Viking: Variational Bayesian Variance Tracking

Joseph de Vilmarest, Olivier Wintenberger

Viking: Variational Bayesian Variance Tracking

Joseph de Vilmarest and Olivier Wintenberger

***Abstract*—We consider the problem of time series forecasting in an adaptive setting. We focus on the inference of state-space models under unknown and potentially time-varying noise variances. We introduce an augmented model in which the variances are represented as auxiliary gaussian latent variables in a tracking mode. As variances are nonnegative, a transformation is chosen and applied to these latent variables. The inference relies on the online variational Bayesian methodology, which consists in minimizing a Kullback-Leibler divergence at each time step. We observe that the minimum of the Kullback-Leibler divergence is an extension of the Kalman filter taking into account the variance uncertainty. We design a novel algorithm, named Viking, using these optimal recursive updates. For auxiliary latent variables, we use second-order bounds whose optimum admit closed-form solutions. Experiments on synthetic data show that Viking behaves well and is robust to misspecification.**

***Index Terms*—adaptive forecasting, state-space model, time series, variance estimation**

I. Introduction

LINEAR state-space models have been widely used to model a time series as a gaussian random variable whose mean is a time-varying linear function of covariates. The linear parameter is a latent variable called state, and the hyperparameters of the state-space model are the covariance matrices of the state and space noises. When these variances are known, the recursive estimation is realized by Kalman filtering [1].

However the state and space noise variances are unknown in most practical applications. A wide literature has emerged to choose them. The estimation of unknown fixed variances on a historical data set is generally realized maximizing the likelihood (see for instance [2], [3]). Another approach is to estimate these variances (fixed or not) in an online fashion, that is adaptive filtering [4].

Recently, recursive variational Bayesian (VB) methods as introduced in [5], [6] have gathered attention in the Kalman filtering community. The objective is the online estimation of potentially time-variant parameters. The difference with the classical Bayesian method is that an approximation is realized at each step in order to make the inference tractable: the distribution of the parameters is estimated by simple factorized distributions. The best factorized distribution is defined as the one minimizing its Kullback-Leibler divergence with the posterior.

A VB approach was first applied to estimate the observation noise covariance matrix in a Kalman filter [7], then extended in [8] to be robust to non-gaussian noise and in [9] to nonlinear state-space models. The covariance matrix is assumed diagonal and the prior used is a product of inverse gamma distributions. To allow for a dynamical noise variance the author use a forgetting factor, multiplying the variances of the inverse gamma posterior distributions by a constant. The method was extended with an inverse Wishart prior [10]. At the same time the authors apply the VB approach to correct the covariance matrix of the state after applying Kalman recursions with an inaccurate state noise covariance matrix. The inverse Wishart distribution appears as a nice conjugate prior to generalize the inverse gamma distribution. More recently another adaptive Kalman filter was proposed in [11] to estimate simultaneously the state and space noise covariance matrices. The method uses Kalman filtering and smoothing on a slide window and could be described as an online Expectation-Maximization algorithm. In all these methods the dynamics of the variances is introduced through a forgetting factor.

Up to our knowledge, to deal with unknown covariance matrices in state-space models all existing methods apply at each step the standard Kalman filter with an estimate of the variances updated in an adaptive fashion. We claim that it is suboptimal and that the recursive update of the state estimates should leverage the variance uncertainty. In this article we treat the variances as auxiliary latent variables yielding an important degree of freedom in an augmented latent representation. We apply the VB approach and we rely on second-order upper-bounds to tackle the intractability of the VB step.

A. Overview

We present in Section II the state-space inference problem, we introduce our augmented dynamical model and the VB principle. As the minimization problem derived in the VB approach doesn't admit a closed-form solution, we derive in Section III an approximation. The algorithm is detailed in Section IV, and we provide experimental results in Section V.

B. Notations

Besides canonical notations we define:

- $\mathcal{N}(x \mid \mu, \Sigma)$ the probability density function at point x of the distribution $\mathcal{N}(\mu, \Sigma)$.
- For any distribution p and function Φ, $\mathbb{E}_{x\sim p}[\Phi(x)]$ is defined as $\int p(x)\Phi(x)dx$.
- For any matrix M, Δ_M is the vector composed of the diagonal coefficients of M. Reciprocally, for any vector v, D_v is the diagonal matrix whose diagonal is composed of the coefficients of v.
- If $\phi : \mathbb{R} \to \mathbb{R}$ and $x \in \mathbb{R}^d$, $\phi(x)$ is the d-dimensional vector obtained by applying ϕ to each coordinate of x.

J. de Vilmarest (joseph.de_vilmarest@sorbonne-universite.fr) and O. Wintenberger (olivier.wintenberger@sorbonne-universite.fr) are with the Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, CNRS.
J. de Vilmarest is also affiliated with Électricité de France R&D.

II. Variance Tracking

We consider the problem of time series forecasting in the univariate setting for simplicity. At each time t we aim at forecasting $y_t \in \mathbb{R}$. To that end we have access to covariates $x_t \in \mathbb{R}^d$ as well as the past observations $x_1, y_1, \dots x_{t-1}, y_{t-1}$. We focus on a state-space representation where y_t is modelled as a linear function of x_t whose linear parameter evolves dynamically:

$$\theta_t = K\theta_{t-1} + \eta_t\,,$$
$$y_t = \theta_t^\top x_t + \varepsilon_t\,,$$

where $\eta_t \sim \mathcal{N}(0, Q_t)$ and $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ are the state and space noises, and the state follows the initial distribution $\theta_0 \sim \mathcal{N}(\hat{\theta}_0, P_0)$. When σ_t^2 and Q_t are known, the state vector θ_t given the past observations follows a gaussian distribution whose mean and covariance can be estimated recursively by the standard Kalman filter [1]. We focus on the setting where these variances are unknown and need to be estimated jointly with the state.

A. Dynamical Variances

A way to introduce a dynamical estimation of σ_t^2 and Q_t is to treat them as latent variables in addition to the state vector. Gaussian distributions are appealing to model a dynamic latent variable. Therefore we choose a gaussian prior for the variances as for the state vector. However a variance is necessarily nonnegative, thus we consider transforms of gaussian distributions. Precisely $\sigma_t^2 = \exp(a_t)$ and $Q_t = f(b_t)$, where a_t, b_t follow gaussian distributions. We detail the choice of f in Section II-E where we define either scalar covariance matrices (proportional to I) or diagonal ones. Note that b_t can be of any dimension, as long as $f(b_t)$ is a $d \times d$ positive semidefinite matrix. Our dynamical model is summarized as follows:

$$\begin{aligned}
&\theta_0 \sim \mathcal{N}(\hat{\theta}_0, P_0)\,, \quad a_0 \sim \mathcal{N}(\hat{a}_0, s_0)\,, \quad b_0 \sim \mathcal{N}(\hat{b}_0, \Sigma_0)\,,\\
&a_t - a_{t-1} \sim \mathcal{N}(0, \rho_a)\,, \quad b_t - b_{t-1} \sim \mathcal{N}(0, \rho_b I)\,,\\
&\theta_t - K\theta_{t-1} \sim \mathcal{N}(0, f(b_t))\,,\\
&y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \exp(a_t))\,.
\end{aligned}$$

In these equations we implicitly assume that we have

$$\begin{aligned}
&p(\theta_t, a_t, b_t \mid \theta_{t-1}, a_{t-1}, b_{t-1})\\
&\quad = p(\theta_t \mid \theta_{t-1}, b_t) p(a_t \mid a_{t-1}) p(b_t \mid b_{t-1})\,.
\end{aligned}$$

B. Bayesian Approach

We apply a Bayesian approach in order to estimate jointly the state θ_t and the latent variables a_t, b_t given the past observations. Note however that the problem at hand is the forecast of y_t thus the latent variable of interest is θ_t. The estimation of a_t is necessary for a probabilistic forecast of y_t since it drives the noise variance. The latent variable b_t is added to open enough flexibility for the estimation of the other variables in a dynamical way. Formally we introduce the filtration of the past observations $\mathcal{F}_t = \sigma(x_1, y_1, \dots, x_t, y_t)$. At each iteration t, the Bayesian approach consists in a prediction step using the model's assumptions and a filtering step using Bayes' rule:

$$\begin{aligned}
&\text{Prediction:} && p(\theta_t, a_t, b_t \mid \mathcal{F}_{t-1})\,,\\
&\text{Filtering:} && p(\theta_t, a_t, b_t \mid \mathcal{F}_t)\,.
\end{aligned}$$

In the case of known variances resolved by the Kalman filter, the prediction step yields $\hat{\theta}_{t|t-1}$ and $P_{t|t-1}$, the expected value and covariance matrix of θ_t given the filtration $\mathcal{F}_{t-1}$. Furthermore we have $p(\theta_t \mid \mathcal{F}_{t-1}) = \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t-1}, P_{t|t-1})$. Then the filtering step yields $\hat{\theta}_{t|t}$ and $P_{t|t}$ such that the posterior distribution is $p(\theta_t \mid \mathcal{F}_t) = \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})$.

However in our variance tracking model the posterior distribution $p(\cdot \mid \mathcal{F}_t)$ is analytically intractable, thus we estimate it with simple distributions.

C. Variational Bayesian Approach

A standard approach, referred to as recursive Variational Bayes (VB), is to approximate recursively the posterior distribution with a factorized distribution where each component is of a simple form [6]. We look for $\hat{\theta}_{t|t}, P_{t|t}, \hat{a}_{t|t}, s_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$ such that the product of gaussian distributions $\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})$ is the best approximation of the posterior distribution. The approximation is quantified by the Kullback-Leibler (KL) divergence:

$$KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; p(\cdot \mid \mathcal{F}_t)\Big), \quad (1)$$

where $KL(p \;||\; q) = \mathbb{E}_{x\sim p(x)}[\log(p(x)/q(x))]$. At each step, the VB approach yields a coupled optimization problem in the three gaussian distributions.

Propagating the factorized approximation

$$\begin{aligned}
p(\theta_{t-1}, a_{t-1}, b_{t-1} \mid \mathcal{F}_{t-1}) \approx\; &\mathcal{N}(\theta_{t-1} \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1})\\
&\mathcal{N}(a_t \mid \hat{a}_{t|t}, s_{t|t})\mathcal{N}(b_{t-1} \mid \hat{b}_{t-1|t-1}, \Sigma_{t-1|t-1})\,,
\end{aligned}$$

the prediction step becomes:

$$\begin{aligned}
&p(\theta_t, a_t, b_t \mid \mathcal{F}_{t-1})\\
&\approx \int \mathcal{N}(\theta_t - K\theta_{t-1} \mid 0, f(b_t))\mathcal{N}(a_t - a_{t-1} \mid 0, \rho_a)\\
&\qquad \mathcal{N}(b_t - b_{t-1} \mid 0, \rho_b I)\mathcal{N}(\theta_{t-1} \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1})\\
&\qquad \mathcal{N}(a_t \mid \hat{a}_{t-1|t-1}, s_{t-1|t-1})\mathcal{N}(b_t \mid \hat{b}_{t-1|t-1}, \Sigma_{t-1|t-1})\\
&\qquad\quad d\theta_{t-1} da_{t-1} db_{t-1}\\
&\approx \mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + f(b_t))\\
&\qquad \mathcal{N}(a_t \mid \hat{a}_{t-1|t-1}, s_{t-1|t-1} + \rho_a)\\
&\qquad \mathcal{N}(b_t \mid \hat{b}_{t-1|t-1}, \Sigma_{t-1|t-1} + \rho_b I)\,.
\end{aligned}$$

Treating the approximation at time $t-1$ as a prior at time t we obtain the following posterior distribution:

$$\begin{aligned}
p(\theta_t, a_t, b_t \mid \mathcal{F}_t) = \;&\mathcal{N}(y_t \mid \theta_t^\top x_t, \exp(a_t))\\
&\mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + f(b_t))\\
&\mathcal{N}(a_t \mid \hat{a}_{t-1|t-1}, s_{t-1|t-1} + \rho_a)\\
&\mathcal{N}(b_t \mid \hat{b}_{t-1|t-1}, \Sigma_{t-1|t-1} + \rho_b I)\frac{p(x_t, \mathcal{F}_{t-1})}{p(\mathcal{F}_t)}\,. \quad (2)
\end{aligned}$$

This last equation defines the posterior that we plug in (1) to obtain the optimization problem that we would like to solve recursively.

The term $\mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + Q_t)$ on the second line, which would appear with any model for Q_t, makes a conjugate prior for Q_t impractical in a VB method to estimate the posterior distribution of the state and the variances. The approach proposed by [11] consists in applying a few iterations of Kalman smoothing with the previous estimates of the variances $\hat{\sigma}^2_{t-1}$ and $\hat{Q}_{t-1}$. Then the authors estimate the posterior distribution of σ^2_t, Q_t given $\mathcal{F}_t$ and the distribution of θ_{t-k} estimated by Kalman smoothing given $\mathcal{F}_t, \hat{\sigma}^2_{t-1}, \hat{Q}_{t-1}$. In that way they get rid of the crossed factor involving θ_t and Q_t, and they obtain exact estimation of the posterior distribution of the variances. Our approach does the opposite on that part. We build on that crossed factor to avoid Kalman smoothing, at the cost of the need of approximations in the posterior estimation.

D. KL Derivation and Optimum in $\hat{\theta}_{t|t}, P_{t|t}$

We first present a detailed expression of the KL divergence defined in (1) in the following Lemma.

Lemma 1. *There exists a constant c independent of $\hat{\theta}_{t|t}, P_{t|t}, \hat{a}_{t|t}, s_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$ such that*

$$\begin{aligned}
&KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; p(\cdot \mid \mathcal{F}_t)\Big) \\
&\quad = -\frac{1}{2}\log\det P_{t|t} - \frac{1}{2}\log(s_{t|t}) \\
&\qquad + \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)\exp(-\hat{a}_{t|t} + \frac{1}{2}s_{t|t}) \\
&\qquad - \frac{1}{2}\log\det\Sigma_{t|t} + \frac{1}{2}\mathbb{E}_{b_t\sim\mathcal{N}(\hat{b}_{t|t},\Sigma_{t|t})}[\psi_t(b_t)] \\
&\qquad + \frac{1}{2(s_{t-1|t-1}+\rho_a)}(s_{t|t} + (\hat{a}_{t|t} - \hat{a}_{t-1|t-1})^2) + \frac{1}{2}\hat{a}_{t|t} \\
&\qquad + \frac{1}{2}\operatorname{Tr}\Big(\big(\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^\top\big) \\
&\qquad\qquad (\Sigma_{t-1|t-1} + \rho_b I)^{-1}\Big) + c\,,
\end{aligned}$$

where

$$\begin{aligned}
\psi_t(b_t) &= \log\det(KP_{t-1|t-1}K^\top + f(b_t)) \\
&\quad + \operatorname{Tr}\Big((P_{t|t} + (\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})^\top) \\
&\qquad (KP_{t-1|t-1}K^\top + f(b_t))^{-1}\Big)\,.
\end{aligned}$$

We easily obtain a closed-form solution to minimize the KL divergence with respect to $\hat{\theta}_{t|t}, P_{t|t}$.

Theorem 2. *Given $\hat{a}_{t|t}, s_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$, the values of $\hat{\theta}_{t|t}, P_{t|t}$ minimizing the KL divergence are given by*

$$A_t = \mathbb{E}_{b_t\sim\mathcal{N}(\hat{b}_{t|t},\Sigma_{t|t})}[(KP_{t-1|t-1}K^\top + f(b_t))^{-1}]\,, \tag{3}$$

$$P_{t|t} = A_t^{-1} - \frac{A_t^{-1}x_t x_t^\top A_t^{-1}}{x_t^\top A_t^{-1}x_t + \exp(\hat{a}_{t|t} - \frac{1}{2}s_{t|t})}\,, \tag{4}$$

$$\hat{\theta}_{t|t} = K\hat{\theta}_{t-1|t-1} + \frac{P_{t|t}x_t}{e^{\hat{a}_{t|t} - s_{t|t}/2}}(y_t - x_t^\top K\hat{\theta}_{t-1|t-1})\,. \tag{5}$$

Note that the updates defined above are the ones of the Kalman filter with known variances σ_t^2 and Q_t, where we have replaced σ_t^2 with $\exp(\hat{a}_{t|t} - \frac{1}{2}s_{t|t})$ which is $\mathbb{E}_{a_t\sim\mathcal{N}(\hat{a}_{t|t},s_{t|t})}[\exp(a_t)^{-1}]^{-1}$ and $KP_{t-1|t-1}K^\top + Q_t$ with $\mathbb{E}_{b_t\sim\mathcal{N}(\hat{b}_{t|t},\Sigma_{t|t})}[(KP_{t-1|t-1}K^\top + f(b_t))^{-1}]^{-1}$. If $s_{t|t} = 0, \Sigma_{t|t} = 0$ then we know the variances and we obtain the Kalman filter with $\sigma_t^2 = \exp(\hat{a}_{t|t})$ and $Q_t = f(\hat{b}_{t|t})$. Otherwise if $\Sigma_{t|t} \neq 0$, the result states that the update of the Kalman filter with unbiased estimated variances in place of the unknown variances is suboptimal in the sense of the Kullback-Leibler divergence. It implies also that we don't expect to obtain unbiased estimates of the variances. The same conclusion would follow if one adapted the classical VB approach of [12] to our framework.

It is important to remark that as long as $\rho_b > 0$ we do not have the convergence of $\Sigma_{t|t}$ to 0. Therefore we do not recover the standard Kalman filter asymptotically. On the contrary, existing adaptive Kalman filters use the standard Kalman recursive updates with estimates of the variances [7]–[11]. Therefore, in a well-specified setting where the state-space model is the underlying generating process, our method should be outperformed by adaptive Kalman filters whose variance estimates are consistent. We believe this drawback is a reasonable price to pay to get robustness to misspecification.

Furthermore note that (5) may be interpreted as a gradient step on the quadratic loss, where instead of a gradient step size we have the preconditioning matrix $P_{t|t}/\exp(\hat{a}_{t|t} - \frac{1}{2}s_{t|t})$. Therefore the algorithm derived in this article may be seen as a way to parameterize a second order stochastic gradient algorithm.

E. Choice of f

The natural transform for the latent variables a_t and b_t is the exponential, see [13] for a filter on latent variables lying in a Riemannian manifold. We use the exponential to represent σ_t^2. However setting $f(b_t) = \exp(b_t)I$ for a unidimensional b_t contradicts a careful property that we define as follows using the gradient interpretation of Section II-D. We claim that the algorithm should be more careful with uncertainty ($\Sigma_{t|t} \succ 0$) than without ($\Sigma_{t|t} = 0$). By more careful we mean smaller gradient steps, that is formally $A_t^{-1} \preccurlyeq KP_{t-1|t-1}K^\top + f(\hat{b}_{t|t})$. By Jensen's inequality we have

$$A_t \succcurlyeq \Big(KP_{t-1|t-1}K^\top + \mathbb{E}_{b_t\sim\mathcal{N}(\hat{b}_{t|t},\Sigma_{t|t})}[f(b_t)]\Big)^{-1}\,.$$

Therefore a sufficient (but not necessary) condition providing the careful property is f concave, again thanks to Jensen, and that is the contrary of the exponential. Unfortunately we cannot have both f concave and $f \succcurlyeq 0$. We propose to use a function which is zero on negative numbers and concave elsewhere:

$$\phi(b) = \begin{cases} 0 & if\ b < 0\,, \\ \log(1+b) & if\ b \geq 0\,. \end{cases}$$

Then we consider two settings for f: First a scalar setting where $f(b_t) = \phi(b_t)I$ for a unidimensional b_t. Second, a diagonal setting where $b_t \in \mathbb{R}^d$ and $f(b_t) = D_{\phi(b_t)}$ is a diagonal matrix whose diagonal coefficients are defined by the ϕ transform applied to each coefficient of b_t.

III. Approximate Variational Bayes

Theorem 2 realizes the exact optimum of the KL divergence with respect to $\hat{\theta}_{t|t}, P_{t|t}$. To obtain closed-form solutions of the minimum with respect to the other parameters we need additional approximations. In this section, we use the first two moments of gaussian distributions in second-order upper-bounds. That yields closed-form approximations to the VB recursive step with respect to $\hat{a}_{t|t}, s_{t|t}$ and $\hat{b}_{t|t}, \Sigma_{t|t}$. Minimizing the upper-bounds does not necessarily lead to minimizing the KL divergence, but it yields the guarantee of decreasing the instantaneous KL divergence at each step.

A. Optimum in $\hat{a}_{t|t}, s_{t|t}$

We first present recursive updates for $\hat{a}_{t|t}, s_{t|t}$.

1) Optimum in $s_{t|t}$: We are looking for $s_{t|t} \geq 0$ minimizing the KL divergence. As the conditional variance of a_t given $\mathcal{F}_{t-1}$ is $s_{t-1|t-1} + \rho_a$, we look for $s_{t|t}$ in the interval $[0, s_{t-1|t-1}+\rho_a]$. In this interval we simply use a linear upper-bound for the exponential:

Proposition 3. *For any $s_{t|t} \in [0, s_{t-1|t-1} + \rho_a]$ we have*

$$\begin{aligned} &KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; p(\cdot \mid \mathcal{F}_t)\Big) \\ &\quad \leq \frac{1}{4}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t) e^{-\hat{a}_{t|t}} s_{t|t} \\ &\qquad + \frac{1}{2}(s_{t-1|t-1} + \rho_a)^{-1} s_{t|t} - \frac{1}{2}\log(s_{t|t}) + c_s \,, \end{aligned}$$

where c_s is a constant independent of $s_{t|t}$. Furthermore, the upper-bound is minimized by:

$$\begin{aligned} s_{t|t} = \Big((s_{t-1|t-1} + \rho_a)^{-1} \\ + \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t) e^{-\hat{a}_{t|t}}\Big)^{-1} . \end{aligned} \quad (6)$$

2) Optimum in $\hat{a}_{t|t}$: To upper-bound the exponential with a polynomial form also in $\hat{a}_{t|t}$ we need to bound $\hat{a}_{t|t}$, and we consider the segment $[\hat{a}_{t-1|t-1} - M_a, \hat{a}_{t-1|t-1} + M_a]$ (we set arbitrarily $M_a = 3 s_{t-1|t-1}$).

Proposition 4. *For any $\hat{a}_{t|t} \in [\hat{a}_{t-1|t-1} - M_a, \hat{a}_{t-1|t-1} + M_a]$ we have*

$$\begin{aligned} &KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; p(\cdot \mid \mathcal{F}_t)\Big) \\ &\quad \leq \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t) e^{-\hat{a}_{t-1|t-1} + s_{t|t}/2} \\ &\qquad \Big(- (\hat{a}_{t|t} - \hat{a}_{t-1|t-1}) + \frac{e^{M_a}}{2}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1})^2\Big) \\ &\qquad + \frac{1}{2}(s_{t-1|t-1} + \rho_a)^{-1}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1})^2 + \frac{1}{2}\hat{a}_{t|t} + c_a \,, \end{aligned}$$

where c_a is a constant independent of $\hat{a}_{t|t}$. Furthermore the upper-bound is minimized by:

$$\begin{aligned} \hat{a} &= \hat{a}_{t-1|t-1} + \frac{1}{2}\Big(\frac{1}{s_{t-1|t-1} + \rho_a} \\ &\quad + \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t) e^{-\hat{a}_{t-1|t-1} + s_{t|t}/2 + M_a}\Big)^{-1} \\ &\qquad \Big(((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t) e^{-\hat{a}_{t-1|t-1} + s_{t|t}/2} - 1\Big) , \\ \hat{a}_{t|t} &= \max(\min(\hat{a}, \hat{a}_{t-1|t-1} + M_a), \hat{a}_{t-1|t-1} - M_a) \,. \end{aligned} \quad (7)$$

We note that $((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t) e^{-\hat{a}_{t-1|t-1} + s_{t|t}/2} - 1$ is the gradient with respect to $\hat{a}$ of

$$\begin{aligned} &\mathbb{E}_{(\theta_t, a_t) \sim \mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times \mathcal{N}(\hat{a}, s_{t|t})}[\log \mathcal{N}(y_t \mid \theta_t^\top x_t, \exp(a_t))] \\ &\quad = -\frac{1}{2}\hat{a} - \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t) e^{-\hat{a} + s_{t|t}/2} \,, \end{aligned}$$

therefore (7) may be seen as a projected gradient step on an expected log-likelihood.

B. Optimum in $\hat{b}_{t|t}, \Sigma_{t|t}$

The minimum of the Kullback-Leibler is also intractable in $\hat{b}_{t|t}, \Sigma_{t|t}$ due to the absence of analytical form for the expected value of ψ_t. In the following we focus on the specific settings that are introduced in Section II-E, namely the scalar setting $f(b_t) = \phi(b_t) I$ and the diagonal setting $f(b_t) = D_{\phi(b_t)}$. For these two possible choices of f we have the following second-order upper-bound for ψ_t:

Proposition 5. *In the scalar and diagonal settings defined in Section II-E, for any t such that $f(\hat{b}_{t-1|t-1}) \succ 0$, the following holds for any b_t in a neighbourhood of $\hat{b}_{t-1|t-1}$:*

$$\begin{aligned} \psi_t(b_t) \leq \psi_t(\hat{b}_{t-1|t-1}) + \frac{\partial \psi_t}{\partial b_t}\Big|_{\hat{b}_{t-1|t-1}}^\top (b_t - \hat{b}_{t-1|t-1}) \\ + \frac{1}{2}(b_t - \hat{b}_{t-1|t-1})^\top H_t (b_t - \hat{b}_{t-1|t-1}) \,, \end{aligned}$$

where $B_t = P_{t|t} + (\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})^\top$, $C_t = K P_{t-1|t-1} K^\top + f(\hat{b}_{t-1|t-1})$, and then

$$\begin{aligned} \frac{\partial \psi_t}{\partial b_t}\Big|_{\hat{b}_{t-1|t-1}} &= \operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1}))\phi'(\hat{b}_{t-1|t-1}) \,, \\ H_t &= -\operatorname{Tr}(C_t^{-1} B_t C_t^{-1})\phi''(\hat{b}_{t-1|t-1}) \\ &\quad + 2\operatorname{Tr}(C_t^{-2} B_t C_t^{-1})\phi'(\hat{b}_{t-1|t-1})^2 \,, \end{aligned}$$

in the scalar setting, and

$$\begin{aligned} \frac{\partial \psi_t}{\partial b_t}\Big|_{\hat{b}_{t-1|t-1}} &= \Delta_{C_t^{-1}(I - B_t C_t^{-1})} \odot \phi'(\hat{b}_{t-1|t-1}) \,, \\ H_t &= -\Big(C_t^{-1} B_t C_t^{-1} D_{\phi''(\hat{b}_{t-1|t-1})}\Big) \odot I \\ &\quad + 2 C_t^{-1} B_t C_t^{-1} \odot C_t^{-1} \odot \phi'(\hat{b}_{t-1|t-1}) \phi'(\hat{b}_{t-1|t-1})^\top \,, \end{aligned}$$

in the diagonal setting, with $\odot$ the Hadamard (pointwise) product.

The upper-bound of the Kullback-Leibler divergence obtained thanks to the proposition above admits a closed-form minimum:

Proposition 6. *In the scalar and diagonal settings, for any t such that $f(\hat{b}_{t-1|t-1}) \succ 0$ and any $\hat{b}_{t|t}, \Sigma_{t|t}$,*

$$\begin{aligned} &KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; p(\cdot \mid \mathcal{F}_t)\Big) \\ &\quad \leq -\frac{1}{2}\log\det \Sigma_{t|t} + \frac{1}{2}\frac{\partial \psi_t}{\partial b_t}\Big|_{\hat{b}_{t-1|t-1}}^\top (\hat{b}_{t|t} - \hat{b}_{t-1|t-1}) \\ &\qquad + \frac{1}{2}\operatorname{Tr}\Big((\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^\top) \\ &\qquad\qquad \Big((\Sigma_{t-1|t-1} + \rho_b I)^{-1} + \frac{1}{2}H_t\Big)\Big) + c_b \,, \end{aligned}$$

where H_t is defined in Proposition 5 and c_b is a constant independent of $\hat{b}_{t|t}, \Sigma_{t|t}$. The minimum of the upper-bound detailed above is obtained with:

$$\Sigma_{t|t} = \left((\Sigma_{t-1|t-1} + \rho_b I)^{-1} + \frac{1}{2} H_t \right)^{-1}, \tag{8}$$

$$\hat{b}_{t|t} = \hat{b}_{t-1|t-1} - \frac{1}{2} \Sigma_{t|t} \frac{\partial \psi_t}{\partial b_t}\Big|_{\hat{b}_{t-1|t-1}}. \tag{9}$$

Similarly as (7) we can interpret (9) as a gradient step on ψ_t and we can remark that $\psi_t(\hat{b})$ is the following expected log-likelihood:

$$\mathbb{E}_{\theta_t \sim \mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})} [\log \mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, K P_{t-1|t-1} K^\top + f(\hat{b}))] .$$

Thus, except the exact recursive steps on $\hat{\theta}_{t|t}, P_{t|t}$ which are extensions of the Kalman filter steps, our recursive steps resemble Stochastic Gradient Variational Bayes (SGVB) algorithm steps as described in [14]. This novel class of algorithms is very popular for tuning complex deep learning networks, see for instance [15], [16]. There, the expectation of the log-likelihood is approximated by using Monte-Carlo simulation and only the first order of the gradient is used.

IV. Viking

We now introduce the algorithm following from the recursive updates described in the previous sections.

A. Definition of the Algorithm

Theorem 2 yields exact recursive updates for $\hat{\theta}_{t|t}, P_{t|t}$ but A_t^{-1} does not admit an explicit form. We propose to run Monte-Carlo estimation of A_t with very small samples ($n_{\mathrm{mc}} = 10$ draws by default). As the KL optimization is a coupled problem we solve it in a classical iterative fashion, that is, we repeat N times the updates alternately ($N = 2$ is a good default value). We summarize the procedure in Algorithm 1. We name it Viking (**V**ariational Bayes**i**an Variance Trac**king**).

B. Complexity

We decompose the number of operations of Viking in Table I. Although matrix multiplication and inversion have the same asymptotic complexity, in practice inversion is more costly.

We suggest the default $N = 2$ and $n_{\mathrm{mc}} = 10$, therefore the complexity of Viking is essentially driven by the complexity of matrix inversion. Consequently it is proportional to the one of methods relying on Kalman smoothing as in [11].

V. Experiments

We run several experiments, and we argue that our method behaves well for misspecified data. We begin with well-specified data generated under a state-space model with smoothly varying variances. Then we focus on misspecified data.

Algorithm 1: Viking at time step t

Time-invariant parameters: $\rho_a, \rho_b, n_{\mathrm{mc}}, f$.
Default: $\rho_a = e^{-9}, \rho_b = e^{-6}, n_{\mathrm{mc}} = 10, f(\cdot) = D_{\phi(\cdot)}$.
Inputs: $\hat{\theta}_{t-1|t-1}$, $P_{t-1|t-1}$, $\hat{a}_{t-1|t-1}$, $s_{t-1|t-1}$, $\hat{b}_{t-1|t-1}$, $\Sigma_{t-1|t-1}$, x_t, y_t.
Initialize:
Set $\hat{a}_{t|t}^{(0)} = \hat{a}_{t-1|t-1}$, $s_{t|t}^{(0)} = s_{t-1|t-1} + \rho_a$.
Set $\hat{b}_{t|t}^{(0)} = \hat{b}_{t-1|t-1}$, $\Sigma_{t|t}^{(0)} = \Sigma_{t-1|t-1} + \rho_b$.
Iterate: for $i = 1, \ldots, N$:

- 1. Set A_t then compute A_t^{-1} using (3) with Monte-Carlo from n_{mc} samples of $\mathcal{N}(\hat{b}_{t|t}^{(i-1)}, \Sigma_{t|t}^{(i-1)})$.
  2. Set $P_{t|t}^{(i)}, \hat{\theta}_{t|t}^{(i)}$ using (4) and (5), with A_t^{-1} from step 1 and $\hat{a}_{t|t}^{(i-1)}, s_{t|t}^{(i-1)}$.
- **If we learn σ_t^2:**
  3. Set $s_{t|t}^{(i)}$ using (6) with $\hat{\theta}_t^{(i)}, P_t^{(i)}, \hat{a}_{t|t}^{(i-1)}$.
  4. Set $\hat{a}_{t|t}^{(i)}$ using (7) with $\hat{\theta}_t^{(i)}, P_t^{(i)}, s_{t|t}^{(i)}$.
- **If we learn Q_t:**
  5. Set $\Sigma_{t|t}^{(i)}, \hat{b}_{t|t}^{(i)}$ using (8) and (9). Apply threshold $\hat{b}_{t|t}^{(i)} = \max(\hat{b}_{t|t}^{(i)}, 0)$.

Outputs: $\hat{\theta}_{t|t} = \hat{\theta}_{t|t}^{(N)}, P_{t|t} = P_{t|t}^{(N)}, \hat{a}_{t|t} = \hat{a}_{t|t}^{(N)}, s_{t|t} = s_{t|t}^{(N)}, \hat{b}_{t|t} = \hat{b}_{t|t}^{(N)}, \Sigma_{t|t} = \Sigma_{t|t}^{(N)}$.

TABLE I
Complexity of Algorithm 1.

Steps	Operations
1	$n_{\mathrm{mc}} S + (n_{\mathrm{mc}} + 1) I(d) + \mathcal{O}(M(d))$
2	$\mathcal{O}(d^2)$
3 and 4	$\mathcal{O}(d^2)$
5	$3I(d) + \mathcal{O}(M(d))$
Whole	$N\big(n_{\mathrm{mc}} S + (n_{\mathrm{mc}} + 4) I(d) + \mathcal{O}(M(d))\big)$

S denotes the complexity of gaussian draw, $M(d)$ and $I(d)$ denote the complexity of matrix multiplication and inversion.

A. Well-Specified Data with Unknown σ_t^2 and Known Q_t

We reproduce the experiment presented in [7] on the stochastic resonator model:

$$\theta_{t+1} - \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\omega \Delta t) & \frac{\sin(\omega \Delta t)}{\omega} \\ 0 & -\omega \sin(\omega \Delta t) & \cos(\omega \Delta t) \end{pmatrix} \theta_t \sim \mathcal{N}(0, Q) ,$$

$$y_t - (\theta_{t,1} + \theta_{t,2}) \sim \mathcal{N}(0, \sigma_t^2) ,$$

where we set $\omega = 0.05$ and $\Delta t = 0.1$ and the known covariance of the process noise is $Q = D_{(0.01,0,0.0001)}$. We display the variance trajectory for one simulation in Figure 1. Running the experiment 100 times we observe that both methods almost coincide in terms of root-mean-square error: 0.6859 for Viking and 0.6858 for VB-AKF [7]. In this comparison we take the best value of ρ_a for Viking as well as the best ρ for the VB-AKF in the list $e^{-i}, 1 \leq i \leq 10$.

B. Well-Specified Data with Unknown σ_t^2 and Q_t

We run a second simulation inspired by [11] in a well-specified setting. We generate $x_t \in [0, 1]^5$ using two possible alternatives:

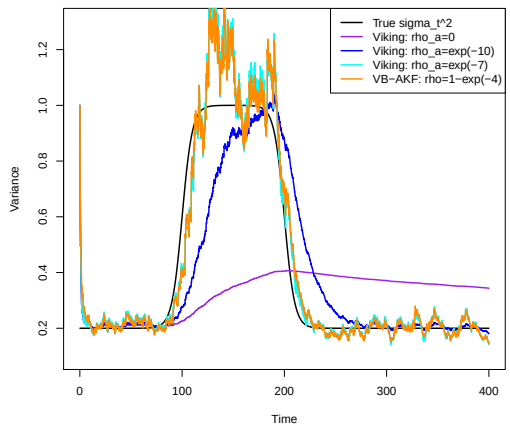

Fig. 1. Trajectory of the observation variance σ_t^2 estimated by our algorithm and compared to the estimate provided in [7].

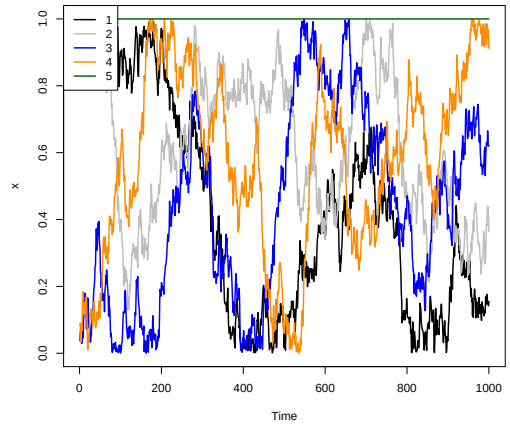

Fig. 2. Example of trajectory of the 5 components of the vector x_t considered in the setting *uniform non-iid*.

1) **Uniform i.i.d. design:** (x_t) is independent identically distributed. For each t, x_t is composed of 4 independent coefficients generated with uniform distributions on $[0,1]$ and one deterministic 1 coefficient.
2) **Uniform non-i.i.d. design:** (x_t) has the same distribution but is not i.i.d., a sample is displayed in Figure 2. Precisely x_1 is generated as before. Then for $j \in \{1,2,3,4\}$ and $t \geq 2$, we consider $z_{t,j} = x_{t-1,j} + \varepsilon_{t,j}$ where $\varepsilon_{t,j} \sim \mathcal{N}(0, 10^{-3})$ and we generate

$$x_{t,j} = \begin{cases} z_{t,j} \text{ if } 0 \leq z_{t,j} \leq 1\,, \\ \lceil z_{t,j} \rceil - z_{t,j} \text{ otherwise.} \end{cases}$$

Then we generate y_t by the following state-space model:

$$\begin{aligned} &\theta_0 \sim \mathcal{N}(0, I)\,, \\ &\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,, \\ &y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma_t^2)\,, \end{aligned}$$

where

$$\begin{aligned} \sigma_t^2 &= 1 + 0.1 \cos \frac{4\pi t}{n}\,, \\ Q_t &= \left(0.25 + 0.2 \cos \frac{4\pi t}{n}\right) D_{(0,0,1,1,1)}\,. \end{aligned}$$

The simulation time is $n = 10^3$. In Figure 3 we compare Viking to the slide window variational adaptive Kalman filter (SWVAKF) introduced in [11], which we tune in several ways. First we increase the window length from 5 to 20, resulting in a significant improvement at the cost of more computations. Second we tune the forgetting factor, and to play fair with Viking we define different forgetting factors for the estimation of σ_t^2 and the estimation of Q_t. We select the best *a posteriori*, and we do the same for Viking. Third, we enforce diagonal and scalar variants of the SWVAKF: the diagonal variant is defined by replacing by 0 each non-diagonal coefficient after each update, and on top of that in the scalar variant we replace each diagonal coefficient by the averaged diagonal.

C. Misspecified Data with Unknown σ_t^2 and Q_t

To experiment misspecification we consider a state-space model with two states evolving independently with identical processes, and the observation is generated using one of them uniformly at random. That is summarized by the following set of equations:

$$\begin{aligned} &\theta_0^{(i)} \sim \mathcal{N}(0, I)\,, && i \in \{0,1\}\,, \\ &\theta_t^{(i)} - 0.9\,\theta_t^{(i)} \sim \mathcal{N}(0, Q_t)\,, && i \in \{0,1\}\,, \\ &i_t \sim \mathcal{B}(1/2)\,, \\ &y_t - \theta_t^{(i_t)\top} x_t \sim \mathcal{N}(0, \sigma_t^2)\,, \end{aligned}$$

where we assume all gaussian noises to be independent of each other and of (i_t). We consider the same settings for x_t as well as the same variances σ_t^2, Q_t defined in Section V-B.

The contraction (here by a coefficient 0.9) is necessary to have the convergence of the distribution of y_t as well as of the conditional distribution of y_t given the filtration $\mathcal{F}_{t-1}$. In the tracking mode (no contraction) the variance of the conditional distribution would diverge to ∞, and therefore the error of any forecasting strategy would also diverge to ∞.

We refer to Figure 3 for the evaluation in mean squared error. We observe that Viking in the diagonal setting behaves poorly compared to the SWVAKF for well-specified data with i.i.d. design but better in the other 3 experiments. As mentioned in Section II-D we believe it is natural that a consistent adaptive Kalman filter should be closer to the true Kalman filter than our algorithm which cannot be written using Kalman recursion. However the careful property (see the design of f in Section II-E) allows us to outperform existing methods for misspecified data. This interpretation of the observation generation may to a minor extent be transposed to the design generation. Indeed, in our non-i.i.d. design a shift in the data should be harder to attribute to one coefficient of the state, and therefore it should be harder to learn the variances, that is why the difference between the two Kalman filters with constant variances is smaller. Thus the model should not be trusted too much.

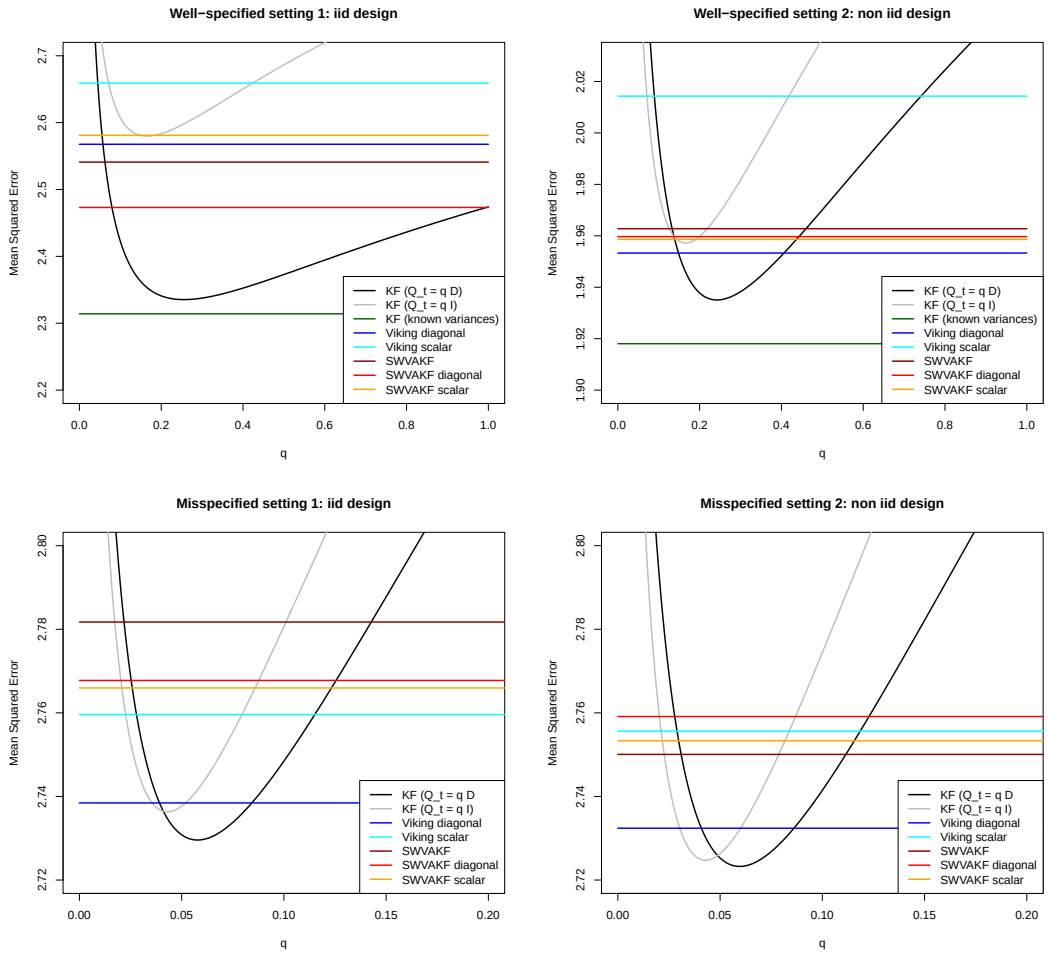

Fig. 3. Mean Squared Errors in the four settings introduced in Sections V-B and V-C: i.i.d. (left) or non-i.i.d. (right) design, well-specified (top) or misspecified (bottom). We compare Viking to the SWVAKF of [11] in the scalar and diagonal settings. For Viking we set $n_{\mathrm{mc}} = 10$. The oracles to which we compare are the Kalman filter with known variances when they exist (well-specified settings) and two Kalman filters with constant variances: the state noise covariance is either $Q = q \cdot D_{(0,0,1,1,1)}$ or $Q = q \cdot I$ and in both we set the space noise variance to $\sigma^2 = 1$. We evaluate through the mean squared error on the second half of the experiment in order to not depend too much on the initialization (even if we have same initial expected variances for Viking and SWVAKF).

D. Impact of n_{mc}

The number of Monte-Carlo samples used at each step to compute A_t^{-1} is a crucial factor of the complexity of Viking. It is therefore necessary to evaluate its impact on the performance in order to reach the best compromise between forecasting and computational efficiencies. We refer to Figure 4 for an evaluation of the error with different values of n_{mc}. The default value $n_{\mathrm{mc}} = 10$ seems reasonable.

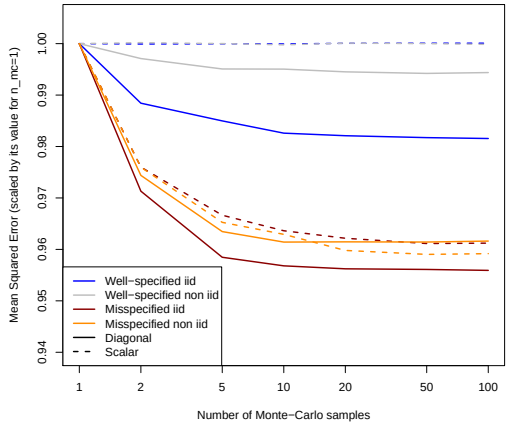

Fig. 4. Mean Squared Error of Viking as a function of n_{mc}. We scale by the mean squared error of the algorithm with $n_{\mathrm{mc}} = 1$ in order to fit the different algorithms (diagonal and scalar settings) as well as the different experiments (i.i.d. or non-i.i.d. design, well-specified or misspecified) in the same graph.

VI. Conclusion

We have introduced Viking, an algorithm for adaptive time series forecasting relying on state-space models with unknown state and space variances. We derived an augmented latent model, and we apply variational Bayes for the inference. We extend the Kalman filter to uncertain environment. For the additional latent variables, we use approximative steps close to SGVB recursive ones. The prediction performances are better than the state of the art in misspecified settings at the same computational cost.

The choice of the function applied to the latent variable to obtain the state noise covariance matrix is a perspective of future research. We provide a specific choice leading to promising experimental results on simulations in both well-specified and misspecified settings. However we wrote most of the article considering this function is a parameter of Viking, because we believe other functions may be of interest.

APPENDIX

We provide the proofs for all the claims of the article.

Proof of Lemma 1. We start from the expression of (1) that we can decompose as follows:

$$\begin{aligned}
&KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; p(\cdot \mid \mathcal{F}_t)\Big) \\
&\quad = \mathbb{E}_{\theta_t \sim \mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})}[\log \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})] \\
&\qquad + \mathbb{E}_{a_t \sim \mathcal{N}(\hat{a}_{t|t}, s_{t|t})}[\log \mathcal{N}(a_t \mid \hat{a}_{t|t}, s_{t|t})] \\
&\qquad + \mathbb{E}_{b_t \sim \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[\log \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})] \\
&\qquad - \mathbb{E}_{(\theta_t, a_t, b_t) \sim \mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})} \\
&\qquad\quad [\log p(\theta_t, a_t, b_t \mid \mathcal{F}_t)] \,.
\end{aligned}$$

The last term can be split using the factorized form of (2). We observe that on the one hand,

$$\begin{aligned}
&\mathbb{E}_{(\theta_t, a_t, b_t) \sim \mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})} \\
&\qquad [\log \mathcal{N}(y_t \mid \theta_t^\top x_t, \exp(a_t))] \\
&\quad = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\hat{a}_{t|t} \\
&\qquad - \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)\exp(-\hat{a}_{t|t} + \frac{1}{2}s_{t|t}) \,,
\end{aligned}$$

and on the other hand,

$$\begin{aligned}
&\mathbb{E}_{(\theta_t, a_t, b_t) \sim \mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})} \\
&\qquad [\log \mathcal{N}(\theta_t \mid K\hat{\theta}_{t-1|t-1}, KP_{t-1|t-1}K^\top + f(b_t))] \\
&\quad = -\frac{d\log(2\pi)}{2} - \frac{1}{2}\mathbb{E}_{b_t \sim \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[\psi_t(b_t)] \,,
\end{aligned}$$

where ψ_t is defined in the lemma. Combining the last equations with the value of the entropy of gaussian random variables yields the result. □

Proof of Theorem 2. Thanks to Lemma 1 we have

$$\begin{aligned}
&KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; p(\cdot \mid \mathcal{F}_t)\Big) \\
&\quad = \frac{1}{2}\operatorname{Tr}\Big((P_{t|t} + (\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})^\top)A_t\Big) \\
&\qquad + \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)\exp(-\hat{a}_{t|t} + \frac{1}{2}s_{t|t}) \\
&\qquad - \frac{1}{2}\log\det P_{t|t} + c_\theta \,,
\end{aligned}$$

where c_θ is a constant independent of $\hat{\theta}_{t|t}, P_{t|t}$, and A_t is defined in the theorem. To conclude we write the first order conditions:

$$\begin{aligned}
&-\frac{1}{2}P_{t|t}^{-1} + \frac{1}{2}\Big(A_t + \frac{x_t x_t^\top}{\exp(\hat{a}_{t|t} - \frac{1}{2}s_{t|t})}\Big) = 0 \,, \\
&-\frac{(y_t - \hat{\theta}_{t|t}^\top x_t)x_t}{\exp(\hat{a}_{t|t} - \frac{1}{2}s_{t|t})} + A_t(\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1}) = 0 \,.
\end{aligned}$$

□

Proof of Proposition 3. Thanks to Lemma 1, we have

$$\begin{aligned}
&KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; p(\cdot \mid \mathcal{F}_t)\Big) \\
&\quad = \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)e^{-\hat{a}_{t|t} + s_{t|t}/2} \\
&\qquad + \frac{1}{2}(s_{t-1|t-1} + \rho_a)^{-1}s_{t|t} - \frac{1}{2}\log(s_{t|t}) + c_s \,,
\end{aligned}$$

where c_s is a constant independent of $s_{t|t}$. Moreover, if $0 \leq s_{t|t} \leq s_{t-1|t-1} + \rho_a$ then

$$e^{s_{t|t}/2} \leq e^{(s_{t-1|t-1} + \rho_a)/2} + \frac{1}{2}(s_{t|t} - (s_{t-1|t-1} + \rho_a)) \,.$$

The last two equations yield the upper-bound of the proposition. To obtain (6) we write the first order condition of optimality:

$$\begin{aligned}
&\frac{1}{4}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)e^{-\hat{a}_{t|t}} - \frac{1}{2}s_{t|t}^{-1} \\
&\qquad + \frac{1}{2}(s_{t-1|t-1} + \rho_a)^{-1} = 0 \,.
\end{aligned}$$

□

Proof of Proposition 4. Thanks to Lemma 1 we have

$$\begin{aligned}
&KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; p(\cdot \mid \mathcal{F}_t)\Big) \\
&\quad \leq \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)e^{-\hat{a}_{t|t} + s_{t|t}/2} \\
&\qquad + \frac{1}{2}(s_{t-1|t-1} + \rho_a)^{-1}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1})^2 + \frac{1}{2}\hat{a}_{t|t} + c_a \,,
\end{aligned}$$

with c_a a constant independent of $\hat{a}_{t|t}$. Moreover, if $\hat{a}_{t|t} \in [\hat{a}_{t-1|t-1} - M_a, \hat{a}_{t-1|t-1} + M_a]$ we have the following upper-bound:

$$\begin{aligned}
e^{-\hat{a}_{t|t}} \leq e^{-\hat{a}_{t-1|t-1}}\Big(&1 - (\hat{a}_{t|t} - \hat{a}_{t-1|t-1}) \\
&+ \frac{e^{M_a}}{2}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1})^2\Big) \,.
\end{aligned}$$

The last two equations yield the upper-bound of the proposition. To obtain (7) we write the first-order condition:

$$\begin{aligned}
&\frac{1}{s_{t-1|t-1} + \rho_a}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1}) + \frac{1}{2} \\
&\qquad + \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)e^{-\hat{a}_{t-1|t-1} + s_{t|t}/2} \\
&\qquad\quad \Big(-1 + e^{M_a}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1})\Big) = 0 \,,
\end{aligned}$$

□

To prove Propositions 5 and 6 we first compute the first and second derivatives of ψ_t for the scalar and diagonal settings:

Lemma 7. *Let $C_t = KP_{t-1|t-1}K^\top + f(b_t)$ and $B_t = P_{t|t} + (\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - K\hat{\theta}_{t-1|t-1})^\top$.*

- *If $f(\cdot) = \phi(\cdot)I$ then for any b_t, we have*

$$\begin{aligned}
\psi_t'(b_t) &= \operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1}))\phi'(b_t) \,, \\
\psi_t''(b_t) &= \operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1}))\phi''(b_t) \\
&\qquad + 2\operatorname{Tr}(C_t^{-2}(B_t C_t^{-1} - I/2))\phi'(b_t)^2 \,.
\end{aligned}$$

- *If $f(\cdot) = D_{\phi(\cdot)}$ then for any b_t, we have*

$$\frac{\partial \psi_t}{\partial b_t} = \Delta_{C_t^{-1}(I - B_t C_t^{-1})} \odot \phi'(b_t)\,,$$
$$\frac{\partial^2 \psi_t}{\partial b_t^2} = C_t^{-1}(I - B_t C_t^{-1}) D_{\phi''(b_t)} \odot I + 2C_t^{-1}(B_t C_t^{-1} - I/2) \odot C_t^{-1} \odot \phi'(b_t)\phi'(b_t)^\top\,,$$

where $\odot$ is the Hadamard (pointwise) product.

Proof. • In the scalar setting we recall that

$$\psi_t(b) = \log\det(KP_{t-1|t-1}K^\top + \phi(b)I) + \operatorname{Tr}(B_t(KP_{t-1|t-1}K^\top + \phi(b)I)^{-1})\,.$$

We denote by $\log$ and $\exp$ the univariate logarithm and exponential and by Log the matrix logarithm. Note that if $A \succ 0$, it holds $\det A = \exp\operatorname{Tr}(\operatorname{Log} A)$. We define $C_t = KP_{t-1|t-1}K^\top + \phi(b_t)I$ and we obtain:

$$\begin{aligned}
&\log\det(KP_{t-1|t-1}K^\top + \phi(b)I) - \operatorname{Tr}\operatorname{Log}(C_t)\\
&\quad= \operatorname{Tr}\operatorname{Log}(KP_{t-1|t-1}K^\top + \phi(b)I) - \operatorname{Tr}\operatorname{Log}(C_t)\\
&\quad= \operatorname{Tr}\operatorname{Log}\left(I + (\phi(b) - \phi(b_t))C_t^{-1}\right)\\
&\quad= \operatorname{Tr}\Big(\Big(\phi'(b_t)(b-b_t) + \frac{1}{2}\phi''(b_t)(b-b_t)^2\Big)C_t^{-1}\\
&\qquad - \frac{1}{2}\big(\phi'(b_t)(b-b_t)C_t^{-1}\big)^2 + o((b-b_t)^2)\Big)\,.
\end{aligned}$$

The last line follows from the series expansion of the Logarithm. We apply another series expansion for the second term of ψ_t: we have

$$\begin{aligned}
&\operatorname{Tr}(B_t(KP_{t-1|t-1}K^\top + \phi(b)I)^{-1})\\
&\quad= \operatorname{Tr}\left(B_t C_t^{-1}\left(I + (\phi(b) - \phi(b_t))C_t^{-1}\right)^{-1}\right)\\
&\quad= \operatorname{Tr}\Big(B_t C_t^{-1}\\
&\qquad\Big(I - \Big(\phi'(b_t)(b-b_t) + \frac{1}{2}\phi''(b_t)(b-b_t)^2\Big)C_t^{-1}\\
&\qquad + \big(\phi'(b_t)(b-b_t)C_t^{-1}\big)^2 + o((b-b_t)^2)\Big)\Big)\,.
\end{aligned}$$

Summing the last two equations, and using the identity $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$, we can identify the first and second derivatives of ψ_t.

- We develop a similar argument in the diagonal setting:

$$\psi_t(b) = \log\det(KP_{t-1|t-1}K^\top + D_{\phi(b)}) + \operatorname{Tr}(B_t(KP_{t-1|t-1}K^\top + D_{\phi(b)})^{-1})\,,$$

then we apply the series expansion of the Logarithm:

$$\begin{aligned}
&\log\det(KP_{t-1|t-1}K^\top + D_{\phi(b)}) - \operatorname{Tr}\operatorname{Log}(C_t)\\
&\quad= \operatorname{Tr}\operatorname{Log}\left(I + D_{\phi(b)-\phi(b_t)}C_t^{-1}\right)\\
&\quad= \operatorname{Tr}\Big(D_{\phi'(b_t)(b-b_t)+\frac{1}{2}\phi''(b_t)(b-b_t)^2}C_t^{-1}\\
&\qquad - \frac{1}{2}(D_{\phi'(b_t)(b-b_t)}C_t^{-1})^2 + o(\|b-b_t\|^2)\Big)\,,
\end{aligned}$$

where $C_t = KP_{t-1|t-1}K^\top + D_{\phi(b_t)}$ and $\phi'(b_t), \phi''(b_t)$ denote the coefficient-wise application of the first and second derivatives of ϕ to the vector b_t. We apply another series expansion for the second term of ψ_t:

$$\begin{aligned}
&\operatorname{Tr}\left(B_t(KP_{t-1|t-1}K^\top + D_{\phi(b)})^{-1}\right)\\
&\quad= \operatorname{Tr}\left(B_t C_t^{-1}\left(I + D_{\phi(b)-\phi(b_t)}C_t^{-1}\right)^{-1}\right)\\
&\quad= \operatorname{Tr}\Big(B_t C_t^{-1}\\
&\qquad\Big(I - D_{\phi'(b_t)(b-b_t)+\frac{1}{2}\phi''(b_t)(b-b_t)^2}C_t^{-1}\\
&\qquad + (D_{\phi'(b_t)(b-b_t)}C_t^{-1})^2 + o(\|b-b_t\|^2)\Big)\Big)\,.
\end{aligned}$$

Summing the last two equations we obtain

$$\begin{aligned}
\psi_t(b) &= \operatorname{Tr}\operatorname{Log}(C_t) + \operatorname{Tr}(B_t C_t^{-1})\\
&\quad + \operatorname{Tr}\Big(C_t^{-1}(I - B_t C_t^{-1})\\
&\qquad D_{\phi'(b_t)(b-b_t)+\frac{1}{2}\phi''(b_t)(b-b_t)^2}\Big)\\
&\quad + \operatorname{Tr}\Big(C_t^{-1}(B_t C_t^{-1} - I/2)\\
&\qquad D_{\phi'(b_t)(b-b_t)}C_t^{-1}D_{\phi'(b_t)(b-b_t)}\Big)\\
&\quad + o(\|b-b_t\|^2)\,.
\end{aligned}$$

Then we use the identity $Tr(AD_vBD_v) = v^\top(A \odot B^\top)v$. We have

$$\begin{aligned}
\psi_t(b) &= \operatorname{Tr}\operatorname{Log}(C_t) + \operatorname{Tr}(B_t C_t^{-1})\\
&\quad + \frac{1}{2}(b-b_t)^\top\Big(C_t^{-1}(I - B_t C_t^{-1})D_{\phi''(b_t)} \odot I\\
&\qquad + 2C_t^{-1}(B_t C_t^{-1} - I/2) \odot\\
&\qquad C_t^{-1} \odot \phi'(b_t)\phi'(b_t)^\top\Big)(b-b_t)\\
&\quad + (\Delta_{C_t^{-1}(I - B_t C_t^{-1})} \odot \phi'(b_t))^\top(b-b_t) + o(\|b-b_t\|^2)\,.
\end{aligned}$$

Thus we can identify the first and second derivatives of ψ_t.

□

Proof of Proposition 5. As long as $f(\hat{b}_{t-1|t-1}) \succ 0$ we know that f is twice differentiable in $\hat{b}_{t-1|t-1}$ and the local upper-bound property of Proposition 5 holds if $\frac{\partial^2 \psi_t}{\partial b_t^2}|_{\hat{b}_{t-1|t-1}} \prec H_t$. We bound the expressions obtained in Lemma 7.

- In the scalar setting,

$$\psi_t''(\hat{b}_{t-1|t-1}) = \operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1}))\phi''(\hat{b}_{t-1|t-1}) + 2\operatorname{Tr}(C_t^{-2}(B_t C_t^{-1} - I/2))\phi'(\hat{b}_{t-1|t-1})^2\,.$$

Furthermore, $C_t \succ 0$ thus $C_t^{-1} \succ 0$, $\operatorname{Tr}(C_t^{-1}) > 0$, and $\operatorname{Tr}(C_t^{-2}) > 0$. $\phi''(\hat{b}_{t-1|t-1}) = -1/(1+\hat{b}_{t-1|t-1})^2 < 0$ and $\phi'(\hat{b}_{t-1|t-1})^2 > 0$, therefore we obtain

$$\psi_t''(\hat{b}_{t-1|t-1}) < -\operatorname{Tr}(C_t^{-1}B_t C_t^{-1})\phi''(\hat{b}_{t-1|t-1}) + 2\operatorname{Tr}(C_t^{-2}B_t C_t^{-1})\phi'(\hat{b}_{t-1|t-1})^2\,.$$

- In the diagonal setting,

$$\frac{\partial^2 \psi_t}{\partial b_t^2}\Big|_{\hat{b}_{t-1|t-1}} = C_t^{-1}(I - B_t C_t^{-1})D_{\phi''(\hat{b}_{t-1|t-1})} \odot I + 2C_t^{-1}(B_t C_t^{-1} - I/2) \odot C_t^{-1} \odot \phi'(\hat{b}_{t-1|t-1})\phi'(\hat{b}_{t-1|t-1})^\top\,.$$

Similarly we have $C_t^{-1} \succ 0, D_{\phi''(\hat{b}_{t-1|t-1})} \prec 0$ and as diagonal coefficients of C_t^{-1} are positive, it yields $(C_t^{-1} D_{\phi''(\hat{b}_{t-1|t-1})}) \odot I \prec 0$.
Moreover $\phi'(\hat{b}_{t-1|t-1})\phi'(\hat{b}_{t-1|t-1})^\top \succ 0$, and we can apply Schur product theorem: $C_t^{-1} \odot C_t^{-1} \odot \phi'(\hat{b}_{t-1|t-1})\phi'(\hat{b}_{t-1|t-1})^\top \succ 0$. Eventually:

$$\frac{\partial^2 \psi_t}{\partial b_t^2}\Big|_{\hat{b}_{t-1|t-1}} \prec -C_t^{-1} B_t C_t^{-1} D_{\phi''(\hat{b}_{t-1|t-1})} \odot I + 2C_t^{-1} B_t C_t^{-1} \odot C_t^{-1} \odot \phi'(\hat{b}_{t-1|t-1})\phi'(\hat{b}_{t-1|t-1})^\top .$$

□

Proof of Proposition 6. Thanks to Lemma 1 we have:

$$\begin{aligned} &KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; p(\cdot \mid \mathcal{F}_t)\Big) \\ &= -\frac{1}{2}\log\det\Sigma_{t|t} + \frac{1}{2}\mathbb{E}_{b_t \sim \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[\psi_t(b_t)] \\ &\quad + \frac{1}{2}\operatorname{Tr}\Big((\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^\top) \\ &\qquad\qquad (\Sigma_{t-1|t-1} + \rho_b I)^{-1}\Big) + c_b \,, \end{aligned}$$

where c_b is a constant independent of $\hat{b}_{t|t}, \Sigma_{t|t}$. Combining the last equation and Proposition 5, then using the first two moments of the gaussian distribution we obtain:

$$\begin{aligned} &KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; p(\cdot \mid \mathcal{F}_t)\Big) \\ &\leq -\frac{1}{2}\log\det\Sigma_{t|t} + \frac{1}{2}\psi_t(\hat{b}_{t-1|t-1}) \\ &\quad + \frac{1}{2}\frac{\partial \psi_t}{\partial b_t}\Big|^\top_{\hat{b}_{t-1|t-1}} (\hat{b}_{t|t} - \hat{b}_{t-1|t-1}) \\ &\quad + \frac{1}{4}\operatorname{Tr}\Big(H_t(\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^\top)\Big) \\ &\quad + \frac{1}{2}\operatorname{Tr}\Big((\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^\top) \\ &\qquad\qquad (\Sigma_{t-1|t-1} + \rho_b I)^{-1}\Big) + c_b \,. \end{aligned}$$

This yields the upper-bound of Proposition 6. The recursive updates follow from the first order conditions:

$$\begin{aligned} &-\frac{1}{2}\Sigma_{t|t}^{-1} + \frac{1}{2}\Big((\Sigma_{t-1|t-1} + \rho_b I)^{-1} + \frac{1}{2}H_t\Big) = 0 \,, \\ &\Big((\Sigma_{t-1|t-1} + \rho_b I)^{-1} + \frac{1}{2}H_t\Big)(\hat{b}_{t|t} - \hat{b}_{t-1|t-1}) \\ &\qquad + \frac{1}{2}\frac{\partial \psi_t}{\partial b_t}\Big|_{\hat{b}_{t-1|t-1}} = 0 \,. \end{aligned}$$

□

REFERENCES

[1] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Journal of basic engineering*, vol. 83, no. 1, pp. 95–108, 1961.

[2] P. J. Brockwell, R. A. Davis, and S. E. Fienberg, *Time series: theory and methods: theory and methods*. Springer Science & Business Media, 1991.

[3] J. Durbin and S. J. Koopman, *Time series analysis by state space methods*. Oxford university press, 2012.

[4] R. Mehra, "Approaches to adaptive filtering," *IEEE Transactions on Automatic Control*, vol. 17, no. 5, pp. 693–698, 1972.

[5] M. J. Beal, *Variational algorithms for approximate Bayesian inference*. University of London, University College London (United Kingdom), 2003.

[6] V. Šmídl and A. Quinn, *The variational Bayes method in signal processing*. Springer Science & Business Media, 2006.

[7] S. Sarkka and A. Nummenmaa, "Recursive noise adaptive kalman filtering by variational bayesian approximations," *IEEE Transactions on Automatic Control*, vol. 54, no. 3, pp. 596–600, 2009.

[8] G. Agamennoni, J. I. Nieto, and E. M. Nebot, "Approximate inference in state-space models with heavy-tailed noise," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5024–5037, 2012.

[9] S. Särkkä and J. Hartikainen, "Non-linear noise adaptive kalman filtering via variational bayes," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2013, pp. 1–6.

[10] Y. Huang, Y. Zhang, Z. Wu, N. Li, and J. Chambers, "A novel adaptive kalman filter with inaccurate process and measurement noise covariance matrices," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 594–601, 2017.

[11] Y. Huang, F. Zhu, G. Jia, and Y. Zhang, "A slide window variational adaptive kalman filter," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 12, pp. 3552–3556, 2020.

[12] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for bayesian inference," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, 2008.

[13] A. Tyagi and J. W. Davis, "A recursive filter for linear systems on riemannian manifolds," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[14] D. A. Knowles, "Stochastic gradient variational bayes for gamma approximating distributions," *arXiv preprint arXiv:1509.01631*, 2015.

[15] D. P. Kingma and M. Welling, "Stochastic gradient vb and the variational auto-encoder," in *Second International Conference on Learning Representations, ICLR*, vol. 19, 2014, p. 121.

[16] A. Tjandra, S. Sakti, S. Nakamura, and M. Adriani, "Stochastic gradient variational bayes for deep learning-based asr," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 175–180.