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 broad 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 Jordani *et al.*, 2011, for extensive reviews).

*Corresponding author: david.frazier@monash.edu.

The key feature of SSMs is their dependence on hidden, or latent, 'local' variables, or states, which govern the dependence of the observed data, in conjunction with a vector of unknown 'global' parameters. This feature leads to inferential challenges with, for example, the likelihood function 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 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 likelihood function 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 assumed data generating process (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 data generating processes; with certain VB methods unable to produce similar out-of-sample accuracy results to exact Bayes.

Throughout the remainder, we make use of the following notational conventions. Generic p, g are used to denote densities, and π is used to denote posteriors conditioned only on data, and where the conditioning is made explicit depending on the situation. For any arbitrary collection of data $(z_1, \ldots, z_n)$, we abbreviate this collection as z_1^n. For a sequence a_n, the terms $O_p(a_n)$, $o_p(a_n)$ and $\rightarrow_p$ have their usual meaning. We let $d(\cdot, \cdot)$ denote a metric on $\Theta \subseteq \mathbb{R}^{d_\theta}$.

2 State Space Models: Exact Inference

A state space model (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 state and measurement equations: for a vector of unknown random parameters θ taking values in the probability space $(\Theta, \mathcal{F}_\theta, P_\theta)$, where P_θ admits the density function p_θ,

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

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

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 ν. The system in (1) and (2) gives rise to the following conditional and transition densities:

$$\begin{aligned} Y_t \; |X_t, \theta &\sim \; g_\theta(y_t|x_t) \\ X_{t+1} \; |X_t, \theta &\sim \; \chi_\theta(x_{t+1}, x_t), \end{aligned}$$

where $\chi_\theta(\cdot, \cdot)$ denotes the [illegible]sition kernel wrt the measure μ. For simplicity, throughout the remainder we disregard terms dependence on the initial measure ν and the invariant measure μ, when no confusion will result.

15

haun McDonald

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

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

The (average) observed data log-likelihood is thus

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

and the maximum likelihood estimator (MLE) of θ is defined as

$$\widehat{\theta}_n^{MLE} := \underset{\theta \in \Theta}{\operatorname{argmax}}\ \ell_n(\theta). \tag{4}$$

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 $p(\theta)$ the prior density, the exact (marginal) posterior for θ, defined as

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

where

16
2021-07-02
haun McDonald

$$\pi(\theta, x_1^n|y_1^n) \propto p(y_1^n|x_1^n, \theta) p(x_1^n|\theta) p(\theta) \propto p_\theta(y_1^n, x_1^n) p(\theta) \tag{6}$$

is available (e.g. via straightforward MCMC methods) only in limited cases, the LGSSM being one such case. In more complex settings and/or settings where either θ or $\{(X_t, Y_t)\}$, or both, are high-dimensional, accessing (5) can be difficult, with standard MCMC methods leading to slow mixing, and potentially unreliable inferences (Betancourt, 2018).

17
haun McDonald

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 (5), depending on the mode of inference being adopted. The focus of this paper, as already highlighted, is on variational *Bayes* and, in particular, on the accuracy of such methods in SSMs. However, as part of the following section we also demonstrate the asymptotic behaviour of frequentist variational point estimators of θ, as this result will ultimately help us interpret the behavior of the variational posterior in SSMs.

18-19
2 notes:

3 State Space Models: Variational Inference

3.1 Overview

The idea of VB is to produce an approximation to the joint posterior in (6) by searching in a given family of distributions for the member that minimizes a user-chosen divergence measure between the posterior of interest and the family. This replaces the posterior sampling problem with one of optimization over the family of densities used to implement the approximation. We now review the use of variational methods in SSMs, paying particular attention to the Markovian nature of the states.

VB approximates the posterior $\pi(x_1^n, \theta|y_1^n)$ by minimizing the KL divergence between a family of densities $\mathcal{Q}$, with generic element $q(\theta, x_1^n)$, and π:

$$\mathrm{KL}(q||p) = \int q(x_1^n, \theta) \log \frac{q(x_1^n, \theta)}{\pi(x_1^n, \theta|y_1^n)} \mathrm{d}x_1^n \mathrm{d}\theta. \tag{7}$$

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

$$\mathrm{ELBO}(q||p) := \int q(\theta, x_1^n) \log \frac{p(y_1^n|x_1^n, \theta)p(x_1^n|\theta)p(\theta)}{q(\theta, x_1^n)} \mathrm{d}x_1^n \mathrm{d}\theta. \tag{8}$$

Hence, we may define the variational approximation as follows: for a given class $\mathcal{Q}$,

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

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

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

with the class sometimes further restricted to be a mean-field approximation

$$\mathcal{Q} = \left\{ q : q(\theta, x_1^n) = \prod_{i=1}^{d_\theta} q_\theta(\theta_i) \prod_{i=1}^{n} q_x(x_i) \right\},$$

or where we make use of the Markovianity of the states via the class

$$\mathcal{Q} = \left\{ q : q(\theta, x_1^n) = \prod_{i=1}^{d_\theta} q_\theta(\theta_i) q_x(x_1) \prod_{i=2}^{n} q_x(x_i|x_{i-1}) \right\}.$$

Regardless of which precise form of variational family is adopted, the KL divergence (and hence the ELBO) involves both parameters θ and states x_1^n. Expressing the KL divergence in (7) as

$$\mathrm{KL}(q||\pi) = -\int q(\theta, x_1^n) \log \frac{p(y_1^n|x_1^n, \theta)p(x_1^n|\theta)p(\theta)}{q(\theta, x_1^n)} \mathrm{d}x_1^n \mathrm{d}\theta + \log p(y_1^n),$$

we define:

$$\mathrm{KL}_c(q||\pi) := -\int q(\theta, x_1^n) \log \frac{p(y_1^n|x_1^n, \theta)p(x_1^n|\theta)p(\theta)}{q(\theta, x_1^n)} \mathrm{d}x_1^n \mathrm{d}\theta \tag{9}$$

as a quantity that is equivalent to $\mathrm{KL}(q||\pi)$ up to $\log p(y_1^n)$. The product form of $\mathcal{Q}$ then allows us to

20
2021-07-02
haun McDonald

write:

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

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

$$\mathcal{L}_n(\theta) := \int_{\mathcal{X}} q_x(x_1^n|\theta) \log \frac{p(y_1^n|x_1^n,\theta)p(x_1^n|\theta)}{q_x(x_1^n|\theta)} \mathrm{d}x_1^n, \tag{10}$$

21 aun McDonald

ensen's inequality

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

22-23 2 notes:

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

$$\Upsilon_n(q) := \int_\Theta \{\log p_\theta(y_1^n) - \mathcal{L}_n(\theta)\} q_\theta(\theta) \mathrm{d}\theta, \tag{11}$$

$\mathrm{KL}_c(q||\pi)$ in (9) can then be expressed as

$$\mathrm{KL}_c(q||\pi) = -\int_\Theta \log p_\theta(y_1^n) q_\theta(\theta) \mathrm{d}\theta + \Upsilon_n(q) + \mathrm{KL}(q_\theta(\theta)||p(\theta)). \tag{12}$$

This representation decomposes $\mathrm{KL}_c(q||\pi)$ into three components, two of which only depend on the variational approximation of the global parameters θ, and a third component, $\Upsilon_n(q)$, that Yang *et al.* (2020) refer to as the average (wrt $q_\theta(\theta)$) "Jensen's gap", which encapsulates the error introduced by approxi- ng 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 nately determines the behavior of the variational approximation.

24 aun McDonald

25 aun McDonald

3.2 Consistency of variational point estimators

The decomposition in (12) 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 tates, and consider point estimation of the parameters θ. In this case, we can think of the variational family as $\mathcal{Q} := \{q : q(\theta, x_1^n) = \delta_\theta \times q_x(x_1^n|\theta)\}$, where δ_θ is the Dirac delta function at θ, and we can then

26-27 2 notes:

write the (average) $\mathrm{KL}_c(q||\pi)$ as

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

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

$$(\widehat{\theta}_n, \widehat{q}_x) := \operatorname*{arginf}_{\theta\in\Theta,\mathcal{Q}} \frac{1}{n}\mathrm{KL}_c(\theta \times q_x||\pi).$$

At a minimum, we would hope that the variational estimator $\widehat{\theta}_n$ converges to the same point as the maximum likelihood estimator (MLE) in (4). To compare the behavior of $\widehat{\theta}_n$ and $\widehat{\theta}_n^{MLE}$, we employ the following high-level regularity conditions.

28-29
2 notes:

Assumption 3.1. *(i) The parameter space Θ 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 $H(\theta_0) \geq \sup_{d(\theta,\theta_0)>\delta} H(\theta) + \delta$.*

30-31
2 notes:

32-34
3 notes:

Assumption 3.1 is sufficient to deduce consistency of $\widehat{\theta}_n^{MLE}$ for θ_0. Low level regularity conditions which imply Assumption 3.1 are given in Douc *et al.* (2011). Since the main thrust of this paper is to deduce the accuracy of variational methods in SSMs, and not to focus on the technical details of the SSMs in particular, we make use of high-level conditions to simplify the exposition and reduce necessary technicalities that may otherwise obfuscate the main point.

The following result shows that consistency of $\widehat{\theta}_n$ (for θ_0) is only guaranteed when the variational family for the states is 'good enough'. As such, the limit of the MLE and that of the variational point estimator will not coincide in general.

35-36
2 notes:

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 $\kappa_n = o(1)$, then $\widehat{\theta}_n \to_p \theta_0$.*

The above result demonstrates that for the variational point estimator $\widehat{\theta}_n$ to be consistent, the scaled (feasible) 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 $\kappa_n = o(1)$ is stated at the true value, θ_0, rather than at the estimated value, as it will often be easier to deduce satisfaction of the condition, or otherwise, at convenient points in the parameter space.

37-42
6 notes:

43
haun McDonald

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

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

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

$$Y_t = \alpha X_t + \sigma_0 \eta_t,$$

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

We make use of the autoregressive nature of the state process to approximate the posterior for $\pi(x_1^n|\theta, y_1^n)$ via the variational family

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

family $\mathcal{Q}_x$ does not depend on any variational parameters, and given ρ_0, is the *actual (infeasible) joint distribution* of the states.

Lemma 3.2. *Consider that $\sigma_0 > 0$ and let $0 \leq |\rho_0| < 1$, $0 \leq |\alpha_0| < M$.*

1. *Consider that $\rho_0 = 0$ and known, then the variational estimator $\widehat{\alpha}$ is consistent if and only if $\alpha_0 = 0$.*
2. *Consider that $\alpha_0 = 0$ and known, then the variational estimator for $\widehat{\rho}$ is consistent if and only if $\rho_0 = 0$.*

Remark 3.1. The result of Lemma 3.2 is similar to the results in g and Titterington (2004). However, our result (and proof) is more direct, uses a more realistic class of variational families, and deals with a more complex inference problem.

3.3 Lack of Bayes consistency of the variational posterior

While the above results are based on the variational point estimator for θ_0, a similar result can be stated in terms of so-called 'idealized' variational posterior for θ. To state this result, again consider that we approximate the states only, for any given θ, using the class of variational approximations,

$$q(x_1^n|\theta) := q_\lambda(x_1^n|\theta), \; \lambda \in \Lambda,$$

where $\lambda \in \Lambda$ denotes the vector of so-called 'ational parameters' that characterize the elements in Q. With reference to (10), and now making explicit the dependence of $q_\lambda(x_1^n|\theta)$ on the variational parameter λ, we re-define the criterion $\mathcal{L}_n(\theta)$ in (10) as

$$\mathcal{L}_n(\theta, \lambda) = \int_{\mathcal{X}} q_\lambda(x_1^n|\theta) \log \frac{p(y_1^n|x_1^n, \theta) p(x_1^n|\theta)}{q_\lambda(x_1^n|\theta)} \mathrm{d}x_1^n, \tag{13}$$

and its maximized value as

$$\widehat{L}_n(\theta) := \mathcal{L}_n[\theta, \widehat{\lambda}_n(\theta)] \equiv \sup_{\lambda \in \Lambda} \mathcal{L}_n[\theta, \lambda]. \tag{14}$$

An 'idealized' variational posterior for θ can then be stated as

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

Note that, unlike with the frequentist optimization problem, the idealized VB posterior incorporates the component of Jensen's gap directly into the definition of that posterior. A sufficient condition for the 'VB ideal' to concentrate onto θ_0 is that θ_0 is the maximum of a well-defined limit counterpart to $\widehat{L}_n(\theta)$.

Assumption 3.2. *The following conditions are satisfied.*

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

Lemma 3.3. *Under Assumptions 3.2, for any* $\epsilon > 0$ *and* $A_\epsilon := \{\theta \in \Theta : d(\theta, \theta_\star) > \epsilon\}$,

$$\widehat{Q}(\theta \in A_\epsilon | y_1^n) = o_p(1).$$

Assumptions 3.2(2.b) implies that for any $\theta \in \Theta$, $\mathcal{L}_n[\theta, \lambda]/n$ converges to $\mathcal{L}[\theta, \lambda]$; while part (2.a) is an identification condition and states that $\mathcal{L}[\theta, \lambda]$ is maximized at some $\theta_\star$, which may differ from θ_0. This identification makes clear that if $\theta_\star \neq \theta_0$, then

$$\mathcal{L}[\theta_\star, \lambda(\theta_\star)] > \mathcal{L}[\theta_0, \lambda(\theta_0)]$$

and the idealized posterior for θ will not concentrate onto θ_0. This can be interpreted explicitly in terms of Jensen's gap as defined in (11) by recalling that under Assumption 3.1, $H(\theta_0) = \operatorname{plim}_n \ell_n(\theta_0)$, and by considering the limit of (the scaled) Jensen's gap evaluated at θ_0:

$$\begin{aligned} \operatorname*{plim}_{n\to\infty} \frac{1}{n} \Upsilon_n\left(\theta_0, q_{\widehat{\lambda}_n(\theta_0)}\right) &= \operatorname*{plim}_{n\to\infty} \frac{1}{n} \left\{\log p_{\theta_0}(y_1^n) - \mathcal{L}_n\left[\theta_0, \widehat{\lambda}_n(\theta_0)\right]\right\} \\ &= H(\theta_0) - \mathcal{L}[\theta_0, \lambda(\theta_0)] \\ &\geq H(\theta_0) - \mathcal{L}[\theta_\star, \lambda(\theta_\star)] + \delta, \end{aligned}$$

for some $\delta \geq 0$. If Assumption 3.2(2.a) is satisfied at $\theta_\star \neq \theta_0$, then $\delta > 0$, and $\kappa_n := \Upsilon_n(\theta_0, q_{\widehat{\lambda}_n(\theta_0)})/n = O_p(1)$ and not $o_p(1)$.

Taken together, Lemmas 3.1 and 3.3 show that, regardless of whether one conducts frequentist or Bayesian variational inference in SSMs, consistent inference for θ_0 will require that a version of Jensen's gap converges to zero. Moreover, as Example 3.1 has demonstrated, this is not likely to occur even in simple SSMs.

4 Existing Variational Approaches: Implications for Inference

As the previous discussion illustrates, the need to approximate the posterior of x_1^n 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_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. 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, θ_0. It is then clear that VB methods must somehow solve the *incidental parameter problem* if they wish to obtain reliable inference for θ_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 θ_0. We begin with reference to a variational approach to point estimation of θ_0, followed by brief outlines of two classes of approach that target the posterior for θ 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 Rotnitzky et al. (2019) splits the optimization procedure into two parts: one for x_1^n conditional on θ; and one for θ conditional on x_1^n. 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, $q_x(x_1^n)$, which depend on parameters $\lambda \in \Lambda$, and proposes to estimate θ by solving for

$$\widehat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}} \, \mathcal{L}_n[\theta, \widehat{\lambda}_n(\theta)],$$

where $\mathcal{L}_n[\theta, \widehat{\lambda}_n(\theta)]$ is defined in (14). As detailed in Section 3, the optimizer of the limit criterion $\mathcal{L}[\theta, \lambda]$ will not in general coincide with the minimizer of $H(\theta)$, i.e. θ_0. Hence, consistency of the variational estimator $\widehat{\theta}_n$ for θ_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 'integrate out' the latent states so that there is no need to perform joint inference on (θ, x_1^n). Such an approach can be motivated by the fact that the KL divergence in (7) can be rewritten as

$$\mathrm{KL}(q||\pi) = \int_\Theta \int_{\mathcal{X}} q_\theta(\theta) q_x(x_1^n|\theta) \log \frac{q_\theta(\theta) q_x(x_1^n|\theta)}{\pi(x_1^n, \theta|y_1^n)} \mathrm{d}x_1^n \mathrm{d}\theta.$$

From the above representation of $\mathrm{KL}(q||\pi)$ it becomes clear that if $q_x(x_1^n|\theta)$ is of the form

$$q_x(x_1^n|\theta) = \pi(x_1^n|y_1^n, \theta),$$

(i.e. is equivalent to the exact posterior for x_1^n conditional on θ), then we can rewrite $\mathrm{KL}(q||\pi)$ as

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

The above demonstrates that if we are able to use as our variational family for the states the actual (conditional) posterior, then we can transform a variational problem for (θ, x_1^n) into a variational problem for θ alone. This is the approach adopted by Loaiza-Maya *et al.* (2021), and is applicable in any case where draws from $p(x_1^n|y_1^n)$ 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 $\mathrm{ELBO}[q_\theta||\pi(\theta|y_1^n)]$ (equivalent, in turn, to the gradient of the joint ELBO in (8), by the above argument). This, in turn, allows optimization over q_θ to produce an approximation to the posterior $\pi(\theta|y_1^n)$. Indeed, such an approach can be applied in many SSMs, such as unobserved component models like the LGSSM, in which draws from $\pi(x_1^n|y_1^n, \theta)$ can be generated

ctly 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 ropolis 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 abl sample readily from $\pi(x_1^n|y_1^n,\theta)$ it may still be possible to integrate out the states using icle filtering methods. To this end, assume that we can obtain an unbiased estimate of the observed data likelihood $p_\theta(y_1^n)$ using a particle filter, which we denote by $\widehat{p}_\theta(y_1^n)$. We follow *et al.* (2017) and write $\widehat{p}_\theta(y_1^n)$ as $\widehat{p}(y_1^n|\theta,z)$ to make the estimator's dependence o random filtering explicit through the dependence on a random variable z, with z subsequently ed by the condition $\log \widehat{p}(y_1^n|\theta,z) - \log p_\theta(y_1^n)$. For $g(z|\theta)$ denoting the density of $z|\theta$, Tran *et al.* (2017) consider variational inference for the augmented posterior

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

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

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

where in this case

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

For fixed θ, $E_z\left[\widehat{p}(y_1^n|\theta\, z)\right]=p_\theta(y_1^n)$, but in general $\log\widehat{p}(y_1^n|\theta,z)$ is a biased estimator of $\log p_\theta(y_1^n)$, from which it follows that

$$|\Upsilon_n[q(\theta,z)]|\geq 0.$$

However, in contrast to the general approximation of the states discussed in Section 3, which intimately

relies on the choice of the approximating density $q(x_1^n|\theta)$, VBIL can achieve consistent inference on θ_0 by choosing an appropriate number of particles N in the production of $\widehat{p}(y_1^n|\theta, z)$.

To see this, we recall that a maintained assumption in the literature on PMCMC methods is that, for all n and N, the conditional mean and variance of the density $g(z|\theta)$ satisfy

$$\mathbb{E}[z|\theta] = -\gamma(\theta)^2/2N, \text{ and Var}\,[z|\theta] = \gamma(\theta)^2/N,$$

where $\gamma(\theta)^2$ is bounded uniformly over Θ; see, e.g., Assumption 1 in Doucet *et al.* (2015) and Assumption 1 in Tran *et al.* (2017). However, in general, N is assumed to be chosen so that $\mathbb{E}[z|\theta] = -\sigma^2/2$ and $\text{Var}[z|\theta] = \sigma^2 > 0, 0 < \sigma < \infty$. Note that, under this choice for N, for any $\varepsilon > 0$

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

assuming $q_\theta(\theta_0), \sigma^2 < \infty$. From this condition, we see that the VBIL inference problem is asymptotically the same as the VB inference problem for θ alone. Consequently, existing results on the posterior concentration of VB methods for θ alone can be used to deduce posterior concentration of the VBIL posterior for θ.

4.3 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 $\pi(x_1^n|y_1^n, \theta)$ by approximating the relationship between X_t and X_{t-1}, which may in truth be non-linear in θ, by the random walk model $X_t = X_{t-1} + \epsilon_t$, with $\epsilon_t \sim i.i.d.N(0, \sigma_0^2)$, and then use Kalman filtering to update the states in conjunction with a linear approximation to the measurement equation. Using this formulation, the variational approximation is of the form

$$q(x_1^n, \theta) = q_\theta(\theta)q_x(x_1^n), \text{ where } q_x(x_1^n) \propto \prod_{k\geq 1} \exp\left(-\frac{1}{2}\left\{x_k - \widehat{x}_{k|k}\right\}^2 (1 - \mathcal{K}_k)P_{k|k-1}\right),$$

where the terms $\mathcal{K}_k, P_{k|k-1}, \widehat{x}_{k|k}$ are explicitly calculated using the Kalman recursion:

$$\widehat{x}_{k|k} = \widehat{x}_{k|k-1} + \mathcal{K}_k(y_t^\star - \widehat{x}_{k|k-1}),$$

and where $\mathcal{K}_k$ is the Kalman gain, $P_{k|k-1}$ is the predicted variance of the state, and in the application of Koop and Korobilis (2018), $y_t^\star = \log(y_t^2)$.

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

The variational approach of Chan and Yu (2020) can be viewed similarly: the suggested algorithm assumes and exploits a particular dynamic structure for the states that allows for analytical (posterior) updates and thus leads to computationally simple estimates for the variational densities of $q_x(x_1^n)$. As with the VBKF approach, the assumed nature of the state process used by Chan and Yu (2020) to estimate $q_x(x_1^n)$ implies that, in general, it is unlikely that Bayesian consistency can be achieved. 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 θ_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 θ_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 θ is $g_\theta(Y_{n+1}|x_{n+1})$, so that the predictive pdf for Y_{n+1} can be expressed as

$$\begin{aligned} p(Y_{n+1}|y_1^n) &= \int_\Theta \int_\mathcal{X} \cdots \int_\mathcal{X} g_\theta(Y_{n+1}|x_{n+1}) \pi(x_1^{n+1}, \theta|y_1^n) \mathrm{d}x_1^{n+1} \mathrm{d}\theta \\ &= \int_\Theta \int_\mathcal{X} \cdots \int_\mathcal{X} g_\theta(Y_{n+1}|x_{n+1}) p(x_{n+1}|x_1^n, y_1^n, \theta) p(x_1^n|y_1^n, \theta) \pi(\theta|y_1^n) \mathrm{d}x_1^{n+1} \mathrm{d}\theta \\ &= \int_\Theta \int_\mathcal{X} \cdots \int_\mathcal{X} g_\theta(Y_{n+1}|x_{n+1}) \underbrace{p(x_{n+1}|x_n, y_1^n, \theta) p(x_1^n|y_1^n, \theta)}_{(1)} \underbrace{\pi(\theta|y_1^n)}_{(2)} \mathrm{d}x_{n+1} \mathrm{d}x_1^n \mathrm{d}\theta, \quad (15) \end{aligned}$$

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 (15) is infeasible or prohibitive computationally, due to the difficulty of sampling from $\pi(x_1^{n+1}, \theta|y_1^n)$. Instead, VB methods can produce an estimate of $p(Y_{n+1}|y_1^n)$ by approximating, in various ways, the two pieces in equation (15) underlined (1) and (2). All such methods replace the second underlined term by some approximate posterior for θ, but differ in how they access the first underlined term. In all the cases of which we are aware, we can separate VB methods for prediction in SSMs into two classes: a class which makes explicit use of a variational approximation to the states, $\widehat{q}_x$ to replace $p(x_1^n|y_1^n, \theta)$; and a class that uses an accurate simulation-based estimate of $p(x_1^n|y_1^n, \theta)$. We discuss these two strategies in greater detail in the following section.

121-122
2 notes:

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 $\widehat{q}_x$ of $p(x_1^n|y_1^n, \theta)$, in addition to using the structure of the state equation. To illustrate this, it is perhaps easiest to consider the case where we seek to estimate (15) by generating values of Y_{n+1} and using as our estimate of $p(Y_{n+1}|y_1^n)$ a 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:

123-124
2 notes:

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

More precisely, consider a fixed value of $\theta^{(j)}$ drawn from some variational approximation of $\pi(\theta|y_1^n)$, call it $\widehat{q}_\theta$. Given the realization $\theta^{(j)}$, we simulate $x_1^n|y_1^n, \theta^{(j)}$ from the VB approximation of the states $\widehat{q}_x$. Next, given $x_n^{(j)} \sim \widehat{q}_x$, we can generate $x_{n+1}^{(j)}$ from $p(x_{n+1}|x_n^{(j)}, y_1^n, \theta^{(j)})$ by generating from the transition density of the states, $x_{n+1}^{(j)} \sim \chi_\theta(x_{n+1}, x_n^{(j)})$, and under the draws $x_n^{(j)}$ and $\theta^{(j)}$. Lastly, $Y_{n+1}^{(j)}$

is generated according to the conditional distribution $Y_{n+1}^{(j)} \sim g_\theta(y_{n+1}|x_{n+1}^{(j)})$. While the above is simple to implement, the critical point to realize is that since $x_n^{(j)}$ has not been generated from $p(x_1^n|y_1^n, \theta)$, in general $x_{n+1}^{(j)}$ is not a draw from $p(x_{n+1}|x_n, y_1^n, \theta)p(x_1^n|y_1^n, \theta)$. Hence, the draw $Y_{n+1}^{(j)}$ does not correctly reflect the structure of the assumed model, and $Y_{n+1}^{(j)}$ cannot be viewed as being a draw from the exact predictive density in (15).

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 $\widehat{q}_x$ (in addition to the differences in the construction of $\widehat{q}_\theta$) and thus to generate $x_{n+1}^{(j)}$.

125

haun McDonald

5.2.2 Simulation approaches

126

haun McDonald

As an alternative, one may estimate $p(Y_{n+1}|y_1^n)$ using exact draws of Y_{n+1}, x_{n+1} and x_1^n, conditional on the draw of θ from some $\widehat{q}_\theta$. For example, if draws from the exact posterior of the states, $p(x_1^n|y_1^n, \theta)$, are readily available via an efficient MCMC algorithm, $p(Y_{n+1}|y_1^n)$ can be estimated via the same set of steps as delineated above, apart from $x_n^{(j)}$ being drawn directly from $p(x_1^n|y_1^n, \theta)$, rather than some $\widehat{q}_x$; see, for example, Loaiza-Maya *et al.* (2021). In this case, $x_{n+1}^{(j)}$ is a draw from $p(x_{n+1}|x_n, y_1^n, \theta)p(x_1^n|y_1^n, \theta)$ and, consequently, the draw Y_{n+1} 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.

127-131

5 notes:

132

haun McDonald

6 Numerical Assessment of VB methods

6.1 Simulation design

133-134

2 notes:

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 these 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):

135

haun McDonald

$$\mu_t = \bar{\mu} + \rho_\mu (\mu_{t-1} - \bar{\mu}) + \sigma_\mu \varepsilon_t \tag{16}$$

$$h_t = \bar{h} + \rho_h \left(h_{t-1} - \bar{h}\right) + \sigma_h \eta_t \tag{17}$$

$$Y_t = \mu_t + \exp(h_t/2)u_t, \tag{18}$$

where $(\varepsilon_t, \eta_t, u_t)' \overset{i.i.d.}{\sim} N(0, I_3)$. The unobserved component term μ_t is a latent variable that captures the persistence in the conditional mean of Y_t, while the stochastic volatility term h_t captures the persistence

136

haun McDonald

in the conditional variance. We consider the following three set of values for the true parameters:

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

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

For the predictive assessment we compare exact Bayes with the two variational methods above plus the method of Chan and Yu (2020). As noted earlier, the 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 $\rho_{h0} = 1.0$ 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 cause 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

6.2.1 Exact Bayes

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

$$\pi(\theta, \mu_1^n, h_1^n | y_1^n) = \frac{p(y_1^n | \mu_1^n, h_1^n, \theta) p(\mu_1^n, h_1^n | \theta) p(\theta)}{p(y_1^n)}, \tag{19}$$

with prior $p(\theta) = p(\bar{\mu})p(\rho_\mu)p(\sigma_\mu)p(\bar{h})p(\rho_h)p(\sigma_h)$, where $\bar{\mu} \sim N(0, 1000)$, $\rho_\mu \sim \text{U}(0,1)$, ${\sigma_\mu}^2 \sim \text{IG}(1.001, 1.001)$, $\bar{h} \sim N(0, 1000)$, $\rho_h \sim \text{U}(0,1)$ and ${\sigma_h}^2 \sim \text{IG}(1.001, 1.001)$. We draw from (19) using an MCMC algorithm. Specifically, the vector h_1^n is generated using the method proposed in Primiceri (2005), while μ_1^n is generated using the forward-filtering backward-sampling method in Carter and Kohn (1994). Given the choice of priors, the parameters $\bar{\mu}$, σ_μ, $\bar{h}$ and σ_h can be generated directly using Gibbs steps. The parameters ρ_μ and ρ_h are generated using a Metropolis-Hastings step with a Gaussian

proposal distribution. The corresponding predictive (expressed using obvious notation):

$$p(Y_{n+1}|y_1^n) = \int_\Theta \int_\mathcal{H} \int_\mathcal{M} g_\theta(Y_{n+1}|\mu_{n+1}, h_{n+1}) p(\mu_{n+1}, h_{n+1}|\theta, \mu_1^n, h_1^n) \pi(\theta, \mu_1^n, h_1^n|y_1^n) d\mu_1^n dh_1^n d\theta, \quad (20)$$

141 haun McDonald

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

6.2.2 Quiroz et al. (2018)

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

$$q_{\widehat{\lambda}}(\theta, \mu_1^n, h_1^n) = q_{\widehat{\lambda}_1}(\theta) q_{\widehat{\lambda}_2}(x_1^n), \quad (21)$$

142-145 4 notes:

where $\widehat{\lambda} = \left(\widehat{\lambda}_1', \widehat{\lambda}_2'\right)'$, x_t $(\mu_t, h_t)'$ and $x_1^n = \left(x_1', \ldots, x_n'\right)'$. The approximations $q_{\widehat{\lambda}_1}(\theta)$ and $q_{\widehat{\lambda}_2}(x_1^n)$ optimal elements in the variational classes $\mathcal{Q}_1 = \{q_{\lambda_1}(\theta) : \lambda_1 \in \Lambda_1\}$ and $\mathcal{Q}_2 = \{q_{\lambda_2}(\lambda_2 \in \Lambda_2\}$, respectively, where the optimization is performed using a stochastic gradient ascent (SGA) algorithm (Bottou, 2010), and the approximation is based on the same prior as specified above. The elements of the first class are Gaussian densities of the form $q_{\lambda_1}(\theta) = \phi_6\left(\theta; \nu_\theta, BB' + \text{diag}(d^2)\right)$, while the elements of the second class are of the form $q_{\lambda_2}(x_1^n) = \phi_{2n}\left(x_1^n; \nu_x, CC'\right)$, where C is a three diagonal lower triangular matrix, and the subscript on the symbol for the normal pdf, ϕ, denotes the dimension of the density. (For more details on this approximating class see Ong *et al.*, 2018.) Replacing $\pi(\theta, \mu_1^n, h_1^n|y_1^n)$ in (20) by the approximation in (21), the predictive density is then estimated as described in Section 5.2.1.

6.2.3 Loaiza-Maya et al. (2021)

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

$$q_{\widehat{\lambda}}(\theta, \mu_1^n, h_1^n|y_1^n) = q_{\widehat{\lambda}}(\theta) p\left(\mu_1^n, h_1^n|y_1^n, \theta\right), \quad (22)$$

where $q_{\widehat{\lambda}}(\theta)$ is an optimal element in the variational class $\mathcal{Q} = \{q_\lambda(\theta) : \lambda \in \Lambda\}$, once again found via SGA. For $\mathcal{Q}$ the class of multivariate Gaussian densities with a factor structure is employed, so that $q_\lambda(\theta) = \phi_6\left(\theta; \nu, BB' + \text{diag}(d^2)\right)$, and $\lambda = \left(\nu', \text{vec}(B)', d'\right)'$. Replacing $\pi(\theta, \mu_1^n, h_1^n|y_1^n)$ in (20) by the approximation in (22) (once again, with the same underlying prior adopted), the predictive density is then estimated as described in Section 5.2.2. Generation from $p\left(\mu_1^n, h_1^n|y_1^n, \theta\right)$ is achieved via an MCMC algorithm that sequentially draws from $p\left(\mu_1^n|y_1^n, \theta, h_1^n\right)$ using the method in Carter and Kohn (1994); and then draws from $p\left(h_1^n|y_1^n, \theta, \mu_1^n\right)$ using the approach in Primiceri (2005).

146 haun McDonald

6.2.4 Chan and Yu (2020)

The final VB method we consider is that of Chan and Yu (2020) (CY hereafter). This approach has been designed specifically for (vector) autoregressive models with stochastic volatility (SV) and not for the UCSV model in (16)-(18). 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:

$$h_t = h_{t-1} + \sigma_h \eta_t \tag{23}$$
$$Y_t = \exp(h_t/2) u_t. \tag{24}$$

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

$$q_{\widehat{\lambda}}(\theta, h_1^n) = q_{\widehat{\lambda}_1}(\sigma_h^2) q_{\widehat{\lambda}_2}(h_0) q_{\widehat{\lambda}_3}(h_1^n), \tag{25}$$

147-148
2 notes:

where $q_{\widehat{\lambda}_1}(\sigma_h^2)$, $q_{\widehat{\lambda}_2}(h_0)$ and $q_{\widehat{\lambda}_3}(h_1^n)$ are optimal elements in the variational classes $\mathcal{Q}_1 = \{q_{\lambda_1}(\sigma_h^2) : \lambda_1 \in \Lambda_1\}$, $\mathcal{Q}_2 = \{q_{\lambda_2}(h_0) : \lambda_2 \in \Lambda_2\}$ and $\mathcal{Q}_3 = \{q_{\lambda_3}(h_1^n) : \lambda_3 \in \Lambda_2\}$, respectively. The elements of each class are defined respectively as $q_{\lambda_1}(\sigma_h^2) = \mathcal{IG}(\sigma_h^2; \nu, S)$, $q_{\lambda_2}(h_0) = \phi_1(h_0; \mu_0, s_0^2)$ and $q_{\lambda_3}(h_1^n) = \phi_n(h_1^n; m, \hat{K}^{-1})$. The variational parameters $\lambda_1 = (\nu, S)'$, $\lambda_2 = (\mu_0, s_0^2)'$ and $\lambda_3 = m$, are calibrated to produce the elements in $\mathcal{Q}_1$, $\mathcal{Q}_2$ and $\mathcal{Q}_3$ that minimise the KL divergence from $p(\theta, h_1^n | y_1^n)$. The authors use a coordinate ascent algorithm (Blei *et al.*, 2017) to perform the optimization, while the value of $\hat{K}^{-1}$ can be optimally computed as a deterministic function of λ_1, λ_2, λ_3 and y_1^n. In our implementation of the CY method, the priors are set to $p(\sigma_h^2) = \mathcal{IG}(\sigma_h^2; 1.001, 1.001)$ and $p(h_0) = \phi_1(h_0; 0, 1000)$, where $\mathcal{IG}$ denotes the inverse gamma distribution. The predictive density is then estimated as described in Section 5.2.1, with h_1^n playing the role of x_1^n therein.

149

haun McDonald

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 $t = 1, 2, ..., T$, for the unobserved component, μ_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 $\mathbb{E}(\mu_t | \theta_0, y_1^n)$ and $\mathbb{E}[\exp(1/2h_t) | \theta_0, y_1^n]$, $t = 1, 2, ..., T$. Table 1 presents the results.

Table 1: Accuracy in the estimation of the unobserved component and conditional standard deviation. Panel A presents the root mean squared error (RMSE) and mean absolute error (MAE) of the posterior mean estimates of the unobserved component (μ_t). The columns correspond to the three DGP specifications, while the rows correspond to the three predictive methods: exact Bayes, LSND and QNK. Panel B presents the corresponding results for the posterior mean estimates of the conditional standard deviation. The unobserved component error measures are computed relative to $E(\mu_1^n|\theta_0, y_1^n)$, while the conditional standard deviation error measures are computed relative to $E(\exp(h_t/2)|\theta_0, y_1^n)$, where θ_0 denotes the true parameter vector.

Panel A: Unobserved component (μ_t)

	RMSE				MAE		
	DGP 1	DGP 2	DGP 3		DGP 1	DGP 2	DGP 3
Exact Bayes	0.0463	0.0865	0.0203	Exact Bayes	0.0382	0.0075	0.0155
LSND	0.0495	0.0817	0.0271	LSND	0.0405	0.0067	0.0207
QNK	0.1211	0.0353	0.1098	QNK	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	Exact Bayes	0.0470	0.0492	0.0180
LSND	0.0520	0.0519	0.0315	LSND	0.0490	0.0468	0.0246
QNK	0.0984	0.1336	0.2231	QNK	0.0983	0.0925	0.1656

As expected, exact Bayes produces the most accurate point estimates for the two sets of latent variables (both μ_t and $\exp(h_t/2)$), as tallies with the theoretical consistency of this method, which ensures that the posterior concentrates onto θ_0. That is, Bayesian consistency (for θ_0) of the exact posterior (given the required regularity) implies that the average of the RMSE and MAE between the exact posterior mean of μ_t and $\mathbb{E}(\mu_t|\theta_0, y_1^n)$ 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 $\pi(\theta|y_1^n)$ 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 time 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 conducted 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_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 Appendix C. 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 Supplementary Appendix D.

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 Supplementary Appendix D a similar ranking is seen to accord between the methods across 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 alternative 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 (relative to exact Bayes) by a substantial amount.

However, despite there being certain DGP and loss 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

Table 2: Predictive performance fof 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.

Panel A: DGP 1

	LS	CS-10%	CS-20%	CS-80%	CS-90%	CRPS	TWCRPS	IS
True DGP	-1.259	-0.308	-0.508	-0.505	-0.297	-0.481	-0.146	-4.001
Exact Bayes	-1.260	-0.308	-0.508	-0.506	-0.297	-0.481	-0.147	-4.012
LSND	-1.261	-0.308	-0.509	-0.507	-0.298	-0.481	-0.147	-4.015
CY	-	-	-	-	-	-	-	-
QNK	-1.262	-0.309	-0.509	-0.507	-0.298	-0.482	-0.147	-4.030

Panel B: DGP 2

	LS	CS-10%	CS-20%	CS-80%	CS-90%	CRPS	TWCRPS	IS
True DGP	-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

	LS	CS-10%	CS-20%	CS-80%	CS-90%	CRPS	TWCRPS	IS
True DGP	-1.268	-0.304	-0.505	-0.521	-0.300	-0.490	-0.150	-4.424
Exact Bayes	-1.268	-0.305	-0.506	-0.520	-0.299	-0.491	-0.150	-4.423
LSND	-1.271	-0.306	-0.507	-0.521	-0.301	-0.491	-0.150	-4.442
CY	-	-	-	-	-	-	-	-
QNK	-1.301	-0.315	-0.521	-0.536	-0.311	-0.497	-0.152	-4.707

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 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 variation 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, when VB methods cannot close Jensen's gap, the resulting inference is markedly 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 discrepancy between different types of variational approaches depends on the DGP, and 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: structured approximations, such as Chan and Yu (2020), perform only slightly worse that more complicated methods, such as that in Loaiza-Maya *et al.* (2021). Whilst so-called 'all-purpose' approximations, such as those in Quiroz *et al.* (2018) generally performing the worst, in terms of both inference and prediction, it is worth pointing out that they still produce 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.

References

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

Ait-Sahalia, Y., Li, C., and Li, C. (2020). Maximum likelihood estimation of stochastic volatility models. Forthcoming, Journal of Econometrics. 2

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. 2

Andrieu, C., Doucet, A., and Holenstein, R. (2011). Particle Markov chain Monte Carlo. jrssb, 72(2):269–342. With discussion. 2

Bates, D. S. (2006). Maximum Likelihood Estimation of Latent Affine Processes. The Review of Financial Studies, 19(3):909–965. 2

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. https://arxiv.org/abs/1701.02434v2. 4

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. 2, 19

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010, pages 177–186. Springer. 18

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. Biometrika, 81(3):541–553. 12, 17, 18

Chan, J. C. and Yu, X. (2020). Fast and accurate variational inference for large Bayesian vars with stochastic volatility. CAMA Working Paper. 2, 14, 16, 17, 19, 23

Chernozhukov, V. and Hong, H. (2003). An mcmc approach to classical estimation. Journal of Econometrics, 115(2):293–346. 28

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. 2

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. 2

Dean, T., Singh, S., Jasra, A., and Peters, G. (2014). Parameter inference for hidden Markov models with intractable likelihoods. Scand. J. Statist. (to appear). 2

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. 35

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. 7

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. 13

Durbin, J. and Koopman, S. J. (2001). Time Series Analysis by State Space Methods. OUP. 1

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. 2

Flury, T. and Shephard, N. (2011). Bayesian inference based only on a simulated likelihood. Econometric Theory, 27:933–956. 2

Frazier, D. T., Loaiza-Maya, R., Martin, G. M., and Koo, B. (2021). Loss-based variational Bayes prediction. arXiv preprint arXiv:2104.14054. 14

Frazier, D. T., Maneesoonthorn, W., Martin, G. M., and McCabe, B. P. (2019). Approximate Bayesian forecasting. International Journal of Forecasting, 35(2):521–539. 14

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. Biometrika, 105(3):593–607. 2

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. Journal of time series analysis, 15(2):183–202. 12

Gallant, A. R. and Tauchen, G. (1996). Which moments to match? Econometric Theory, 12(4):657–681. 2

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. 1, 2

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477):359–378. 35, 36

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. 36

Gunawan, D., Kohn, R., and Nott, D. (2020). Variational approximation of factor stochastic volatility models. arXiv preprint arXiv:2010.06738. 14

Harvey, A., Koopman, S., and Shephard, N. (2004). State Space and Unobserved Component Models: Theory and Applications. CUP. 1

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. 12

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. 12

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. 12

Koop, G. and Korobilis, D. (2018). Variational Bayes inference in high-dimensional time-varying parameter models. Forthcoming, Journal of Econometrics. 2, 13, 14, 16

Lancaster, T. (2000). The incidental parameter problem since 1948. Journal of econometrics, 95(2):391–413. 2, 10

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. 2, 11, 16, 18, 23

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. 2

Miller, J. W. (2019). Asymptotic normality, concentration, and coverage of generalized posteriors. arXiv preprint arXiv:1907.09611. 28

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. Econometrica, 16(1):1–32. 2

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. 18

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. The Review of Economic Studies, 72(3):821–852. 12, 17, 18

Quiroz, M., Nott, D. J., and Kohn, R. (2018). Gaussian variational approximation for high-dimensional state space models. arXiv preprint arXiv:1801.07873. 2, 14, 16, 18, 21, 23

Ruiz, E. (1994). Quasi-maximum likelihood estimation of stochastic volatility models. Journal of Econometrics, 63(1):289–306. 2

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

Syring, N. and Martin, R. (2020). Gibbs posterior concentration rates under sub-exponential type losses. arXiv preprint arXiv:2012.04505. 28

Tran, M.-N., Nott, D. J., and Kohn, R. (2017). ational Bayes with intractable likelihood. Journal of Computational and Graphical Statistics, 26(4):873–882. 2, 12, 13

Wang, B. and Titterington, D. (2004). Lack of consistency of mean field and variational Bayes approximations for state space models. Neural Processing Letters, 20(3):151–170. 8

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. 10, 11

Yang, Y., Pati, D., Bhattacharya, A., et al. (2020). alpha-variational inference with statistical guarantees. Annals of Statistics, 48(2):886–905. 6

A Proofs of Main Results

Proof of Lemma 3.1. The proof follows a modification of the standard argument in, e.g., Pakes and Pollard Theorem 3.2. Namely, fix $\epsilon > 0$. Then, there exists a $\delta > 0$ 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 $H(\theta) = \operatorname{plim}_n \mathbb{E}\left[m_n(\theta)\right]$, $m_n(\theta) = -\frac{1}{n}\log p_\theta(y_1^n)$ and θ_0 satisfies $H(\theta_0) \leq \inf_{\theta\in\Theta} H(\theta)$. The stated result then follows if the RHS can be shown to be $o(1)$. However,

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

By Assumption 3.1, the first term is $o_p(1)$, and we can concentrate on the second term. Now, from the definition of $\hat{\theta}_n$,

$$0 \leq m_n(\theta_0) + \kappa_n - \left[m_n(\hat{\theta}_n) + \hat{\kappa}_n\right] \leq m_n(\theta_0) - m_n(\hat{\theta}_n) + \kappa_n - \hat{\kappa}_n. \tag{26}$$

However, since the MLE, $\hat{\theta}_{MLE}$ is consistent for θ_0, under Assumption 3.1, then

$$\begin{aligned} m_n(\theta_0) &= m_n(\hat{\theta}_{MLE}) + \{m_n(\theta_0) - m_n(\hat{\theta}_{MLE})\} \\ &= m_n(\hat{\theta}_{MLE}) + o_p(1) \\ &\leq m_n(\hat{\theta}_n) + o_p(1), \end{aligned}$$

where the last line uses the definition of the MLE in (4), and which implies that $m_n(\theta_0) - m_n(\hat{\theta}_n) \leq o_p(1)$, for some positive $o_p(1)$ sequence. Using this in equation (26), and $\hat{\kappa}_n \geq 0$, we have

$$0 \leq m_n(\theta_0) - m_n(\hat{\theta}_n) + \kappa_n - \hat{\kappa}_n \leq \kappa_n - \hat{\kappa}_n + o_p(1) \leq \kappa_n + o_p(1) = o_p(1).$$

Conclude that $H(\theta_0) - H(\hat{\theta}_n) \leq o_p(1)$.

□

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), Syring and Martin (2020) and Miller (2019).

Define $\Pi_n(\Theta) := \int_\Theta \exp\left\{\widehat{L}_n(\theta)\right\} p_\theta(\theta)\mathrm{d}\theta$. By hypothesis, for all $n \geq 1$, $\Pi_n(\Theta) < \infty$. Fix $\epsilon > 0$. For any $\delta > 0$, the idealized posterior can then be stated as

$$\begin{aligned}
\widehat{Q}(A_\epsilon | y_1^n) &= \frac{\int_{A_\epsilon} \exp\left\{\widehat{L}_n(\theta)\right\} p_\theta(\theta)\mathrm{d}\theta}{\int_\Theta \exp\left\{\widehat{L}_n(\theta)\right\} p_\theta(\theta)\mathrm{d}\theta} \\
&= \frac{\int_{A_\epsilon} \exp\left\{-\widehat{L}(\theta_0) + n\delta\right\} \exp\left\{\widehat{L}_n(\theta)\right\} p_\theta(\theta)\mathrm{d}\theta}{\int_\Theta \exp\left\{-\widehat{L}_n(\theta_0) + n\delta\right\} \exp\left\{\widehat{L}_n(\theta)\right\} p_\theta(\theta)\mathrm{d}\theta} \\
&= \frac{\exp\left\{-\widehat{L}_n(\theta_0) + n\delta\right\} \Pi_n(A_\epsilon)}{\exp\left\{-\widehat{L}_n(\theta_0) + n\delta\right\} \Pi_n(\Theta)} = \frac{N_n}{D_n}.
\end{aligned}$$

We treat the numerator and denominator separately.

Write the numeriator as

$$N_n = \int_{A_\epsilon} \exp\left\{n\left[\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_0)/n + \delta\right]\right\} \pi(\theta)\mathrm{d}\theta.$$

Considering the term $\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_0)/n$, we have that

$$\begin{aligned}
\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_0) &\leq 2 \sup_{\theta\in\Theta,\lambda\in\Lambda} |\mathcal{L}_n(\theta,\lambda)/n - \mathcal{L}(\theta,\lambda)| + \mathcal{L}[\theta, \widehat{\lambda}_n(\theta)] - \mathcal{L}[\theta_0, \widehat{\lambda}_n(\theta_0)] \\
&\leq o_p(1) + \left\{\mathcal{L}[\theta, \widehat{\lambda}_n(\theta)] - \mathcal{L}[\theta, \lambda(\theta)]\right\} - \left\{\mathcal{L}[\theta_0, \widehat{\lambda}_n(\theta_0)] - \mathcal{L}[\theta_0, \lambda(\theta_0)]\right\} \\
&\quad + \mathcal{L}[\theta, \lambda(\theta)] - \mathcal{L}[\theta_0, \lambda(\theta_0)] \\
&\leq o_p(1) - \delta
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second from Assumption 3.2(2.a), and

the third follows from consistency of $\widehat{\lambda}_n(\theta)$, uniformly over θ, and 3.2 (3). Thus for any $\epsilon > 0$,

$$\liminf_{n\to\infty} \Pr\left[\sup_{\theta:\|\theta-\theta_0\|>\epsilon} \frac{1}{n}\left\{\widehat{L}_n(\theta) - \widehat{L}_n(\theta_0)\right\} \leq -\delta\right] = 1.$$

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

$$\frac{1}{n}\left\{\widehat{L}_n(\theta) - \widehat{L}_n(\theta_0)\right\} + \delta \leq 0$$

with probability converging to one (wpc1), there exists an n large enough such that

$$\exp\left\{n\left[\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_0)/n + \delta\right]\right\} \leq 1.$$

Consequently, there exists an n large enough such that

$$N_n = \int_{A_\epsilon} \exp\left\{n\left[\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_0)/n + \delta\right]\right\} \pi(\theta)\mathrm{d}\theta \leq \Pi(A_\epsilon) \leq 1.$$

wpc1.

For the denominator, first define $L(\theta) = \mathcal{L}(\theta, \lambda(\theta))$ and the set $G_\delta := \{\theta : L(\theta) - L(\theta_0) < -\delta/2\}$ and note that, for any $\theta \in G_\delta$, by Assumption 3.2(1.b),

$$\left\{\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_0)/n\right\} + \delta/2 \to L(\theta) - L(\theta_0) + \delta/2 < 0,$$

wpc1. Thus, for any $\delta > 0$ and any $\theta \in G_\delta$, $\exp\left\{n\left[\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_0)/n + \delta\right]\right\} \to \infty$ as $n \to \infty$ wpc1. Therefore, from Fatous lemma

$$\begin{aligned}\liminf_{n\to\infty} \exp\left\{-\widehat{L}_n(\theta_0) + n\delta\right\} \Pi_n(G_\delta) &= \liminf_{n\to\infty} \int_{G_\delta} \exp\left\{n\left[\widehat{L}_n(\theta)/n - \widehat{L}_n(\theta_0)/n + \delta\right]\right\} \pi(\theta)\mathrm{d}\theta \\ &\geq \liminf_{n\to\infty} \exp\left[n\delta/4\right] \Pi(G_\delta).\end{aligned}$$

Since $\Pi(G_\delta) > 0$ for any $\delta > 0$, 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). □

Lemma 3.2. The complete data likelihood is proportional to

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

$$= \{2\pi\sigma_0^2\}^{-n} \exp\left\{-\frac{1}{2\sigma_0^2}\left[(x_1^n)'\Omega_n(\theta)x_1^n - 2\alpha(y_1^n)'x_1^n + (y_1^n)'y_1^n\right]\right\},$$

for the matrix

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

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

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

which yields the observable data log-likelihood

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

Following Lemma 3.1, consistency of the variational estimates of θ will require that $n^{-1}\Upsilon_n(q) := \int_\Theta \log p(y_1^n|\theta)q_\theta(\theta)\mathrm{d}\theta - \int_\Theta \mathcal{L}_n(\theta)q_\theta(\theta)\mathrm{d}\theta = o_p(1)$. To this end, consider the infeasible situation where our variational family for θ is

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

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

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

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

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

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

to obtain

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

Similarly, we have that

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

and Jensen's Gap is proportional to

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

To determine whether $\Upsilon_n(q) = o_p(1)$, we must first consider the behavior of the first and second terms in $\Upsilon_n(q)$. For the first term, we note that $|\Omega_n(\theta_0)|$ is a deterministic function of θ_0 and, it can be shown that,[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 $d := \sqrt{a_0^2 - 4\rho_0^2}$, we have

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

and we can define

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

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

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

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

[1] See Lemma B.1 for details.

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

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

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

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

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 $\epsilon > 0$,

$$\Pr\left(|Z_n - \mathbb{E}[Z_n]| > n\epsilon\right) \leq \text{Var}[Z_n]/(n^2\epsilon^2) = \text{Tr}\left[\Omega_n(\theta_0)^{-1}M\Omega_n(\theta_0)^{-1}M\right]/(n^2\epsilon^2) \tag{27}$$

In addition,

$$\text{Var}[Z_n] = \text{Tr}\left[\Omega_n(\theta_0)^{-1}M\Omega_n(\theta_0)^{-1}M\right] \leq n \cdot \sup_{n\geq 1} |\text{diag}\left\{\Omega_n(\theta_0)^{-1}M\Omega_n(\theta_0)^{-1}M\right\}|. \tag{28}$$

Define the sequence $c_n := \sup_{n\geq 1} |\text{diag}\left\{\Omega_n(\theta_0)^{-1}M\Omega_n(\theta_0)^{-1}M\right\}|$. For any $0 \leq \rho_0^2 < 1$ and $0 \leq |\alpha_0| < M < \infty$, the sequence c_n is non-random and bounded for each n, hence we have that $c_n/n \to 0$. From the boundedness of c_n, apply equations (27) and (28) to conclude that, for any $\epsilon > 0$,

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

The above argument and equation (27) allow us to conclude that $C_2(\theta_0) := \frac{\alpha_0^2}{2\sigma_0^2} \lim_{n\to\infty} \text{Tr}[\Omega_n(\theta_0)M]/n$ exists and that

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

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

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

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

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

for α_0 will be consistent iff

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

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

Case 2: $\alpha_0 = 0$**.** Similar to the above, since $H(\theta_0)$ is the minimal entropy, 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 B.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 B.1 as

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

Conclude that VI is consistent iff

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

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

□

B Additional Results

Lemma B.1. *Let*

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

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

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

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

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

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

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

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

$$f_k = a f_{k-1} - bc f_{k-2},$$

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 $f_1 = a$, and $f_2 = a^2 - bc$, so that

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

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

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

Therefore, we can conclude that

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

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

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

□

C Scoring rules

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

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

$$S_{\text{LS}}(P\left(Y_{n+1}|y_1^n\right), y_{n+1}) = \ln p\left(y_{n+1}|y_1^n\right). \tag{29}$$

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

$$S_{\text{CS}}(P\left(Y_{n+1}|y_1^n\right), y_{n+1}) = \ln p\left(y_{n+1}|y_1^n\right) I\left(y_{n+1} \in A\right) + \left[\ln \int_{A^c} p\left(y|y_1^n\right) dy\right] I\left(y_{n+1} \in A^c\right). \tag{30}$$

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

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

$$S_{\text{CRPS}}\left[P(Y_{n+1}|y_1^n), y_{n+1}\right] = -\int_{-\infty}^{\infty}\left[P\left(y|y_1^n\right) - I(y \geq y_{n+1})\right]^2 dy. \tag{31}$$

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

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

$$S_{\text{TWCRPS}}\left[P(Y_{n+1}|y_1^n), y_{n+1}\right] = -\int_0^1 2\left[I(P^{-1}(\alpha|y_1^n) \geq y_{n+1}) - \alpha\right]\left[P^{-1}(\alpha|y_1^n) - y_{n+1}\right](1-\alpha)^2 d\alpha. \tag{32}$$

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

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

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 $\alpha = 0.05$.

D 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 fof 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 fof 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

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

Frazier, David T; Loaiza-Maya, Rubén; Martin, Gael M

01 Shaun McDonald Page 1
30/6/2021 20:33

02 Shaun McDonald Page 1
30/6/2021 20:34

03 Shaun McDonald Page 1
30/6/2021 20:35

04 Shaun McDonald Page 1
30/6/2021 20:36

05 Shaun McDonald Page 2
30/6/2021 20:38

06 Shaun McDonald Page 2
30/6/2021 20:38

07 Shaun McDonald Page 2
30/6/2021 20:38

08 Shaun McDonald Page 2
30/6/2021 20:41
Marginal? Or do there exist SSM's where even the joint likelihood is not analytically known?

09 Shaun McDonald Page 2
30/6/2021 20:41

10 Shaun McDonald Page 2

30/6/2021 20:42

11 Shaun McDonald Page 2

30/6/2021 20:42

12 Shaun McDonald Page 2

30/6/2021 20:42

13 Shaun McDonald Page 2

30/6/2021 20:42

14 Shaun McDonald Page 2

30/6/2021 20:44

15 Shaun McDonald Page 3

30/6/2021 20:50

16 Shaun McDonald Page 4

30/6/2021 22:30
Joint posterior

17 Shaun McDonald Page 4

30/6/2021 22:26

18 Shaun McDonald Page 4

30/6/2021 22:28
Presumably the LA is one such method?

19 Shaun McDonald Page 4

30/6/2021 22:27

20 Shaun McDonald Page 5

30/6/2021 22:35
Isn't this just the negative ELBO?

21 Shaun McDonald Page 6

30/6/2021 22:38

22 Shaun McDonald Page 6

30/6/2021 22:38

23 Shaun McDonald Page 6

30/6/2021 22:38

24 Shaun McDonald Page 6

2/7/2021 18:01

25 Shaun McDonald Page 6

2/7/2021 18:01

26 Shaun McDonald Page 6

5/7/2021 20:34

I think this setup would encompass the LA?

27 Shaun McDonald Page 6

5/7/2021 20:34

28 Shaun McDonald Page 7

5/7/2021 20:38

Is this viable? I suppose you could just assume that some sufficiently large box in R^n would bound all "reasonable" parameter values. In other words, you could just lie!

29 Shaun McDonald Page 7

5/7/2021 20:37

30 Shaun McDonald Page 7

5/7/2021 20:39

Typo here? I assume one of these deltas should be an epsilon.

31 Shaun McDonald Page 7

5/7/2021 20:39

32 Shaun McDonald Page 7

5/7/2021 20:40

33 Shaun McDonald Page 7

5/7/2021 20:39

34 Shaun McDonald Page 7

5/7/2021 20:40

35 Shaun McDonald Page 7

5/7/2021 20:40

36 Shaun McDonald Page 7

5/7/2021 20:42

Does it? Lemma 3.1 is stated as an "if", not an "only if". The proof doesn't seem to go in the other direction, so this doesn't rule out the possibility of consistency when these conditions AREN'T met. TODO: look again after post-vaccine brain fog

37 Shaun McDonald Page 7

5/7/2021 20:43

"Infeasible" in the sense of being minimal over variational distributions?

38 Shaun McDonald Page 7

5/7/2021 20:43

39 Shaun McDonald Page 7

5/7/2021 20:42

40 Shaun McDonald Page 7

5/7/2021 20:44

41 Shaun McDonald Page 7

5/7/2021 20:45

Is there an assumption that the (unscaled) Jensen gap will *grow* with n? I find that surprising.

42 Shaun McDonald Page 7

5/7/2021 20:44

43 Shaun McDonald Page 7

5/7/2021 20:44

44 Shaun McDonald Page 8

5/7/2021 21:01

What is the prior on theta? Without it, doesn't this just reduce to ML?

45 Shaun McDonald Page 8

5/7/2021 21:01

46 Shaun McDonald Page 8

5/7/2021 21:03

This was a little confusing to me. Certainly the actual joint posterior P(X | Y, theta) can be expressed in this form, but there will be various terms that depend on the data Y. Thus, to me it seems a bit apples-and-oranges to compare the quantity \rho in q to the one appearing in the actual model. Thus, who cares if "variational rho" is an inconsistent estimator for "model rho"? I wonder if the optimal "variational rho" can be expressed in closed form as a function of Y and "model rho"?

47 Shaun McDonald Page 8

5/7/2021 20:59

48 Shaun McDonald Page 8

5/7/2021 21:03

This seems important enough that I should come back to it later for some pencil-and-paper reasoning.

49 Shaun McDonald Page 8

5/7/2021 22:45

May be post-vaccine brain fog, but as far as I can tell the proof of Lemma 3.2 uses Lemma 3.1, which is not an "iff" statement. An error?

50 Shaun McDonald Page 8

5/7/2021 21:00

51 Shaun McDonald Page 8

5/7/2021 21:00

52 Shaun McDonald Page 8

5/7/2021 21:04

53 Shaun McDonald Page 8

5/7/2021 21:04

54 Shaun McDonald Page 9

5/7/2021 21:13
This assumes some type of "closure" of \Lambda, such that \hat{\lambda_n} exists

55 Shaun McDonald Page 9

5/7/2021 21:13

56 Shaun McDonald Page 9

5/7/2021 21:08
Does this optimize (12)? If not, in what sense is it "idealized"?

By construction, the variational distribution for states optimizes L_n (which is the only component of (12) that depends on states), but I'm not sure if this sort of "two-step" thing implies optimization of (12) overall.

If this does NOT optimize (12), does a similar theoretical result exist for the q that does? Presumably, this all depends on the way in which the variational distribution is assumed to factorize.

57 Shaun McDonald Page 9

5/7/2021 21:08

58 Shaun McDonald Page 9

5/7/2021 21:08

59 Shaun McDonald Page 9

5/7/2021 21:17
Strange. Usually in a setup like this, the delta and epsilon would be in the opposite places. Perhaps a typo?

60 Shaun McDonald Page 9

5/7/2021 21:17

61 Shaun McDonald Page 9

5/7/2021 22:48

Again, not an "iff" statement, and the proof only goes in one direction. How do we know it doesn't work WITHOUT these assumptions?

62 Shaun McDonald Page 9

5/7/2021 21:22

Variational cdf?

63 Shaun McDonald Page 9

5/7/2021 21:22

64 Shaun McDonald Page 9

6/7/2021 20:47

This uses the optimal variational distribution for the latent states as defined in (14), and a point mass at \theta_0 for q_\theta.

I don't see how this tells us anything about the "idealized" variational posterior, given that it's not even involved in this Jensen's gap? Maybe there'd be a clearer connection if Lemma 3.3 was an "iff" statement, but it's not.

65 Shaun McDonald Page 10

5/7/2021 21:15

3.1 = "frequentist" (variational point estimate)? Although the restated KL_c in Section 3.2 still has a term for the prior of theta, so I'm not sure if the terminology here is entirely sound.
3.3 = Bayesian ("idealized posterior")

66 Shaun McDonald Page 10

5/7/2021 21:15

67 Shaun McDonald Page 10

5/7/2021 22:48

Again, this (central?) point seems to be predicated on the idea that the converses of Lemmas 3.1 and 3.3 are also true. Am I missing something here?

68 Shaun McDonald Page 10

5/7/2021 22:51

69 Shaun McDonald Page 10

5/7/2021 22:52

70 Shaun McDonald Page 10

5/7/2021 22:52

71 Shaun McDonald Page 10

5/7/2021 22:57

TODO: read this to explore connection to LA

72 Shaun McDonald Page 10

5/7/2021 22:53

73 Shaun McDonald Page 10

5/7/2021 22:53

74 Shaun McDonald Page 10

5/7/2021 22:53

75 Shaun McDonald Page 10

5/7/2021 22:53

This approach appears similar to the LA, in that it has this nested optimization thing going on.

Of course, to make this connection you'd need q such that L_n (10) worked out equal to the LA. Bit of pencil-and-paper could take care of that.

76 Shaun McDonald Page 11

5/7/2021 22:54

77 Shaun McDonald Page 11

5/7/2021 22:59

78 Shaun McDonald Page 11

5/7/2021 23:00

Stochastic gradient ascent - generate latent variables from P(X|Y, theta), use simulations to obtain stochastic estimates of gradients, optimize ELBO for variational density of theta using these. See Section 8.3 of BMC notes.

79 Shaun McDonald Page 11

5/7/2021 23:00

80 Shaun McDonald Page 12

5/7/2021 23:02

81 Shaun McDonald Page 12

5/7/2021 23:02

82 Shaun McDonald Page 12

5/7/2021 23:03

I barely know what "particle filtering" is. That's another todo.

83 Shaun McDonald Page 12

5/7/2021 23:02

84 Shaun McDonald Page 12

6/7/2021 17:14

I didn't find this explanation very clear. The Tran et al. paper explains things better.

85 Shaun McDonald Page 12

5/7/2021 23:03

86 Shaun McDonald Page 12

6/7/2021 17:24

In other words, you obtain an unbiased estimate of the log-likelihood (via particle filtering, I guess), THEN define an (actually unobserved) variable z in terms of that

87 Shaun McDonald Page 12

6/7/2021 17:23

88 Shaun McDonald Page 12

5/7/2021 23:03

89 Shaun McDonald Page 12

6/7/2021 17:07

90 Shaun McDonald Page 12

6/7/2021 17:20

Jensen's inequality on the second term (-log)

91 Shaun McDonald Page 13

6/7/2021 17:21

92 Shaun McDonald Page 13

6/7/2021 17:24

TODO: read more about SMC and PMCMC

93 Shaun McDonald Page 13

6/7/2021 17:24

94 Shaun McDonald Page 13

6/7/2021 17:26

This paper, and the references therein, seem to explain it a bit more

95 Shaun McDonald Page 13

6/7/2021 17:42

As a function of theta and sigma^2

96 Shaun McDonald Page 13

6/7/2021 17:26

97 Shaun McDonald Page 13

6/7/2021 17:28

98 Shaun McDonald Page 13

6/7/2021 20:55

This is just the outer integrand in the Jensen gap formula above, evaluated at \theta = \theta_0. What definition of q(\theta_0, z) allows this?

I thought maybe it involved a point mass on \theta_0 like in Section 3, but given q_\theta IS the variational distribution here that doesn't make any sense.

99 Shaun McDonald Page 13

6/7/2021 20:53

What does this mean? Isn't the Jensen's gap an integral over theta?

100 Shaun McDonald Page 13

6/7/2021 20:51

101 Shaun McDonald Page 13

6/7/2021 20:54

102 Shaun McDonald Page 13

6/7/2021 20:56

They would, if this derivation made sense - but I'm not convinced it does.

103 Shaun McDonald Page 13

6/7/2021 20:56

104 Shaun McDonald Page 13

6/7/2021 20:59

105 Shaun McDonald Page 13

6/7/2021 20:59

106 Shaun McDonald Page 13

6/7/2021 21:02

Why no slope? Presumably that'd allow for more flexible approximation to the true nonlinear relationship.

107 Shaun McDonald Page 13

6/7/2021 21:01

108 Shaun McDonald Page 13

6/7/2021 21:01

109 Shaun McDonald Page 14

6/7/2021 21:02

110 Shaun McDonald Page 14

6/7/2021 21:10

111 Shaun McDonald Page 14

6/7/2021 21:10

112 Shaun McDonald Page 14

6/7/2021 21:03

113 Shaun McDonald Page 14

6/7/2021 21:04

114 Shaun McDonald Page 14

6/7/2021 21:04

As I recall, this is why we sort of fell out of love with VB a couple of years ago

115 Shaun McDonald Page 14

6/7/2021 21:04

116 Shaun McDonald Page 14

6/7/2021 21:04

117 Shaun McDonald Page 14

6/7/2021 21:04

118 Shaun McDonald Page 14

6/7/2021 21:06

Again, Lemma 3.3 was about the "idealized" variational posterior for \theta. I'm not clear if that turns out to be the one that's optimal w.r.t. the ELBO.

119 Shaun McDonald Page 14

6/7/2021 21:05

120 Shaun McDonald Page 14

6/7/2021 21:06

121 Shaun McDonald Page 15

6/7/2021 21:09

What about point estimation for \theta, as explored in the previous sections?

122 Shaun McDonald Page 15

6/7/2021 21:09

123 Shaun McDonald Page 15

6/7/2021 21:18

This feels somehow immoral but I can't articulate why

124 Shaun McDonald Page 15

6/7/2021 21:18

125 Shaun McDonald Page 16

6/7/2021 21:19

126 Shaun McDonald Page 16

6/7/2021 21:19

127 Shaun McDonald Page 16

6/7/2021 21:19

128 Shaun McDonald Page 16

6/7/2021 21:20

129 Shaun McDonald Page 16

6/7/2021 21:20

Right, the particle filtering thing

130 Shaun McDonald Page 16

6/7/2021 21:20

131 Shaun McDonald Page 16

6/7/2021 21:21

Guess this depends what you mean by "model structure". Technically, the prior for \theta is misspecified; does that not count as "model structure"?

132 Shaun McDonald Page 16

6/7/2021 21:20

133 Shaun McDonald Page 16

6/7/2021 21:21

134 Shaun McDonald Page 16

6/7/2021 21:22

Air quotes

135 Shaun McDonald Page 16

6/7/2021 21:22

136 Shaun McDonald Page 16

6/7/2021 21:23

137 Shaun McDonald Page 17

6/7/2021 22:22

138 Shaun McDonald Page 17

6/7/2021 21:25

139 Shaun McDonald Page 17

6/7/2021 21:25

140 Shaun McDonald Page 17

6/7/2021 21:26

141 Shaun McDonald Page 18

6/7/2021 21:30

142 Shaun McDonald Page 18

6/7/2021 21:31

Typo

143 Shaun McDonald Page 18

6/7/2021 21:50

This just seems like a fancy way of saying "the variational density factorizes over \theta and X". I glanced at the QNK paper, and nothing in my skimming suggested it was anything other than that. Seems weird to phrase it like this.

144 Shaun McDonald Page 18

6/7/2021 21:51

145 Shaun McDonald Page 18

6/7/2021 21:31

146 Shaun McDonald Page 18

6/7/2021 21:54

147 Shaun McDonald Page 19

6/7/2021 21:56

Again, how is this any different from saying its the ELBO-optimizing density within the class that factorizes in this way?

148 Shaun McDonald Page 19

6/7/2021 21:56

149 Shaun McDonald Page 19

6/7/2021 21:57

150 Shaun McDonald Page 20

6/7/2021 22:08

Inferential accuracy (and consistency, I guess) is assessed here by comparing the means of the [X | \theta] draws to the true means [X | \theta_0].
Wouldn't it be more direct to compare the means of the posterior \theta draws to the true value \theta_0? Most of the paper talks about "inference on \theta", and then all of a sudden here we're doing inference on the states.

151 Shaun McDonald Page 21

6/7/2021 22:20

152 Shaun McDonald Page 21

6/7/2021 22:23

153 Shaun McDonald Page 21

6/7/2021 22:24

154 Shaun McDonald Page 21

6/7/2021 22:24

155 Shaun McDonald Page 21

6/7/2021 22:24

156 Shaun McDonald Page 23

6/7/2021 22:25

157 Shaun McDonald Page 23

6/7/2021 22:25

158 Shaun McDonald Page 23

6/7/2021 22:25

But not necessary, as I noted above

159 Shaun McDonald Page 23

6/7/2021 22:25

160 Shaun McDonald Page 23

6/7/2021 22:26

161 Shaun McDonald Page 23

6/7/2021 22:26

This is certainly shown numerically, but (assuming I haven't missed something, which I very well might have) not theoretically

162 Shaun McDonald Page 23

6/7/2021 22:26

163 Shaun McDonald Page 23

6/7/2021 22:26

164 Shaun McDonald Page 23

6/7/2021 22:27

165 Shaun McDonald Page 23

6/7/2021 22:27

Sort of like how non-Gaussian functions can have a "correct" LA?

166 Shaun McDonald Page 23

6/7/2021 22:27

167 Shaun McDonald Page 23

6/7/2021 22:27

168 Shaun McDonald Page 27

6/7/2021 17:13