied Volatility Surfaces for Stochastic Volatility Models with Jumps,” Journal of Econometrics, 222, 364–392. [793]

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, 328–352. [793]

Andrews, C., Doucet, A., and Holenstein, R. (2011), “Particle Markov chain Monte Carlo,” (with discussion), Journal of the Royal Statistical Society, Series B, 72, 269–342. [793]

Bardenet, R., Doucet, A., and Holmes, C. (2017), “On Markov Chain Monte Carlo Methods for Tall Data,” J. Machine Learning Res. 18, 1515–1557. [793]

Bates, D. S. (2006), “Maximum Likelihood Estimation of Latent Affine Processes,” The Review of Financial Studies, 19, 909–965. [793]

Betancourt, M. (2018), “A Conceptual Introduction to Hamiltonian Monte Carlo,” https://arxiv.org/abs/1701.02434v2. [793,794]

Bissiri, P.G., Holmes, C. C., and Walker, S. G. (2016), “A General Framework for Updating Belief Distributions,” Journal of the Royal Statistical Society, Series B, 78, 1103–1130. [796]

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), “Variational Inference: A Review for Statisticians,” Journal of the American statistical Association, 112, 859–877. [793]

Braun, M., and McAuliffe, J. (2010), “Variational Inference for Large-Scale Models of Discrete Choice,” Journal of the American Statistical Association, 105, 324–335. [793]

Carter, C. K., and Kohn, R. (1994), “On Gibbs Sampling for State Space Models,” Biometrika, 81, 541–553. [798]

Chan, J. C., and Yu, X. (2022), “Fast and Accurate Variational Inference for Large Bayesian VARs with Stochastic Volatility,” Journal of Economic Dynamics and Control, 143, 104505. [794,799,800,802]

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. [793]

Danielsson, J., and Richard, J.-F. (1993), “Accelerated Gaussian Importance Sampler with Application to Dynamic Latent Variable Models,” Journal of Applied Econometrics, 8, S153–S173. [793]

Dean, T. A., Singh, S. S., Jasra, A., and Peters, G. W. (2014), “Parameter Estimation for Hidden Markov Models with Intractable Likelihoods,” Scandinavian Journal of Statistics, 41, 970–987. [793]

Douc, R., Moulines, E., Olsson, J., and Van Handel, R. (2011), “Consistency of the Maximum Likelihood Estimator for General Hidden Markov Models,” The Annals of Statistics, 39, 474–513. [796,797]

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, 295–313. [793]

Durbin, J., and Koopman, S. J. (2001), Time Series Analysis by State Space Methods. Oxford: Oxford University Press. [793]

Earle, P. (2011), “Bayesian Inference for Time Series State Spaces,” in Handbook of Markov Chain Monte Carlo, eds. S. Brooks, A. Gelman, G. Jones, and X. Meng. p. 513–530, Boca Raton, FL: Taylor & Francis. [793]

Flury, T., and Shephard, N. (2011), “Bayesian Inference based only on a Simulated Likelihood,” Econometric Theory, 27, 933–956. [793]

Frazier, D. T., Loaiza-Maya, R., Martin, G. M., and Koo, B. (2021), “Loss-based Variational Bayes Prediction,” arXiv preprint arXiv:2104.14054. [799]

Frazier, D. T., Manes onethorn, W., Martin, G. M., and McCabe, B. P. (2019), “Approximate Bayesian Forecasting,” International Journal of Forecasting, 35, 521–539. [799,800]

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018), “Asymptotic Properties of Approximate Bayesian Computation,” Biometrika, 105, 593–607. [793]
Frühwirth-Schnatter, S. (1994), “Data Augmentation and Dynamic Linear Models,” *Journal of Time Series Analysis*, 15, 183–202. [798]

Gallant, A. R., and Tauchen, G. (1996), “Which Moments to Match?” *Econometric Theory*, 12, 657–681. [793]

Giordani, P., Pitt, M., and Kohn, R. (2011), “Bayesian Inference for Time Series State Space Models,” in *The Oxford Handbook of Bayesian Econometrics*, eds. J. Geweke, G. Koop, and H. van Dijk, pp. 61–124, Oxford: Oxford University Press. [793]

Gunawan, D., Kohn, R., and Nott, D. (2021), “Variational Bayes Approximation of Factor Stochastic Volatility Models,” *International Journal of Forecasting*, 37, 1355–1375. [799]

Harvey, A., Koopman, S., and Shephard, N. (2004), *State Space and Unobserved Component Models: Theory and Applications*. Cambridge: Cambridge University Press. [793]

Huber, F., Koop, G., and Onorante, L. (2020), “Inducing Sparsity and Shrinkage in Time-Varying Parameter Models,” *Journal of Business & Economic Statistics*, 39, 669–683. [798]

Jacquier, E., Polson, N. G., and Rossi, P. E. (2002), “Bayesian Analysis of Stochastic Volatility Models,” *Journal of Business & Economic Statistics*, 20, 69–87. [798]

Johndrow, J. E., Smith, A., Pillai, N., and Dunson, D. B. (2019), “MCMC for Imbalanced Categorical Data,” *Journal of the American Statistical Association*, 114, 1394–1403. [793]

Kim, S., Shephard, N., and Chib, S. (1998), “Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models,” *The Review of Economic Studies*, 65, 361–393. [798]

Koop, G., and Korobilis, D. (2020), “Bayesian Dynamic Variable Selection in High Dimensions,” available at SSRN 3246472. [794,799]

Lancaster, T. (2000), “The Incidental Parameter Problem Since 1948,” *Journal of Econometrics*, 95, 391–413. [794,798]

Loaiza-Maya, R., Smith, M. S., Nott, D. J., and Danaher, P. J. (2022), “Fast and Accurate Variational Inference for Models with Many Latent Variables,” *Journal of Econometrics*, 230, 339–362. [794,798,800,801]

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, 508–522. [793]

Neyman, J., and Scott, E. L. (1948), “Consistent Estimates based on Partially Consistent Observations,” *Econometrica*, 16, 1–32. [794]