Recursive Estimation of State-Space Noise Covariance Matrix by Approximate Variational Bayes

Joseph de Vilmarest, Olivier Wintenberger

Recursive Estimation of State-Space Noise Covariance Matrix by Approximate Variational Bayes

Joseph de Vilmarest[1] and Olivier Wintenberger[2]

April 15, 2021

Abstract

This working paper considers state-space models where the variance of the observation is known but the covariance matrix of the state process is unknown and potentially time-varying. We propose an adaptive algorithm to estimate jointly the state and the covariance matrix of the state process, relying on Variational Bayes and second-order Taylor approximations.

1 Introduction

Linear state-space models have been widely used to model observations as gaussian distributions whose mean is a time-varying linear function of covariates. The linear parameter is a latent variable called state, and the parameters of the state-space model are the variance of the observation and the variance of the state process noise. When these variances are known, the recursive estimation is realized by Kalman filters (Kalman and Bucy, 1961).

However the observation and state noise variances are unknown in most practical applications. A wide literature has emerged for tuning these hyper-parameters. The estimation of unknown fixed variances on a historical dataset is generally realized maximizing the likelihood (Brockwell et al., 1991; Durbin and Koopman, 2012). Another approach estimates these hyper-parameters (fixed or not) in an online fashion. Adaptive filtering methods have been described by Mehra (1972).

Recently, online Variational Bayesian (VB) methods as introduced by Šmídl and Quinn (2006) have gathered attention in the Kalman filtering community. We recall that the objective is the online estimation of potentially time-variant parameters. The difference with 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 factored distributions. The best factored distribution is defined as the one minimizing its Kullback-Leibler divergence with the posterior.

Sarkka and Nummenmaa (2009) apply a VB approach to estimate the observation noise covariance matrix in a Kalman filter. 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 some sort of forgetting factor, multiplying the variances of the inverse gamma posterior distributions by a constant. Huang et al. (2017) extend this method with an inverse Wishart prior. At the same time they apply the VB approach to correct the covariance matrix of the state after applying Kalman recursions with an inaccurate state noise covariance. The inverse Wishart distribution appears as a nice conjugate prior to generalize the inverse gamma distribution. In another adaptive Kalman filter they propose to estimate both the observation and state noise covariance matrices (Huang et al., 2020). Their 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 covariance matrices is introduced via a forgetting factor.

In this working paper we present a new approach to estimate recursively the state noise covariance matrix relying on the VB approach. Instead of using a forgetting factor to impose a dynamical estimation, we consider a random walk model on the covariance matrix. As there doesn't exist a conjugate prior distribution on the state noise covariance, we rely on several second-order Taylor approximations to estimate the Kullback-Leibler divergence.

[1]Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, CNRS and Électricité de France R&D
[2]Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, CNRS

2 Bayesian formulation

We focus on the following state-space model:

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

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ and $\eta_t \sim \mathcal{N}(0, Q_t)$ are the observation and process noises, and the state follows the initial distribution $\theta_0 \sim \mathcal{N}(\hat{\theta}_0, P_0)$. In the case where 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 (Kalman and Bucy, 1961). We focus on a particular setting where $\sigma_t^2 = \sigma^2$ is known, but the covariance matrix Q_t is unknown and needs to be estimated jointly with the state θ_t.

2.1 Dynamical model

A way to introduce a dynamical estimation of Q_t is to treat it as another latent variable. **ref** Gaussian distributions behave well to capture the latent information of a nonlinear state-space model, and they are appealing due to their nice random walk interpretation to model a dynamic latent variable. Similarly we choose to use a gaussian prior for the covariance matrix Q_t. However a variance is necessarily nonnegative, thus we introduce a known transform g such that $Q_t = g(b_t)$, where b_t is a normal distribution. Our dynamical model can be summarized as follows:

$$\theta_0 \sim \mathcal{N}(\hat{\theta}_0, P_0)\,, \qquad b_0 \sim \mathcal{N}(\hat{b}_0, \Sigma_0)\,,$$
$$\theta_t - \theta_{t-1} \sim \mathcal{N}(0, g(b_t))\,, \qquad b_t - b_{t-1} \sim \mathcal{N}(0, r_b I)\,,$$
$$y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma^2)\,.$$

In these equations we implicitly assume the independence of the process noises on the state and on the variance, in particular we have

$$p(\theta_t, b_t \mid \theta_{t-1}, b_{t-1}) = p(\theta_t \mid \theta_{t-1}, b_t) p(b_t \mid b_{t-1})\,. \tag{1}$$

2.2 Bayesian approach

We apply a bayesian approach in order to estimate jointly the state θ_t and the latent vector b_t given the past observations. Formally we introduce the filtration of the past observations $\mathcal{F}_t = \sigma(x_1, y_1, \ldots, x_t, y_t)$. At each iteration t, the bayesian approach consists in a prediction step and a filtering step where we use Bayes' rule:

Prediction: $$p(\theta_t, b_t \mid \mathcal{F}_{t-1}) = \int p(\theta_t \mid \theta_{t-1}, b_t) p(b_t \mid b_{t-1}) p(\theta_{t-1}, b_{t-1} \mid \mathcal{F}_{t-1}) d\theta_{t-1} db_{t-1}\,,$$

Filtering: $$p(\theta_t, b_t \mid \mathcal{F}_t) = p(y_t \mid x_t, \theta_t, b_t) p(\theta_t, b_t \mid \mathcal{F}_{t-1}) \frac{p(x_t, \mathcal{F}_{t-1})}{p(\mathcal{F}_t)}\,.$$

The prediction equation uses Equation (1). The posterior distribution $p(. \mid \mathcal{F}_t)$ is analytically intractable. The objective is to estimate the first and second moments of its marginals in θ_t and b_t, namely $\hat{\theta}_{t|t}, P_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$.

2.3 Variational Bayesian approach

A standard approach, referred to as VB, is to approximate recursively the posterior distribution with a factorized distribution where each component is of a simple form (Šmídl and Quinn, 2006). We look for $\hat{\theta}_{t|t}, P_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$ such that the product of gaussian distributions $\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})$ is the best approximation of the posterior distribution. The best approximation is in the sense of the minimum of the Kullback-Leibler (KL) divergence.

In what follows we use the notation $\mathcal{N}(x \mid \mu, \Sigma)$ for the probability density function at point x of the distribution $\mathcal{N}(\mu, \Sigma)$. Our aim is to minimize the following criterion in $\hat{\theta}_{t|t}, P_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$:

$$KL\Big(\mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \,||\, p(\theta_t, b_t \mid \mathcal{F}_t)\Big)$$
$$= \int \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \log \frac{\mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})}{p(\theta_t, b_t \mid \mathcal{F}_t)} d\theta_t db_t\,. \tag{2}$$

Propagating the factorized approximation

$$p(\theta_{t-1}, b_{t-1} \mid \mathcal{F}_{t-1}) \approx \mathcal{N}(\theta_{t-1} \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1})\mathcal{N}(b_{t-1} \mid \hat{b}_{t-1|t-1}, \Sigma_{t-1|t-1}) \,,$$

the prediction step becomes:

$$\begin{aligned} p(\theta_t, b_t \mid \mathcal{F}_{t-1}) &\approx \int \mathcal{N}(\theta_t - \theta_{t-1} \mid 0, g(b_t))\mathcal{N}(b_t - b_{t-1} \mid 0, r_b I) \\ &\qquad \mathcal{N}(\theta_{t-1} \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1})\mathcal{N}(b_{t-1} \mid \hat{b}_{t-1|t-1}, \Sigma_{t-1|t-1}) d\theta_{t-1} db_{t-1} \\ &\approx \mathcal{N}(\theta_t \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1} + g(b_t))\mathcal{N}(b_t \mid \hat{b}_{t-1|t-1}, \Sigma_{t-1|t-1} + r_b I) \,. \end{aligned}$$

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

$$p(\theta_t, b_t \mid \mathcal{F}_t) = \mathcal{N}(y_t \mid \theta_t^\top x_t, \sigma^2)\mathcal{N}(\theta_t \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1} + g(b_t))\mathcal{N}(b_t \mid \hat{b}_{t-1|t-1}, \Sigma_{t-1|t-1} + r_b I)\frac{p(x_t, \mathcal{F}_{t-1})}{p(\mathcal{F}_t)} \,. \tag{3}$$

This last equation defines the posterior distribution that we plug in Equation 2 to obtain the optimization problem that we would like to solve recursively. At each step, the VB approach yields a coupled optimization problem in the parameters of the two gaussian distributions.

2.4 The iterative optimization solution to the VB problem

The classical iterative method (see for instance Tzikas et al. (2008)) consists in computing alternately $\exp(\mathbb{E}[\log p(\theta_t, b_t \mid \mathcal{F}_t)])$ where the expected value is taken with respect to one of the two latent variables, and identifying the desired first and second moments with respect to the other latent variable. We compute the VB iterative step with respect to $\hat{\theta}_{t|t}, P_{t|t}$:

Theorem 1. *Given $\hat{b}_{t|t}, \Sigma_{t|t}$, the values of $\hat{\theta}_{t|t}, P_{t|t}$ minimizing the KL divergence are given by*

$$\begin{aligned} P^\star_{t|t} &= A_t^{-1} - \frac{A_t^{-1} x_t x_t^\top A_t^{-1}}{x_t^\top A_t^{-1} x_t + \sigma^2} \\ \hat{\theta}^\star_{t|t} &= \hat{\theta}_{t-1} + \frac{A_t^{-1} x_t}{x_t^\top A_t^{-1} x_t + \sigma^2}(y_t - x_t^\top \hat{\theta}_{t-1|t-1}) \,. \end{aligned}$$

with $A_t = \int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})(P_{t-1|t-1} + g(b_t))^{-1} db_t$.

Note that the updates defined above are the ones of the Kalman filter with a known variance Q_t, where we have replaced $P_{t-1|t-1} + Q_t$ with A_t^{-1}.

However the expression $\exp(\mathbb{E}_{\theta_t}[\log p(\theta_t, b_t \mid \mathcal{F}_t)])$ doesn't match a gaussian distribution in b_t and we need additional approximations. Specifically, we use the first two moments in a second-order Taylor expansion to derive an approximation to the VB iterative step with respect to $\hat{b}_{t|t}, \Sigma_{t|t}$.

3 Approximate Variational Bayes

We first present a detailed expression of the KL divergence defined in Equation (2) in the following Lemma:

Lemma 2. *There exists a constant c independent of $\hat{\theta}_{t|t}, P_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$ such that*

$$\begin{aligned} &KL\Big(\mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \,||\, p(\theta_t, b_t \mid \mathcal{F}_t)\Big) \\ &= c - \frac{1}{2}\log\det P_{t|t} + \frac{1}{2}\frac{(y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)}{\sigma^2} + \frac{1}{2}\log\det(\Sigma_{t-1|t-1} + r_b I) - \frac{1}{2}\log\det\Sigma_{t|t} \\ &\quad + \frac{1}{2}\operatorname{Tr}\Big(\Big(\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^\top\Big)(\Sigma_{t-1|t-1} + r_b I)^{-1}\Big) \\ &\quad + \frac{1}{2}\int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})\Big(\log\det(P_{t-1|t-1} + g(b_t)) \\ &\qquad\qquad + \operatorname{Tr}\Big((P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top)(P_{t-1|t-1} + g(b_t))^{-1}\Big)\Big) db_t \,. \end{aligned}$$

We observe that the constraint $g(b_t) \succcurlyeq 0$ is sufficient to obtain a finite integral in Lemma 2. The rest of the Section is devoted to minimize the expression of Lemma 2.

Theorem 1 realizes the exact optimum of the KL divergence with respect to $\hat{\theta}_{t|t}, P_{t|t}$, even though A_t^{-1} does not admit an explicit form. We discuss a second-order Taylor approximation for A_t^{-1} in this iterative step in Section 3.2.

3.1 Second-order Taylor approximation of the Kullback-Leibler divergence

The minimization of the KL divergence is analytically intractable due to the integral under b_t (the last term of the expression provided in Lemma 2). We rewrite this last term of the KL divergence as $\int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})\psi_g(b_t)db_t$ with

$$\psi_g(b_t) = \log\det(P_{t-1|t-1} + g(b_t)) + \mathrm{Tr}((P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top)(P_{t-1|t-1} + g(b_t))^{-1})\,.$$

Therefore we use the first two moments of the gaussian distribution $\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})$, after performing a second-order expansion. Specifically, we make the following approximations for a chosen $\bar{b}_t$:

$$\int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})\psi_g(b_t)db_t \approx \psi_g(\bar{b}_t) + \frac{\partial \psi_g}{\partial b_t}\Big|_{\bar{b}_t}^\top (\hat{b}_{t|t} - \bar{b}_t) + \frac{1}{2}\,\mathrm{Tr}\left(\frac{\partial^2 \psi_g}{\partial b_t^2}\Big|_{\bar{b}_t}(\Sigma_{t|t} + (\hat{b}_{t|t} - \bar{b}_t)(\hat{b}_{t|t} - \bar{b}_t)^\top)\right). \tag{4}$$

Using this approximation, we obtain the iterative updates presented in the following theorem:

Theorem 3. *Given $\hat{\theta}_{t|t}, P_{t|t}$, replacing the last term of the KL divergence in Lemma 2 with its approximation* (4)*, the values of $\hat{b}_{t|t}, \Sigma_{t|t}$ minimizing this expression are:*

$$\Sigma_{t|t} = \left((\Sigma_{t-1|t-1} + r_b I)^{-1} + \frac{1}{2}\frac{\partial^2 \psi_g}{\partial b_t^2}\Big|_{\bar{b}_t}\right)^{-1},$$
$$\hat{b}_{t|t} = \Sigma_{t|t}\Big((\Sigma_{t-1|t-1} + r_b I)^{-1}\hat{b}_{t-1|t-1} + \frac{1}{2}\frac{\partial^2 \psi_g}{\partial b_t^2}\Big|_{\bar{b}_t}\bar{b}_t - \frac{1}{2}\frac{\partial \psi_g}{\partial b_t}\Big|_{\bar{b}_t}\Big).$$

Our approach consists in using this approximation as an alternative to the iterative approach of Šmídl and Quinn (2006) in case the prior is not conjugate.

3.2 The algorithm

We focus on two settings. First we consider the scalar setting where b_t is one-dimensional and $g(b_t) = b_t^2 I$. Second we present the diagonal setting where b_t is of the same dimension as the state and $g(b_t) = D_{b_t^2}$ (D_v is the diagonal matrix whose coefficients are the ones of v). We provide the values of the derivatives used in Theorem 3:

Proposition 4. Scalar setting*: If $g(b_t) = b_t^2 I$, we have*

$$\frac{\partial \psi_g}{\partial b_t}\Big|_{\bar{b}_t} = 2\,\mathrm{Tr}(C_t^{-1}(I - B_t C_t^{-1}))\bar{b}_t\,,$$
$$\frac{\partial^2 \psi_g}{\partial b_t^2}\Big|_{\bar{b}_t} = 2\,\mathrm{Tr}(C_t^{-1}(I - B_t C_t^{-1})) + 8\,\mathrm{Tr}\left(C_t^{-2}\Big(B_t C_t^{-1} - \frac{I}{2}\Big)\right)\bar{b}_t^2\,,$$

where $B_t = P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top$ and $C_t = P_{t-1|t-1} + \bar{b}_t^2 I$.

Proposition 5. Diagonal setting*: If $g(b_t) = D_{b_t^2}$, we have*

$$\frac{\partial \psi_g}{\partial b_t}\Big|_{\bar{b}_t} = 2\Delta_{C_t^{-1}(I - B_t C_t^{-1})} \odot \bar{b}_t\,,$$
$$\frac{\partial^2 \psi_g}{\partial b_t^2}\Big|_{\bar{b}_t} = 2(C_t^{-1}(I - B_t C_t^{-1}) \odot I) + 8\Big(C_t^{-1}\Big(B_t C_t^{-1} - \frac{I}{2}\Big) \odot D_{\bar{b}_t} C_t^{-1} D_{\bar{b}_t}\Big)\,,$$

where $B_t = P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top$, $C_t = P_{t-1|t-1} + D_{\bar{b}_t^2}$ and Δ_M is the vector composed of the diagonal of M.

Furthermore, we recall that in Theorem 1 A_t is defined only implicitly. We use a similar second-order Taylor approximation as in the previous subsection to estimate A_t. In the scalar setting, defining $C_t = P_{t-1|t-1} + \overline{b}_t^2 I$ we have

$$\begin{aligned}
A_t &= \int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})(C_t + (2\overline{b}_t(b_t - \overline{b}_t) + (b_t - \overline{b}_t)^2)I)^{-1} db_t \\
&\approx \int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) C_t^{-1}(I - (2\overline{b}_t(b_t - \overline{b}_t) + (b_t - \overline{b}_t)^2)C_t^{-1} + 4\overline{b}_t^2(b_t - \overline{b}_t)^2 C_t^{-2}) db_t \\
&\approx \int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \Big(C_t^{-1} - C_t^{-2}(b_t^2 - \overline{b}_t^2) + 4C_t^{-3}\overline{b}_t^2(b_t - \overline{b}_t)^2\Big) db_t \\
&\approx C_t^{-1} - C_t^{-2}(\hat{b}_{t|t}^2 - \overline{b}_t^2 + \Delta_{\Sigma_{t|t}}) + 4C_t^{-3}\overline{b}_t^2(\Sigma_{t|t} + (\hat{b}_{t|t} - \overline{b}_t)^2)\,.
\end{aligned}$$

In the diagonal setting, defining $C_t = P_{t-1|t-1} + D_{\overline{b}_t^2}$ we have

$$\begin{aligned}
A_t &= \int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})(C_t + 2D_{\overline{b}_t(b_t - \overline{b}_t)} + D_{(b_t - \overline{b}_t)^2})^{-1} db_t \\
&\approx \int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) C_t^{-1}(I - (2D_{\overline{b}_t(b_t - \overline{b}_t)} + D_{(b_t - \overline{b}_t)^2})C_t^{-1} + 4D_{\overline{b}_t(b_t - \overline{b}_t)} C_t^{-1} D_{\overline{b}_t(b_t - \overline{b}_t)} C_t^{-1}) db_t \\
&\approx \int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \Big(C_t^{-1} - C_t^{-1} D_{b_t^2 - \overline{b}_t^2} C_t^{-1} + 4C_t^{-1}(C_t^{-1} \odot \overline{b}_t \overline{b}_t^\top \odot (b_t - \overline{b}_t)(b_t - \overline{b}_t)^\top) C_t^{-1}\Big) db_t \\
&\approx C_t^{-1} - C_t^{-1} D_{\hat{b}_{t|t}^2 - \overline{b}_t^2 + \Delta_{\Sigma_{t|t}}} C_t^{-1} + 4C_t^{-1}(C_t^{-1} \odot \overline{b}_t \overline{b}_t^\top \odot (\Sigma_{t|t} + (\hat{b}_{t|t} - \overline{b}_t)(\hat{b}_{t|t} - \overline{b}_t)^\top)) C_t^{-1}\,.
\end{aligned}$$

Combining our findings we obtain Algorithm 1. As the KL optimization is a coupled problem we solve it in a classical iterative fashion through the Iterative VB algorithm (Šmídl and Quinn, 2006), that is, we repeat N times the updates of Theorems 1 and 3 (for instance $N = 2$).

4 Conclusion

We have presented in this working paper a recursive estimation of the state-space covariance matrix based on the VB approach. We assumed the variance of the observation to be known and constant and we focused on the joint online estimation of the state together with the time-varying covariance matrix of the state process.

Algorithm 1: Variational Bayesian State-Space Model at time step t

Inputs: $\hat{\theta}_{t-1|t-1}, P_{t-1|t-1}, \hat{b}_{t-1|t-1}, \Sigma_{t-1|t-1}, x_t, y_t$.
Initialize: $\hat{b}_{t|t}^{(0)} = \hat{b}_{t-1|t-1}, \Sigma_{t|t}^{(0)} = \Sigma_{t-1|t-1} + r_b I$.
Iterate: for $i = 1, \ldots, N$:

1. • **If scalar setting:** Set $\bar{b}_t = \sqrt{(\hat{b}_{t|t}^{(i-1)})^2 + \Sigma_{t|t}^{(i-1)}}$ and $C_t = K P_{t-1|t-1} K^\top + \bar{b}_t^2 I$.
   Set $A_t^{-1} = \Big(C_t^{-1} + 4C_t^{-3}\bar{b}_t^2(\Sigma_{t|t}^{(i-1)} + (\hat{b}_{t|t}^{(i-1)} - \bar{b}_t)^2)\Big)^{-1}$.
   • **If diagonal setting:** Set $\bar{b}_t = \sqrt{(\hat{b}_{t|t}^{(i-1)})^2 + \Delta_{\Sigma_{t|t}^{(i-1)}}}$ and $C_t = K P_{t-1|t-1} K^\top + D_{\bar{b}_t^2}$.
   Set $A_t^{-1} = \Big(C_t^{-1} + 4C_t^{-1}(C_t^{-1} \odot \bar{b}_t \bar{b}_t^\top \odot (\Sigma_{t|t}^{(i-1)} + (\hat{b}_{t|t}^{(i-1)} - \bar{b}_t)(\hat{b}_{t|t}^{(i-1)} - \bar{b}_t)^\top))C_t^{-1}\Big)^{-1}$.
2. **Update the posterior for θ_t:**
   $P_{t|t}^{(i)} = A_t^{-1} - \frac{A_t^{-1} x_t x_t^\top A_t^{-1}}{x_t^\top A_t^{-1} x_t + \sigma^2}, \qquad \hat{\theta}_{t|t}^{(i)} = \hat{\theta}_{t-1|t-1} + \frac{A_t^{-1} x_t}{x_t^\top A_t^{-1} x_t + \sigma^2}(y_t - x_t^\top \hat{\theta}_{t-1|t-1})$.
3. **Update the posterior for Q_t:**
   Set $B_t = P_{t|t}^{(i)} + (\hat{\theta}_{t|t}^{(i)} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t}^{(i)} - \hat{\theta}_{t-1|t-1})^\top$.
   • **If scalar setting:**
   $\frac{\partial \psi_g}{\partial b_t}\Big|_{\bar{b}_t} = 2\,\mathrm{Tr}(C_t^{-1}(I - B_t C_t^{-1}))\bar{b}_t$,
   $\frac{\partial^2 \psi_g}{\partial b_t^2}\Big|_{\bar{b}_t} = 2\,\mathrm{Tr}(C_t^{-1}(I - B_t C_t^{-1})) + 8\,\mathrm{Tr}(C_t^{-2}(B_t C_t^{-1} - \frac{I}{2}))\bar{b}_t^2$.
   • **If diagonal setting:**
   $\frac{\partial \psi_g}{\partial b_t}\Big|_{\bar{b}_t} = 2\Delta_{C_t^{-1}(I - B_t C_t^{-1})} \odot \bar{b}_t$,
   $\frac{\partial^2 \psi_g}{\partial b_t^2}\Big|_{\bar{b}_t} = 2(C_t^{-1}(I - B_t C_t^{-1}) \odot I) + 8\Big(C_t^{-1}(B_t C_t^{-1} - \frac{I}{2}) \odot D_{\bar{b}_t} C_t^{-1} D_{\bar{b}_t}\Big)$.

$\Sigma_{t|t}^{(i)} = \Big((\Sigma_{t-1|t-1} + r_b I)^{-1} + \frac{1}{2}\frac{\partial^2 \psi_g}{\partial b_t^2}\Big|_{\bar{b}_t}\Big)^{-1}$.
$\hat{b}_{t|t}^{(i)} = \Sigma_{t|t}\Big((\Sigma_{t-1|t-1} + r_b I)^{-1}\hat{b}_{t-1|t-1} + \frac{1}{2}\frac{\partial^2 \psi_g}{\partial b_t^2}\Big|_{\bar{b}_t}\bar{b}_t - \frac{1}{2}\frac{\partial \psi_g}{\partial b_t}\Big|_{\bar{b}_t}\Big)$.

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

References

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

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

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*, 63(2):594–601, 2017.

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*, 67(12):3552–3556, 2020.

R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Journal of basic engineering*, 83(1): 95–108, 1961.

R. Mehra. Approaches to adaptive filtering. *IEEE Transactions on Automatic Control*, 17(5):693–698, 1972.

S. Sarkka and A. Nummenmaa. Recursive noise adaptive kalman filtering by variational bayesian approximations. *IEEE Transactions on Automatic Control*, 54(3):596–600, 2009.

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

D. G. Tzikas, A. C. Likas, and N. P. Galatsanos. The variational approximation for bayesian inference. *IEEE Signal Processing Magazine*, 25(6):131–146, 2008.

A Proofs

Proof of Lemma 2. We start from the expression of Equation (2) that we can write in the following form:

$$\begin{aligned}
&KL\Big(\mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \mid\mid p(\theta_t, b_t \mid \mathcal{F}_t)\Big) \\
&\quad = \int \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t}) \log \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t}) d\theta_t + \int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \log \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) db_t \\
&\qquad - \int \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \log p(\theta_t, b_t \mid \mathcal{F}_t) d\theta_t db_t \,.
\end{aligned}$$

The entropy of gaussian variables is easily computed. The last term can be split using the factored form of Equation (3) and we observe that

$$\begin{aligned}
&\int \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \log \mathcal{N}(y_t \mid \theta_t^\top x_t, \sigma^2) d\theta_t db_t \\
&\quad = \int \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t}) \Big(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{(y_t - \theta_t^\top x_t)^2}{\sigma^2} \Big) d\theta_t \\
&\quad = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{(y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)}{\sigma^2} \,, \\
&\int \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \log \mathcal{N}(\theta_t \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1} + g(b_t)) d\theta_t db_t \\
&\quad = \int \mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \Big(-\frac{d\log(2\pi)}{2} - \frac{1}{2}\log\det(P_{t-1|t-1} + g(b_t)) \\
&\qquad\qquad - \frac{1}{2}(\theta_t - \hat{\theta}_{t-1|t-1})^\top (P_{t-1|t-1} + g(b_t))^{-1} (\theta_t - \hat{\theta}_{t-1|t-1}) \Big) dl_t d\theta_t \\
&\quad = -\frac{d\log(2\pi)}{2} - \frac{1}{2}\int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \Big(\log\det(P_{t-1|t-1} + g(b_t)) \\
&\qquad\qquad + \operatorname{Tr}\Big((P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top)(P_{t-1|t-1} + g(b_t))^{-1} \Big) \Big) db_t \,.
\end{aligned}$$

Combining the last few equations we obtain

$$\begin{aligned}
&KL\Big(\mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \mid\mid p(\theta_t, b_t \mid \mathcal{F}_t)\Big) \\
&\quad = -\frac{1}{2}(1 + d\log(2\pi) + \log\det P_{t|t}) - \frac{1}{2}(1 + d\log(2\pi) + \log\det \Sigma_{t|t}) \\
&\qquad + \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\frac{(y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)}{\sigma^2} \\
&\qquad + \frac{d\log(2\pi)}{2} + \frac{1}{2}\int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \Big(\log\det(P_{t-1|t-1} + g(b_t)) \\
&\qquad\qquad + \operatorname{Tr}\Big((P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top)(P_{t-1|t-1} + g(b_t))^{-1} \Big) \Big) db_t \\
&\qquad + \frac{1}{2}(d\log(2\pi) + \log\det(\Sigma_{t-1|t-1} + r_b I)) + \frac{1}{2}\operatorname{Tr}\Big(\Big(\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^\top \Big) (\Sigma_{t-1|t-1} + r_b I)^{-1} \Big) \\
&\qquad + \log p(\mathcal{F}_t) - \log p(x_t, \mathcal{F}_{t-1}) \,.
\end{aligned}$$

□

Proof of Theorem 1. Thanks to Lemma 2 we have

$$\begin{aligned}
&KL\Big(\mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \mid\mid p(\theta_t, a_t, b_t \mid \mathcal{F}_t)\Big) \\
&\quad = -\frac{1}{2}\log\det P_{t|t} + \frac{1}{2}\frac{(y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t}{\sigma^2} \\
&\qquad + \frac{1}{2}\operatorname{Tr}\Big((P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top)\int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})(P_{t-1|t-1} + g(b_t))^{-1} db_t\Big) + c_\theta\,,
\end{aligned}$$

where c_θ is a constant independent of $\hat{\theta}_{t|t}, P_{t|t}$. We define $A_t = \int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})(P_{t-1|t-1} + g(b_t))^{-1} db_t$, then the first order conditions are written as

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

It yields

$$\begin{aligned}
P_{t|t}^\star &= \left(\frac{x_t x_t^\top}{\sigma^2} + A_t\right)^{-1} = A_t^{-1} - \frac{A_t^{-1} x_t x_t^\top A_t^{-1}}{x_t^\top A_t^{-1} x_t + \sigma^2} \\
\hat{\theta}_{t|t}^\star &= P_{t|t}^\star\left(\frac{y_t x_t}{\sigma^2} + A_t\hat{\theta}_{t-1|t-1}\right) = \hat{\theta}_{t-1} + \frac{A_t^{-1}x_t}{x_t^\top A_t^{-1} x_t + \sigma^2}(y_t - x_t^\top \hat{\theta}_{t-1|t-1})\,.
\end{aligned}$$

□

Proof of Theorem 3. Combining Lemma 2 and the approximation of Equation (4), we obtain:

$$\begin{aligned}
&KL\Big(\mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \mid\mid p(\theta_t, b_t \mid \mathcal{F}_t)\Big) \\
&\quad = \frac{1}{2}\int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})\psi_g(b_t) db_t - \frac{1}{2}\log\det\Sigma_{t|t} + \frac{1}{2}\operatorname{Tr}\left(\Big(\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^\top\Big)(\Sigma_{t-1|t-1} + r_b I)^{-1}\right) + c \\
&\quad \approx \frac{1}{2}\psi_g(\bar{b}_t) + \frac{1}{2}\frac{\partial\psi_g}{\partial b_t}\Big|_{\bar{b}_t}^\top(\hat{b}_{t|t} - \bar{b}_t) + \frac{1}{4}\operatorname{Tr}\Big(\frac{\partial^2\psi_g}{\partial b_t^2}\Big|_{\bar{b}_t}(\Sigma_{t|t} + (\hat{b}_{t|t} - \bar{b}_t)(\hat{b}_{t|t} - \bar{b}_t)^\top)\Big) \\
&\qquad - \frac{1}{2}\log\det\Sigma_{t|t} + \frac{1}{2}\operatorname{Tr}\left(\Big(\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^\top\Big)(\Sigma_{t-1|t-1} + r_b I)^{-1}\right) + c_b\,,
\end{aligned}$$

where c_b is a constant independent of $\hat{b}_{t|t}, \Sigma_{t|t}$. Therefore the first order condition yield

$$\begin{aligned}
&\frac{1}{4}\frac{\partial^2\psi_g}{\partial b_t^2}\Big|_{\bar{b}_t} - \frac{1}{2}\Sigma_{t|t}^{-1} + \frac{1}{2}(\Sigma_{t-1|t-1} + r_b I)^{-1} = 0\,, \\
&\frac{1}{2}\frac{\partial\psi_g}{\partial b_t}\Big|_{\bar{b}_t} + \frac{1}{2}\frac{\partial^2\psi_g}{\partial b_t^2}\Big|_{\bar{b}_t}(\hat{b}_{t|t} - \bar{b}_t) + (\Sigma_{t-1|t-1} + r_b I)^{-1}(\hat{b}_{t|t} - \hat{b}_{t-1|t-1}) = 0\,,
\end{aligned}$$

and the result follows immediately. □

Proof of Proposition 4. We recall that in the scalar setting,

$$\psi_g(b_t) = \log\det(P_{t-1|t-1} + b_t^2 I) + \operatorname{Tr}((P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top)(P_{t-1|t-1} + b_t^2 I)^{-1})\,.$$

Furthermore, note that if $A \succ 0$, it holds $\det A = \exp\operatorname{Tr}(\operatorname{Log} A)$. We define $C_t = P_{t-1|t-1} + \bar{b}_t^2 I$ and we get

$$\begin{aligned}
\log\det(P_{t-1|t-1} + b_t^2 I) &= \operatorname{Tr}\operatorname{Log}(P_{t-1|t-1} + b_t^2 I) \\
&= \operatorname{Tr}\operatorname{Log}(C_t) + \operatorname{Tr}\operatorname{Log}\Big(I + (2\bar{b}_t(b_t - \bar{b}_t) + (b_t - \bar{b}_t)^2)C_t^{-1}\Big) \\
&= \operatorname{Tr}\operatorname{Log}(C_t) + \operatorname{Tr}\Big((2\bar{b}_t(b_t - \bar{b}_t) + (b_t - \bar{b}_t)^2)C_t^{-1} - \frac{1}{2}(2\bar{b}_t(b_t - \bar{b}_t)C_t^{-1})^2 + o((b_t - \bar{b}_t)^2)\Big)\,.
\end{aligned}$$

The last line follows from the series expansion of the Logarithm. We apply another series expansion for the second term of ψ_g: defining $B_t = P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top$ we have

$$\begin{aligned}\operatorname{Tr}(B_t(P_{t-1|t-1} + b_t^2 I)^{-1}) &= \operatorname{Tr}\left(B_t C_t^{-1}\Big(I + (2\overline{b}_t(b_t - \overline{b}_t) + (b_t - \overline{b}_t)^2)C_t^{-1}\Big)^{-1}\right) \\ &= \operatorname{Tr}\left(B_t C_t^{-1}\Big(I - (2\overline{b}_t(b_t - \overline{b}_t) + (b_t - \overline{b}_t)^2)C_t^{-1} + (2\overline{b}_t(b_t - \overline{b}_t)C_t^{-1})^2 + o((b_t - \overline{b}_t)^2)\Big)\right).\end{aligned}$$

Summing the last two equations, and using the identity $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$, we obtain

$$\begin{aligned}\psi_g(b_t) = \operatorname{Tr}\operatorname{Log}(C_t) + \operatorname{Tr}(B_t C_t^{-1}) + 2\operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1}))\overline{b}_t(b_t - \overline{b}_t) + \operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1}))(b_t - \overline{b}_t)^2 \\ - 4\operatorname{Tr}(C_t^{-2}(\frac{I}{2} - B_t C_t^{-1}))\overline{b}_t^2(b_t - \overline{b}_t)^2 + o((b_t - \overline{b}_t)^2)\,.\end{aligned}$$

We can identify the first and second derivatives of ψ_g, that yields Proposition 4. □

Proof of Proposition 5. The proof is similar to the one of Proposition 4. We recall that in the diagonal setting,

$$\psi_g(b_t) = \log\det(P_{t-1|t-1} + D_{b_t^2}) + \operatorname{Tr}((P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top)(P_{t-1|t-1} + D_{b_t^2})^{-1})\,.$$

Furthermore, note that if $A \succ 0$, it holds $\det A = \exp\operatorname{Tr}(\operatorname{Log} A)$. We define $C_t = P_{t-1|t-1} + D_{\overline{b}_t^2}$ and we get

$$\begin{aligned}\log\det(P_{t-1|t-1} + D_{b_t^2}) &= \operatorname{Tr}\operatorname{Log}(P_{t-1|t-1} + D_{b_t^2}) \\ &= \operatorname{Tr}\operatorname{Log}(C_t) + \operatorname{Tr}\operatorname{Log}\left(I + (2D_{\overline{b}_t}D_{b_t - \overline{b}_t} + D^2_{b_t - \overline{b}_t})C_t^{-1}\right) \\ &= \operatorname{Tr}\operatorname{Log}(C_t) + \operatorname{Tr}\left((2D_{\overline{b}_t}D_{b_t - \overline{b}_t} + D^2_{b_t - \overline{b}_t})C_t^{-1} - \frac{1}{2}(2D_{\overline{b}_t}D_{b_t - \overline{b}_t}C_t^{-1})^2 + o(\|b_t - \overline{b}_t\|^2)\right).\end{aligned}$$

The last line follows from the series expansion of the Logarithm. We apply another series expansion for the second term of ψ_g: defining $B_t = P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top$ we have

$$\begin{aligned}\operatorname{Tr}(B_t(P_{t-1|t-1} + D_{b_t^2})^{-1}) &= \operatorname{Tr}\left(B_t C_t^{-1}\Big(I + (2D_{\overline{b}_t}D_{b_t - \overline{b}_t} + D^2_{b_t - \overline{b}_t})C_t^{-1}\Big)^{-1}\right) \\ &= \operatorname{Tr}\left(B_t C_t^{-1}\Big(I - (2D_{\overline{b}_t}D_{b_t - \overline{b}_t} + D^2_{b_t - \overline{b}_t})C_t^{-1} + (2D_{\overline{b}_t}D_{b_t - \overline{b}_t}C_t^{-1})^2 + o(\|b_t - \overline{b}_t\|^2)\Big)\right).\end{aligned}$$

Summing the last two equations, and using the identity $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$, we obtain

$$\begin{aligned}\psi_g(b_t) = \operatorname{Tr}\operatorname{Log}(C_t) + \operatorname{Tr}(B_t C_t^{-1}) + 2\operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1})D_{\overline{b}_t}D_{b_t - \overline{b}_t}) + \operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1})D^2_{b_t - \overline{b}_t}) \\ - 4\operatorname{Tr}(C_t^{-1}(\frac{I}{2} - B_t C_t^{-1})D_{\overline{b}_t}D_{b_t - \overline{b}_t}C_t^{-1}D_{\overline{b}_t}D_{b_t - \overline{b}_t}) + o(\|b_t - \overline{b}_t\|^2)\,.\end{aligned}$$

To conclude we use the following identity:

$$\operatorname{Tr}\left(AD_v B D_v\right) = \sum_{i,j} a_{i,j}v_j b_{j,i}v_i = v^\top(A^\top \odot B)v\,.$$

Noting $\odot$ the Hadamard product and Δ_M the vector composed of the diagonal coefficient of M, we obtain

$$\begin{aligned}\psi_g(b_t) = \operatorname{Tr}\operatorname{Log}(C_t) + \operatorname{Tr}(B_t C_t^{-1}) + (2\Delta_{C_t^{-1}(I - B_t C_t^{-1})} \odot \overline{b}_t)^\top(b_t - \overline{b}_t) + (b_t - \overline{b}_t)^\top(C_t^{-1}(I - B_t C_t^{-1}) \odot I)(b_t - \overline{b}_t) \\ - 4(b_t - \overline{b}_t)^\top\Big(C_t^{-1}(\frac{I}{2} - B_t C_t^{-1}) \odot D_{\overline{b}_t}C_t^{-1}D_{\overline{b}_t}\Big)(b_t - \overline{b}_t) + o(\|b_t - \overline{b}_t\|^2)\,.\end{aligned}$$

We can identify the first and second derivatives of ψ_g, that yields Proposition 5. □