VIKING: Variational Bayesian Variance Tracking
Winning a Post-Covid Day-Ahead Electricity Load Forecasting Competition

Joseph de Vilmarest [1 2] **Yannig Goude** [1 3] **Olivier Wintenberger** [2]

Abstract

We present a novel variational bayesian approach for time series forecasting following from a state-space representation, named VIKING (**V**ariational Bayes**I**an Variance Trac**KING**). The method is illustrated with the procedure used to win a recent competition on post-covid electricity load forecasting.

1. Introduction

Electricity demand forecasting is a crucial task for grid operators. Indeed the production must balance the consumption as storage capacities are still negligible compared to the load. Time series methods have been applied to address that problem, relying on calendar information and lags of the electricity consumption. Statistical and machine learning models have been designed to use exogenous information such as meteorological forecasts (the load usually depends on the temperature for instance, due to electric heating and cooling). However the behavior of the consumption changed abruptly during the coronavirus crisis, especially during lockdowns imposed by many governments. These changes of consumption mode have been challenging for electricity grid operators as historical forecasting procedures performed poorly. Therefore designing new forecasting strategies to take that evolution into account is important to reduce the cost of forecasting errors and to ensure the stability of the network in the future.

Our methodology extends a previous work on the French electricity load (Obst et al., 2021). We present a state-space approach to adapt statistical and machine learning methods. After applying a standard Kalman Filter we present a novel approach to adaptively estimate the observation and state noise variances based on the Variational Bayes approach as in (Huang et al., 2020). Our procedure resulted in the winning strategy in a competition on post-covid day-ahead electricity demand forecasting[1]. In Section 2 we present the competition along with classical forecasting methods, then in Section 3 we present a generic way to adapt these methods. Finally we present numerical experiments in Section 4. For clarity we apply some pruning compared to the winning submissions with very little performance degradation, and we focus on adaptation methods as it is the core of our strategy.

2. Competition presentation

The objective of the competition was to predict the electricity load of an undisclosed location of average consumption 1.1 GW, that is of the order of one million people in western countries. We had access to the historical load starting from March 18$^{\text{th}}$ 2017 and the evaluation was based on the Mean Average Error (MAE) on the period ranging from January 18$^{\text{th}}$ to February 16$^{\text{th}}$ 2021. The break in the electricity consumption due to coronavirus is presented in Figure 1.

The competition's setting was to forecast the hourly load 16 to 40 hours ahead, precisely during the 30-day evaluation period, we predicted day d with at our disposal the data updated up to day d-1 at 8AM.

2.1. The dataset

To forecast the load, meteorological forecasts are provided (as well as historical realized meteorology): temperature, cloud cover, pressure, wind direction and speed. From the statistical properties of the meteorological residuals, we gather that the forecasts come from physical models that need to be statistically corrected. We thus correct the meteorological forecasts via autoregressive models on the residuals. Finally we use the following explanatory variables:

- calendar variables: the day of the week, the time of year (Toy) growing linearly from 0 on January 1$^{\text{st}}$ to 1 on December 31$^{\text{st}}$, and a variable growing linearly with time to account for a trend,

[1]Électricité de France R&D, Palaiseau, France [2]Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Paris, France [3]Laboratoire de Mathématique d'Orsay, Université Paris-Saclay, France. Correspondence to: Joseph de Vilmarest <joseph.de_vilmarest@upmc.fr>.

Time Series Workshop at the 38^{th} International Conference on Machine Learning, 2021.

[1]https://ieee-dataport.org/competitions/day-ahead-electricity-demand-forecasting-post-covid-paradigm

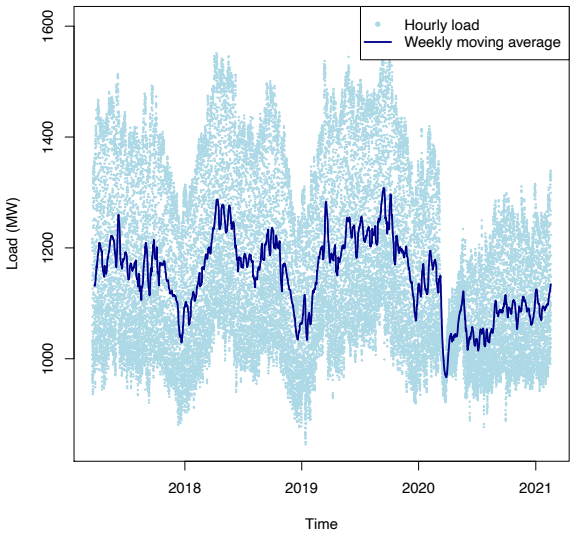

Figure 1. Electricity load from March 18th 2017 to February 16th 2021.

Figure 2. Evolution of the forecasting error for the different models introduced in Section 2.2 trained on the data up to January 1st 2020.

- meteorological forecasts after statistical correction: the temperature along with an exponential smoothing variant of parameter 0.95 ($Temps95$), the cloud cover, the pressure, the wind direction and speed,
- lags of the electricity load: the load a week ago $LoadW$ and the last load available $LoadD$ (a day ago for the forecast before 8AM and two days ago after 8AM, due to the availability of the online data during the competition).

2.2. Statistical and Machine Learning methods

We apply a few classical predictive models. It is usual in electricity load forecasting to define independent models for the different hours of the day. For each model we use the same structure for the different hours but we learn the model parameters independently for each hour of the day. However the correlation between different hours is important and therefore to capture intraday information we fit on the residuals of each model an autoregressive model incorporating lags of the 24 last available hours and optimized for each forecast horizon.

Autoregressive. We consider a seasonal autoregressive model based on the daily and weekly lags of the load.

Linear regression. We use a linear model with the following variables: temperature, cloud cover, pressure, wind direction and speed, day type (7 booleans), the linear trend parameter, Toy, $LoadW$ and $LoadD$.

Generalized additive model (GAM). We propose a Gaussian generalized additive model (Wood, 2017):

$$y_t = \alpha t + \sum_{i=1}^{6} \beta_i \mathbb{1}_{DayType_t=i} + \gamma Temps95_t + f_1(Toy_t) + f_2(LoadD_t) + f_3(LoadW_t) + \beta_0 + \varepsilon_t \,,$$

where y_t is the load, f_1 is obtained by penalized regression on cubic cyclic splines, f_2, f_3 on cubic regression splines.

Multi-Layer Perceptron (MLP). Finally we test a multi-layer perceptron with 2 hidden layers of 15 and 10 neurons, taking as input the linear trend parameter, Toy, $DayType$, $LoadD$, $LoadW$, $Temps95_t$.

3. Model adaptation

Due to the lockdowns the behaviour of the load changed abruptly and therefore the models presented in Section 2.2 perform poorly during Spring 2020 and afterwards, see Figure 2. To adapt the models in time, we rely on linear gaussian state-space models, summarized as

$$\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t) \,,$$
$$y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma_t^2) \,,$$

where σ_t^2 is the observation variance and Q_t the process noise covariance matrix.

This state-space representation is natural for linear regression for which x_t is the vector containing the explanatory variables detailed in Section 2.2. Autoregressive and linear

models fit directly in that framework. To adapt GAM and MLP we linearize the models and x_t is just another feature representation. We freeze the non-linear effects in the GAM as in (Obst et al., 2021), and x_t contains the different effects, linear and non-linear. We apply a similar approach for the MLP, for which we freeze the deepest layers and we learn the last one, that is, x_t is the final hidden state.

3.1. Kalman Filter

The estimation $\hat{\theta}_t$ of the state in linear gaussian state-space model and the associated variance P_t is well-understood assuming that $\hat{\theta}_1, P_1, \sigma_t^2, Q_t$ are known. The best estimator is obtained by the well known Kalman Filter (Kalman & Bucy, 1961). However there is no consensus in the literature as to how to tune the hyper-parameters, see for instance (Brockwell et al., 1991; Durbin & Koopman, 2012; Fahrmeir, 1992). The widely used Expectation-Maximization algorithm is an iterative algorithm that guarantees to converge to a local maximum of the likelihood. However there is no global guarantee and in our case it performs poorly. We propose instead the following settings, building on (Obst et al., 2021):

Static. We consider the degenerate setting where $Q_t = 0$ and $\hat{\theta}_1 = 0, P_1 = I, \sigma_t^2 = 1$.

Static break. We consider a break at March 1st 2020 by setting $\hat{\theta}_1 = 0, P_1 = I, \sigma_t^2 = 1, Q_t = 0$ except $Q_T = I$ where T is March 1st 2020.

Dynamic. We approximate the maximum-likelihood for constant σ_t^2, Q_t. We set $P_1 = \sigma^2 I$ and we observe that for a given Q/σ^2 we have closed-form solutions for $\hat{\theta}_1, \sigma^2$. Then we restrict ourselves to diagonal matrices Q/σ^2 whose nonzero coefficients are in $\{2^j, -30 \leq j \leq 0\}$ and we apply a greedy procedure: starting from $Q/\sigma^2 = 0$ we change at each step the coefficient improving the most the likelihood. That procedure is designed to optimize Q on the training data (up to January 1st 2020).

Dynamic break. We use similar $\hat{\theta}_1, P_1, \sigma_t^2 = \sigma^2, Q_t = Q$ as in the dynamic setting except $Q_T = P_1 = \sigma^2 I$ where T is March 1st 2020.

Dynamic big. We simply use $\sigma^2 = 1$ and a matrix Q proportional to I optimized based on the 2020 data.

3.2. Dynamical variances

We would like to learn the variances σ_t^2, Q_t in an adaptive fashion. We thus treat them as latent variables and we augment the state-space model:

$$
\begin{aligned}
&a_t - a_{t-1} \sim \mathcal{N}(0, \rho_a)\,, \quad b_t - b_{t-1} \sim \mathcal{N}(0, \rho_b)\,,\\
&\theta_t - \theta_{t-1} \sim \mathcal{N}(0, \exp(b_t) I)\,,\\
&y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \exp(a_t))\,.
\end{aligned}
$$

Instead of estimating the state θ_t with known variances, we estimate here both the state and the variances represented as log-normal distributions. We have removed these hyper-parameters, however we now have to set priors on a_0, b_0 along with the parameters ρ_a, ρ_b controlling the smoothness of the dynamics on the variances.

We apply a bayesian approach. At each step, we start from a prior $p(\theta_{t-1}, a_{t-1}, b_{t-1} \mid \mathcal{F}_{t-1})$ obtained at the last iteration, where we introduce the filtration of the past observations $\mathcal{F}_t = \sigma(x_1, y_1, ..., x_{t-1}, y_{t-1})$. Then we apply a prediction step thanks to the dynamical equations yielding $p(\theta_t, a_t, b_t \mid \mathcal{F}_{t-1})$. Finally we apply Bayes' rule to derive the posterior distribution $p(\theta_t, a_t, b_t \mid \mathcal{F}_t)$.

However the posterior distribution is analytically intractable, therefore we apply the classical Variational Bayesian (VB) approach (Šmídl & Quinn, 2006). The idea is to approximate recursively the posterior distribution with a factorized distribution. In our setting we look for the best product $\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})$ approximating $p(\theta_t, a_t, b_t \mid \mathcal{F}_t)$. The criterion we minimize is the Kullback-Leibler (KL) divergence

$$
KL(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \,||\, p(\theta_t, a_t, b_t \mid \mathcal{F}_t))\,,
$$

where $KL(p||q) = \int p \log(p/q) dp$. At each step, the VB approach yields a coupled optimization problem in the three gaussian distributions. The classical iterative method (see for instance (Tzikas et al., 2008)) consists in computing alternately $\exp(\mathbb{E}[\log p(\theta_t, a_t, b_t \mid \mathcal{F}_t)])$ where the expected value is taken with respect to two of the three latent variables, and identifying the desired first two moments with respect to the other latent variable. However the expression $\exp(\mathbb{E}_{\theta_t, b_t}[\log p(\theta_t, a_t, b_t \mid \mathcal{F}_t)])$ doesn't match a gaussian distribution in a_t, and similarly for b_t. We therefore need additional approximations. Specifically, we use the first two moments of the gaussian distribution to derive an upper-bound of the KL divergence for which we have an analytical solution. We refer to Appendix A for the detailed derivation of the algorithm, that we call **V**ariational Bayes**i**an Variance Trac**king** (VIKING).

4. Experiments

We display the performance of the introduced methods that we call experts. Then we use aggregation of experts to leverage specificities of each forecaster.

4.1. Individual experts

We have 4 different models (AR, linear, GAM and MLP). For each one, we try the various adaptation settings (no adaptation, KF and VIKING). We illustrate the different settings in Figure 3 where we display the evolution of the state coefficient for the GAM adaptation strategies. We

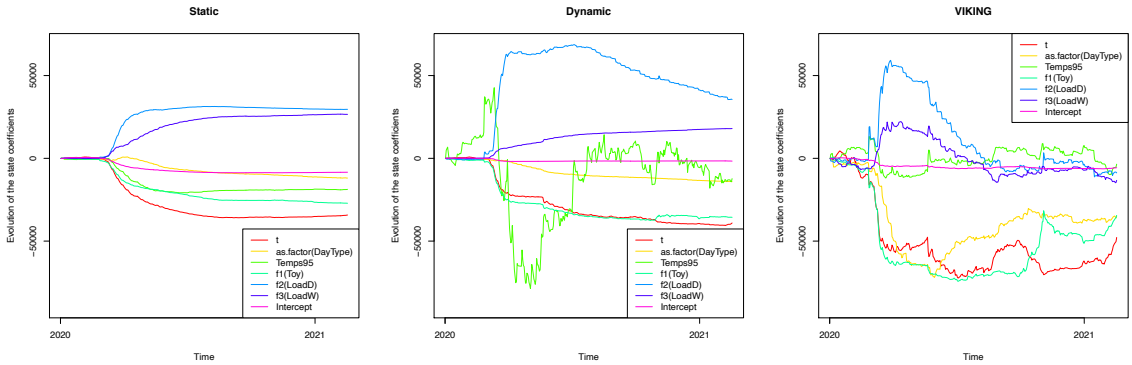

Figure 3. Evolution of the state coefficients for various adaptations of the GAM, see Section 3. On the left, we use the Kalman Filter in the static setting (degenerate covariance matrix $Q_t = 0$). On the middle, the dynamic setting where the variances are constant. On the right, the VIKING setting where we estimate the variances adaptively.

Adaptation	AR	Linear	GAM	MLP
Offline	14.6	22.8	34.3	16.7
Static	20.5	15.7	17.0	22.9
Static break	27.9	14.4	28.4	35.4
Dynamic	14.4	14.9	15.3	13.0
Dynamic break	16.2	13.6	14.3	12.3
Dynamic big	14.3	11.2	13.8	13.7
VIKING	14.4	11.5	12.7	12.5

Table 1. Mean average error of each method (in MW) during the competition evaluation set (2021-01-18 to 2021-02-16).

Adaptation	AR	Linear	GAM	MLP	All
Best expert	14.3	11.2	12.7	12.3	11.2
Aggregation	14.4	11.4	11.6	11.9	**10.9**

Table 2. Mean average error of aggregation strategies (in MW) during the competition evaluation set (2021-01-18 to 2021-02-16).

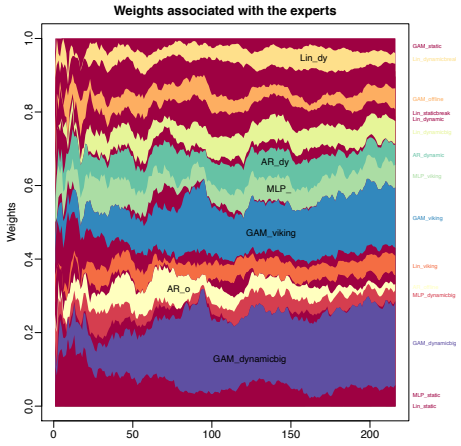

Figure 4. Evolution of the aggregation weights at 3PM from July 1st 2020 to February 16th 2021.

detail the numerical results on the competition dataset in Table 1.

4.2. Aggregation

Online robust aggregation of experts (Cesa-Bianchi & Lugosi, 2006) is a powerful model agnostic approach for time series forecasting, already applied to load forecasting during the lockdown in (Obst et al., 2021). We use the ML-Poly algorithm proposed in (Gaillard et al., 2014) and implemented in the `R` package `opera` (Gaillard & Goude, 2016) to compute these online weights.

The aggregation weights are estimated independently for each hour of the day. We summarize different variants in Table 2. First, for each family of models we compute the aggregation of all the adaptation settings (7 for each). Then we aggregate all of them (28 models). An example of the weights obtained at 3PM is displayed in Figure 4. The aggregation presented in this paper obtains a performance close to our strategy winning the competition (degradation of about 0.05 MW).

5. Conclusion

In this paper we presented a novel time series forecasting approach based on state-space models. Winning a competition on electricity load forecasting motivates its usefulness. Some perspectives have been raised during the competition such as interpretability of the global approach and a better understanding of the error propagation along the different adaptations (AR correction, Kalman filtering, variance tracking and aggregation).

References

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

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

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

Fahrmeir, L. Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, 87 (418):501–509, 1992.

Gaillard, P. and Goude, Y. opera: Online prediction by expert aggregation. *URL: https://CRAN. R-project. org/package= opera. r package version*, 1, 2016.

Gaillard, P., Stoltz, G., and Van Erven, T. A second-order bound with excess losses. In *Conference on Learning Theory*, pp. 176–196, 2014.

Huang, Y., Zhu, F., Jia, G., and Zhang, Y. A slide window variational adaptive kalman filter. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(12):3552–3556, 2020.

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

Obst, D., De Vilmarest, J., and Goude, Y. Adaptive methods for short-term electricity load forecasting during covid-19 lockdown in france. *IEEE Transactions on Power Systems*, 2021.

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

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

Wood, S. N. *Generalized additive models: an introduction with R*. CRC press, 2017.

A. Approximate Variational Bayes

We first present the prediction step and the filtering step in order to obtain the posterior distribution. Propagating the factorized approximation

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

the prediction step becomes:

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

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

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

We then derive a detailed expression of the KL divergence 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*

$$\begin{aligned}
& KL\Big(\mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(a_t \mid \hat{a}_{t|t}, s_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \,||\, p(\theta_t, a_t, b_t \mid \mathcal{F}_t)\Big) \\
& = -\frac{1}{2}\log\det P_{t|t} + \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)\exp(-\hat{a}_{t|t} + \frac{1}{2}s_{t|t}) \\
& \quad + \frac{1}{2}\int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})\Big(\log\det(P_{t-1|t-1} + \exp(b_t)I) \\
& \qquad\qquad + \mathrm{Tr}\left((P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top)(P_{t-1|t-1} + \exp(b_t)I)^{-1}\right)\Big)db_t \\
& \quad - \frac{1}{2}\log(s_{t|t}) + \frac{1}{2(s_{t-1|t-1} + \rho_a)}((\hat{a}_{t|t} - \hat{a}_{t-1|t-1})^2 + s_{t|t}) + \frac{1}{2}\hat{a}_{t|t} \\
& \quad - \frac{1}{2}\log\det\Sigma_{t|t} + \frac{1}{2}\Big(\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^2\Big)(\Sigma_{t-1|t-1} + \rho_b)^{-1} + c\,.
\end{aligned}$$

Proof. We start from the definition of the KL divergence:

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

The entropy of the gaussian variables are easily computed. The last term can be split using the factorized form of Equation

(1) and we observe that

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

Combining the last equations we obtain

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

□

The rest of the Section is devoted to minimize the expression of Lemma 1, which is the criterion for the VB approach. We first obtain the exact minimum with respect to $\hat{\theta}_{t|t}, P_{t|t}$ given fixed $\hat{a}_{t|t}, s_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$:

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*

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

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

Theorem 2 realizes the exact optimum of the KL divergence with respect to $\hat{\theta}_{t|t}, P_{t|t}$, however A_t^{-1} does not admit an explicit form. We discuss an approximation in Section A.2.

Proof. Thanks to Lemma 1 we have

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

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

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

It yields

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

□

A.1. Second-order upper-bound of the Kullback-Leibler divergence

We minimize a second-order upper-bound of the KL divergence. Minimizing the upper-bound does not necessarily lead to minimizing the KL divergence, but we obtain the decrease at each step of the instantaneous KL divergence.

A.1.1. Minimizing an upper-bound in $\hat{a}_{t|t}, s_{t|t}$

We take the following expression of the KL divergence in $\hat{a}_{t|t}, s_{t|t}$:

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

To optimize the previous quantity in $s_{t|t}$ we consider the following upper-bound if $-\rho_a \leq s_{t|t} - s_{t-1|t-1} \leq \rho_a$:

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

and thus we can optimize the upper-bound of the expression on $[s_{t-1|t-1} - \rho_a, s_{t-1|t-1} + \rho_a]$. We obtain the following first order condition:

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

This yields

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

Then we use the following upper-bound for $\hat{a}_{t|t}$: if $|\hat{a}_{t|t} - \hat{a}_{t-1|t-1}| \leq M$ then

$$e^{-\hat{a}_{t|t}} \leq e^{-\hat{a}_{t-1|t-1}}(1 - (\hat{a}_{t|t} - \hat{a}_{t-1|t-1}) + \frac{e^M}{2}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1})^2)\,.$$

Thus the first order condition becomes

$$\frac{1}{s_{t-1|t-1} + \rho_a}(\hat{a}_{t|t} - \hat{a}_{t-1|t-1}) + \frac{1}{2} + \frac{1}{2}((y_t - \hat{\theta}_{t|t}^\top x_t)^2 + x_t^\top P_{t|t} x_t)e^{-\hat{a}_{t-1|t-1} + s_{t|t}/2}\Big(-1 + e^M(\hat{a}_{t|t} - \hat{a}_{t-1|t-1})\Big) = 0\,,$$

yielding

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

We take for instance $M = 100\rho_a$.

A.1.2. Optimization in $\hat{b}_{t|t}, \Sigma_{t|t}$

We now focus on the update of $\hat{b}_{t|t}, \Sigma_{t|t}$. We fix $\hat{\theta}_{t|t}, P_{t|t}, \hat{a}_{t|t}, s_{t|t}$, then the KL divergence is written

$$\begin{aligned}&KL\Big(\mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(a_t \mid \hat{a}_{t|t}, s_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \,||\, p(\theta_t, a_t, b_t \mid \mathcal{F}_t)\Big) \\ &\quad = -\frac{1}{2}\log\det\Sigma_{t|t} + \frac{1}{2}\int \mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t})\psi(b_t)db_t + \frac{1}{2}\Big(\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^2\Big)(\Sigma_{t-1|t-1} + \rho_b)^{-1} + c\,,\end{aligned}$$

where

$$\psi(b_t) = \log\det(P_{t-1|t-1} + \exp(b_t)I) + \mathrm{Tr}\left((P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top)(P_{t-1|t-1} + \exp(b_t)I)^{-1}\right).$$

As the integral is analytically intractable we use the following second-order upper bound:

$$\psi(b_t) \leq \psi(\hat{b}_{t-1|t-1}) + \psi'(\hat{b}_{t-1|t-1})(b_t - \hat{b}_{t-1|t-1}) + \frac{1}{2}\max\psi''([\hat{b}_{t-1|t-1}, b_t])(b_t - \hat{b}_{t-1|t-1})^2\,.$$

This yields the following upper-bound on the KL divergence: if $\max_{b\in\mathbb{R})}\psi''(b) \leq H$, we have

$$\begin{aligned}&KL\Big(\mathcal{N}(\theta_t \mid \hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(a_t \mid \hat{a}_{t|t}, s_{t|t})\mathcal{N}(b_t \mid \hat{b}_{t|t}, \Sigma_{t|t}) \,||\, p(\theta_t, a_t, b_t \mid \mathcal{F}_t)\Big) \\ &\quad \leq -\frac{1}{2}\log\det\Sigma_{t|t} + \frac{1}{2}\Big(\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^2\Big)(\Sigma_{t-1|t-1} + \rho_b)^{-1} + c \\ &\qquad + \psi(\hat{b}_{t-1|t-1}) + \psi'(\hat{b}_{t-1|t-1})(\hat{b}_{t|t} - \hat{b}_{t-1|t-1}) + \frac{1}{2}H(\Sigma_{t|t} + (\hat{b}_{t|t} - \hat{b}_{t-1|t-1})^2)\,.\end{aligned} \tag{2}$$

Therefore we obtain the following theorem under the assumption that we know H:

Theorem 3. *Given $\hat{\theta}_{t|t}, P_{t|t}, \hat{a}_{t|t}, s_{t|t}$, the minimum of the upper-bound on the KL divergence of Equation (2) is obtained with:*

$$\begin{aligned}\Sigma_{t|t} &= \Big((\Sigma_{t-1|t-1} + \rho_b)^{-1} + \frac{1}{2}H\Big)^{-1}\,, \\ \hat{b}_{t|t} &= \hat{b}_{t-1|t-1} - \frac{1}{2}\Sigma_{t|t}\psi'(\hat{b}_{t-1|t-1})\,,\end{aligned}$$

where $\max_{b\in\mathbb{R})}\psi''(b) \leq H$.

Proof. Combining Lemma 1 and the second-order upper-bound, the quantity that we minimize is:

$$-\frac{1}{2}\log\det\Sigma_{t|t}+\frac{1}{2}\psi(\hat{b}_{t-1|t-1})+\frac{1}{2}\psi'(\hat{b}_{t-1|t-1})(\hat{b}_{t|t}-\hat{b}_{t-1|t-1})\\+\frac{1}{4}H(\Sigma_{t|t}+(\hat{b}_{t|t}-\hat{b}_{t-1|t-1})^2)+\frac{1}{2}\Big(\Sigma_{t|t}+(\hat{b}_{t|t}-\hat{b}_{t-1|t-1})^2\Big)(\Sigma_{t-1|t-1}+\rho_b)^{-1}\,.$$

where $\max_{b\in\mathbb{R})}\psi''(b)\leq H$. Therefore the first order conditions are

$$\frac{1}{4}H-\frac{1}{2}\Sigma_{t|t}^{-1}+\frac{1}{2}(\Sigma_{t-1|t-1}+\rho_b)^{-1}=0\,,\\\frac{1}{2}\psi'(\hat{b}_{t-1|t-1})+\frac{1}{2}H(\hat{b}_{t|t}-\hat{b}_{t-1|t-1})+(\Sigma_{t-1|t-1}+\rho_b)^{-1}(\hat{b}_{t|t}-\hat{b}_{t-1|t-1})=0\,,$$

and the result follows. □

We provide in the following proposition the first and second derivatives of ψ:

Proposition 4. *We have for any b_t,*

$$\psi'(b_t)=\mathrm{Tr}(C_t^{-1}(I-B_tC_t^{-1}))\exp(b_t)\,,\\\psi''(b_t)=\mathrm{Tr}(C_t^{-1}(I-B_tC_t^{-1}))\exp(b_t)+2\,\mathrm{Tr}(C_t^{-2}(B_tC_t^{-1}-I/2))\exp(2b_t)\,.$$

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 $C_t=P_{t-1|t-1}+\exp(b_t)I$.

Proof. We recall that

$$\psi(b)=\log\det(P_{t-1|t-1}+\exp(b_t)I)+\mathrm{Tr}(B_t(P_{t-1|t-1}+\exp(b_t)I)^{-1})\,,$$

where $B_t=P_{t|t}+(\hat{\theta}_{t|t}-\hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t}-\hat{\theta}_{t-1|t-1})^\top$. Furthermore, note that if $A\succ 0$, it holds $\det A=\exp\mathrm{Tr}(\mathrm{Log}\,A)$. We define $C_t=P_{t-1|t-1}+\exp(b_t)I$ and we get

$$\begin{aligned}&\log\det(P_{t-1|t-1}+\exp(b)I)\\&\quad=\mathrm{Tr}\,\mathrm{Log}(P_{t-1|t-1}+\exp(b)I)\\&\quad=\mathrm{Tr}\,\mathrm{Log}(C_t)+\mathrm{Tr}\,\mathrm{Log}\left(I+(\exp(b)-\exp(b_t))C_t^{-1}\right)\\&\quad=\mathrm{Tr}\,\mathrm{Log}(C_t)+\mathrm{Tr}\left((\exp(b_t)(b-b_t)+\frac{1}{2}\exp(b_t)(b-b_t)^2)C_t^{-1}-\frac{1}{2}(\exp(b_t)(b-b_t)C_t^{-1})^2+o((b-b_t)^2)\right).\end{aligned}$$

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

$$\begin{aligned}&\mathrm{Tr}(B_t(P_{t-1|t-1}+\exp(b)I)^{-1})\\&\quad=\mathrm{Tr}\left(B_tC_t^{-1}\Big(I+(\exp(b)-\exp(b_t))C_t^{-1}\Big)^{-1}\right)\\&\quad=\mathrm{Tr}\left(B_tC_t^{-1}\Big(I-(\exp(b_t)(b-b_t)+\frac{1}{2}\exp(b_t)(b-b_t)^2)C_t^{-1}+(\exp(b_t)(b-b_t)C_t^{-1})^2+o((b-b_t)^2)\Big)\right).\end{aligned}$$

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

$$\psi(b)=\mathrm{Tr}\,\mathrm{Log}(C_t)+\mathrm{Tr}(B_tC_t^{-1})+\mathrm{Tr}(C_t^{-1}(I-B_tC_t^{-1}))(\exp(b_t)(b-b_t)+\frac{1}{2}\exp(b_t)(b-b_t)^2)\\-\mathrm{Tr}(C_t^{-2}(\frac{I}{2}-B_tC_t^{-1}))\exp(2b_t)(b-b_t)^2+o((b-b_t)^2)\,.$$

Thus we can identify the first and second derivatives of ψ, that yields Proposition 4. □

A.2. The algorithm

Theorem 3 immediately follows from the second-order upper-bound, however we need to define H. Can we obtain an explicit bound for ψ'' ?

True upper-bound: if we don't use the compensations but we just bound each of the four terms, we obtain $\psi'' \leq d + 2\operatorname{Tr}(B_t(P_{t-1|t-1})^{-1})$.

Optimistic bound: we use instead an optimistic upper-bound. We have

$$\begin{aligned}
\psi'''(b_t) &= \operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1}))\exp(b_t) + 6\operatorname{Tr}(C_t^{-2}(B_t C_t^{-1} - I/2))\exp(2b_t) \\
&\quad + 2\operatorname{Tr}(C_t^{-3})\exp(3b_t) - 6\operatorname{Tr}(C_t^{-4}B_t)\exp(3b_t) \\
&= \exp(b_t)\Big(\operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1})) + 6\operatorname{Tr}(C_t^{-2}(B_t C_t^{-1} - I/2))\exp(b_t) + 6\operatorname{Tr}(C_t^{-3}(I/3 - B_t C_t^{-1})\exp(2b_t)\Big) \\
&= 6\operatorname{Tr}(C_t^{-3}(I/3 - B_t C_t^{-1})\exp(b_t)\Big(\alpha + \beta\exp(b_t) + \exp(2b_t)\Big) \\
&= 6\operatorname{Tr}(C_t^{-3}(I/3 - B_t C_t^{-1})\exp(b_t)\big((\exp(b_t) + \frac{1}{2}\beta)^2 - \frac{1}{4}\beta^2 + \alpha\Big)
\end{aligned}$$

where $\alpha = \frac{\operatorname{Tr}(C_t^{-1}(I - B_t C_t^{-1}))}{6\operatorname{Tr}(C_t^{-3}(I/3 - B_t C_t^{-1}))}$ and $\beta = \frac{\operatorname{Tr}(C_t^{-2}(B_t C_t^{-1} - I/2))}{\operatorname{Tr}(C_t^{-3}(I/3 - B_t C_t^{-1}))}$. If $\beta^2 \geq 4\alpha$ then we can write

$$\psi'''(b_t) = 6\operatorname{Tr}(C_t^{-3}(I/3 - B_t C_t^{-1})\exp(b_t)\Big(\exp(b_t) + \frac{1}{2}\beta - \sqrt{\frac{1}{4}\beta^2 - \alpha}\Big)\Big(\exp(b_t) + \frac{1}{2}\beta + \sqrt{\frac{1}{4}\beta^2 - \alpha}\Big)$$

Thus fixing a constant $C_t = (P_{t-1|t-1} + \exp(\hat{b}_{t-1|t-1} - \rho_b/2)I)^{-1}$ we have two possible roots for ψ''' (if they exist) which are $b_t^{(1)} = \log\Big(-\frac{1}{2}\beta + \sqrt{\frac{1}{4}\beta^2 - \alpha}\Big), b_t^{(2)} = \log\Big(-\frac{1}{2}\beta - \sqrt{\frac{1}{4}\beta^2 - \alpha}\Big)$. Our optimistic constant is thus the maximum of $0 = \psi''(\infty), \psi''(\hat{b}_{t-1|t-1} - \rho_b/2), \psi''(b_t^{(1)})$ and $\psi''(b_t^{(2)})$.

We recall that in Theorem 2 A_t is defined only implicitly. We use the following second-order approximation to estimate A_t:

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

where $C_t = P_{t-1|t-1} + \exp(\hat{b}_{t|t})I$.

Combining our findings we obtain Algorithm 1. 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 (in our experiments $N = 2$).

B. Additional numerical results

We present in Table 3 the numerical performance of our methods before intraday correction, that is the models built entirely independently at each time of day.

Algorithm 1 Variational Bayesian Variance Tracking (VIKING) at time step t

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: $\hat{a}_{t|t}^{(0)} = \hat{a}_{t-1|t-1}, s_{t|t}^{(0)} = s_{t-1|t-1} + \rho_a, \hat{b}_{t|t}^{(0)} = \hat{b}_{t-1|t-1}, \Sigma_{t|t}^{(0)} = \Sigma_{t-1|t-1} + \rho_b$.
Iterate: for $i = 1, \dots, N$:

1. Set $A_t = C_t^{-1} - \frac{1}{2}\exp(\hat{b}_{t|t})\Sigma_{t|t}C_t^{-2} + \exp(2\hat{b}_{t|t})\Sigma_{t|t}C_t^{-3}$ with $C_t = P_{t-1|t-1} + \exp(\hat{b}_{t|t}^{(i-1)})I$.
   Update $P_{t|t}^{(i)} = A_t^{-1} - \frac{A_t^{-1}x_t x_t^\top A_t^{-1}}{x_t^\top A_t^{-1} x_t + \exp(\hat{a}_{t|t}^{(i-1)} - \frac{1}{2}s_{t|t}^{(i-1)})}$.
   Update $\hat{\theta}_{t|t}^{(i)} = \hat{\theta}_{t-1|t-1} + \frac{A_t^{-1}x_t}{x_t^\top A_t^{-1} x_t + \exp(\hat{a}_{t|t}^{(i-1)} - \frac{1}{2}s_{t|t}^{(i-1)})}(y_t - x_t^\top \hat{\theta}_{t-1|t-1})$.

2. **If we learn σ_t^2:**
   Update $s_{t|t}^{(i)}, \hat{a}_{t|t}^{(i)}$ thanks to Section A.1.1 using $\hat{\theta}_t^{(i)}, P_t^{(i)}$.

3. **If we learn Q_t:**
   Update $\Sigma_{t|t}^{(i)}, \hat{b}_{t|t}^{(i)}$ thanks to Section A.1.2 using $\hat{\theta}_t^{(i)}, P_t^{(i)}$.

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

Adaptation	AR	Linear	GAM	MLP
Offline	17.6	39.9	48.3	38.8
Static	18.3	16.5	19.8	23.1
Static break	23.5	15.8	17.3	35.4
Dynamic	14.9	15.0	15.4	13.1
Dynamic break	17.8	14.3	16.2	12.8
Dynamic big	17.1	11.8	13.5	12.7
VIKING	16.4	12.0	13.2	12.8

Table 3. Mean average error of each method (in MW) during the competition evaluation set (2021-01-18 to 2021-02-16) without intraday correction.