VIKING: Variational Bayesian Variance Tracking

Application to Adaptive Electricity Load Forecasting

Joseph de Vilmarest [1,2], Olivier Wintenberger [1], Yannig Goude [2]

October 4th 2021

[1]LPSM, Sorbonne Université
[2]EDF R&D

1. Introduction

2. State-Space Representation: Kalman Filtering

3. Variance Tracking: VIKING

Adaptive Time Series Forecasting

We aim at forecasting $y_t \in \mathbb{R}$ given explanatory variables $x_t \in \mathbb{R}^d$.

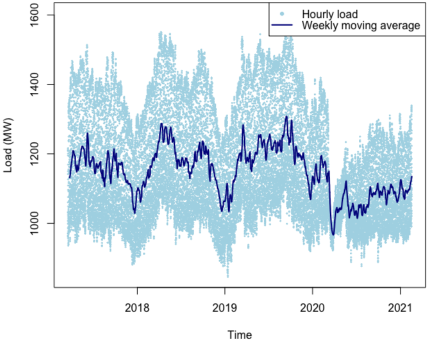
- **Non adaptive**: we predict $\hat{y}_t = f(x_t)$ where f is optimized on a historical data set.
- **Adaptive**: we predict $\hat{y}_t = f_t(x_t)$ and then we update the forecasting model: $f_{t+1} = \Phi(f_t, x_t, y_t)$.

Motivation: Electricity Load Forecasting

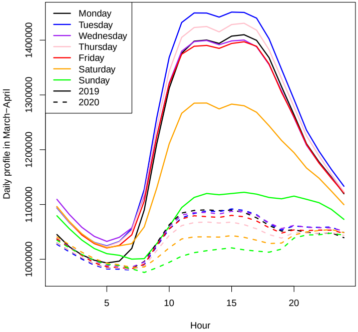
Competition from IEEE DataPort: *Day-Ahead Electricity Demand Forecasting: Post-COVID Paradigm*.

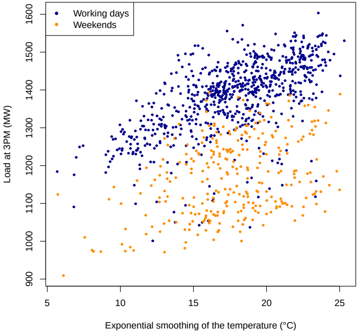
y_t: electricity load.

x_t: meteorological forecasts, calendar variables ...



Dependence to Covariates





Forecasting Methods

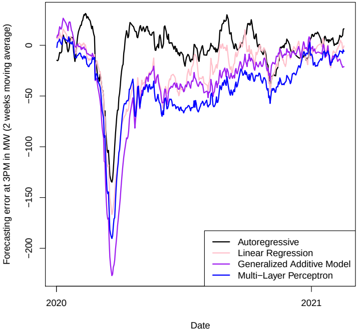
- Seasonal Auto-Regressive,
- Linear Regression,
- Generalized Additive Model:

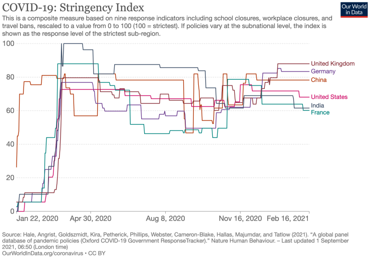
$$y_t = \alpha t + \sum_{i=1}^{6} \beta_i \mathbb{1}_{DayType_t = i} + \gamma \, Temps95_t \\ + f_1(Toy_t) + f_2(LoadD_t) + f_3(LoadW_t) + \beta_0 + \varepsilon_t \,,$$

- Multi-Layer Perceptron (2 hidden layers of 15 and 10 neurons).

All forecasting models are defined by hour of the day.

Motivation to Model Adaptation: Bias Evolution





1 Introduction

2 State-Space Representation: Kalman Filtering

3 Variance Tracking: VIKING

State-Space Model with Constant Variances

We consider the linear gaussian state-space model in the tracking mode:

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

where σ^2, Q are the hyper-parameters of the model, and x_t is defined differently for the different models. We restrict ourselves to a diagonal matrix Q.

State-Space Model with Constant Variances

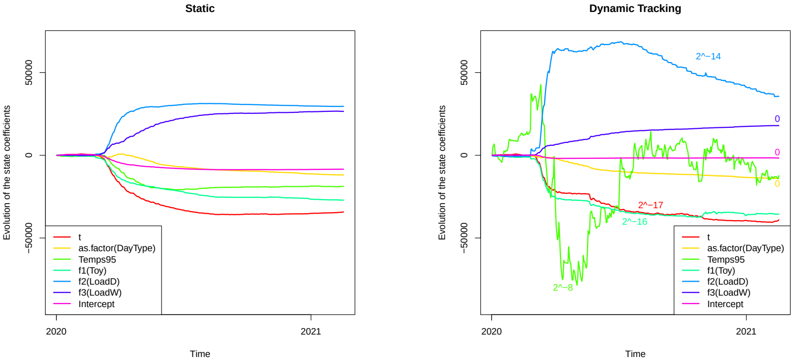
We consider the linear gaussian state-space model in the tracking mode:

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

where σ^2, Q are the hyper-parameters of the model, and x_t is defined differently for the different models. We restrict ourselves to a diagonal matrix Q.
Kalman filtering: estimation of $\theta_t \sim \mathcal{N}(\hat\theta_{t|t}, P_{t|t})$ with

$$P_{t|t-1} = P_{t-1|t-1} + Q\,, \qquad P_{t|t} = P_{t|t-1} - \frac{P_{t|t-1} x_t x_t^\top P_{t|t-1}}{x_t^\top P_{t|t-1} x_t + \sigma^2}\,,$$
$$\hat\theta_{t|t} = \hat\theta_{t-1|t-1} - \frac{P_{t|t}}{\sigma^2}\left(x_t(\hat\theta_t^\top x_t - y_t)\right).$$

Kalman Adaptation of GAM: static vs dynamic



Static: $\theta_t = \theta_{t-1}$, i.e. $Q = 0$.

Dynamic Tracking: $\theta_t = \theta_{t-1} + \eta_t$ i.e. $Q \succcurlyeq 0$.

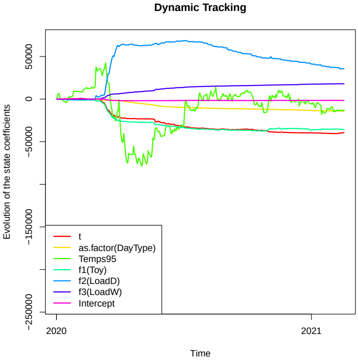
To Dynamical Variances ?

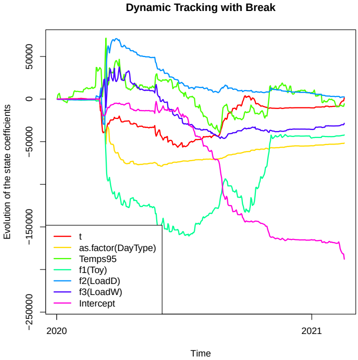
Time-varying variances:

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

First test: break at a specified time T (March 1st 2020).
Tracking with break: $\sigma_t^2 = \sigma^2, Q_t = Q$ except $Q_T \gg Q$.

Dynamic With vs Without Break





1 Introduction

2 State-Space Representation: Kalman Filtering

3 Variance Tracking: VIKING

Augmented Latent Representation

We consider time-varying variances:

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

Augmented Latent Representation

We consider time-varying variances:

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

We treat the variances σ_t^2, Q_t as other latent variables (tracking mode):

$$\sigma_t^2 = \exp(a_t)\,, \qquad Q_t = diag(\phi(b_t))\,,$$
$$a_t - a_{t-1} \sim \mathcal{N}(0, \rho_a)\,, \qquad b_t - b_{t-1} \sim \mathcal{N}(0, \rho_b I)\,.$$

Variational Bayes

Bayesian: we start from a prior $p(\theta_0, a_0, b_0)$, then at each time step t:

- **Prior**: $p(\theta_{t-1}, a_{t-1}, b_{t-1} \mid \mathcal{F}_{t-1})$,
- **Prediction step**: $p(\theta_t, a_t, b_t \mid \mathcal{F}_{t-1})$,
- **Filtering step** (Bayes rule):

$$p(\theta_t, a_t, b_t \mid \mathcal{F}_t) \propto p(x_t, y_t \mid \theta_t, a_t, b_t) p(\theta_t, a_t, b_t \mid \mathcal{F}_{t-1}) .$$

[3] Smidl and Quinn (2006): The variational Bayes method in signal processing

Variational Bayes

Bayesian: we start from a prior $p(\theta_0, a_0, b_0)$, then at each time step t:

- **Prior**: $p(\theta_{t-1}, a_{t-1}, b_{t-1} \mid \mathcal{F}_{t-1})$,
- **Prediction step**: $p(\theta_t, a_t, b_t \mid \mathcal{F}_{t-1})$,
- **Filtering step** (Bayes rule):

$$p(\theta_t, a_t, b_t \mid \mathcal{F}_t) \propto p(x_t, y_t \mid \theta_t, a_t, b_t) p(\theta_t, a_t, b_t \mid \mathcal{F}_{t-1}) .$$

Variational Bayes[3]: as the bayesian approach is intractable we estimate the posterior distribution with the best factorized distribution of the form

$$\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \mathcal{N}(\hat{a}_{t|t}, s_{t|t}) \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) .$$

[3] Smidl and Quinn (2006): The variational Bayes method in signal processing

Inference

Lemma (Posterior distribution of the Variance Tracking model)

If we have the prior

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

then the posterior distribution is expressed as follows:

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

Kullback-Leibler Divergence

We use the best factorized distribution in the sense of the Kullback-Leibler divergence: we minimize

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

where

$$KL(p \,||\, q) = \int \log\left(\frac{dp}{dq}\right) dp\,.$$

The KL doesn't have closed-form solutions and we derive upper-bounds easier to optimize.

Comparison to Kalman Filter

Theorem

Given all the other parameters, the minimum of the KL is achieved with the following:

VIKING

$$P_{t|t-1} = \mathbb{E}_{b_t}\left[\left(P_{t-1|t-1} + diag(\phi(b_t)) \right)^{-1} \right]^{-1},$$

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

$$\hat{\theta}_{t|t} = \hat{\theta}_{t-1|t-1} - \frac{P_{t|t}}{\exp(\hat{a}_{t|t} - \frac{1}{2} s_{t|t})} \left(x_t (\hat{\theta}_{t-1|t-1}^\top x_t - y_t) \right),$$

Kalman

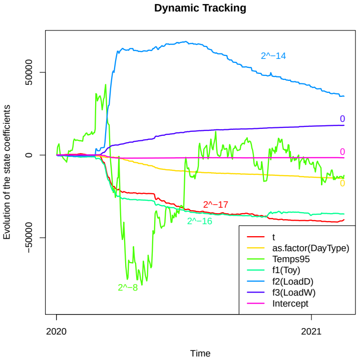
$$P_{t|t-1} = P_{t-1|t-1} + Q_t,$$

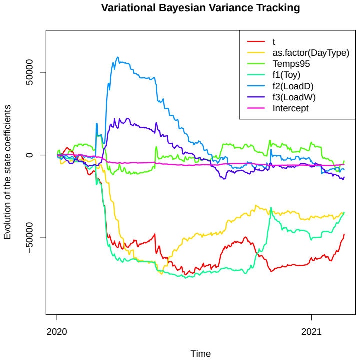
$$\square = \square - \frac{\square}{\square + \sigma_t^2},$$

$$\square = \square - \frac{\square}{\sigma_t^2} \square.$$

Jensen: if ϕ concave then $P_{t|t-1} \preccurlyeq P_{t-1|t-1} + diag(\phi(\hat{b}_{t|t}))$.

Kalman dynamique vs VIKING





Conclusion

- The method presented allows to adapt linear models, but also GAM and MLP. It yields a compromise between complex dependence to covariates and time-varying models,
- 1st place in the competition using a preliminary version of VIKING (we used aggregation of various models),
- 1st place also in *Competition on building energy consumption forecasting*: state-space methods to forecast at a much smaller scale.

https://josephdevilmarest.github.io/

de Vilmarest, J., Goude, Y. and Wintenberger, O. VIKING: Variational Bayesian Variance Tracking Winning a Post-Covid Day-Ahead Electricity Load Forecasting Competition at the Time series Workshop ICML (2021)

Generalized Additive Model

Generalized Additive Model:

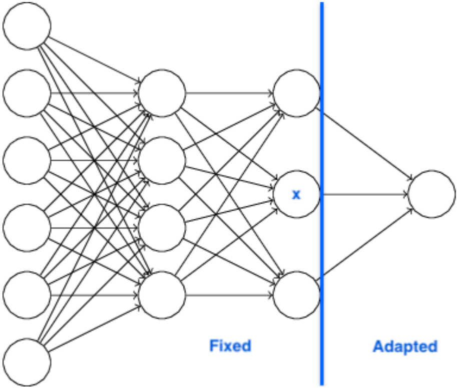
$$y_t = f_1(z_t^{(1)}) + f_2(z_t^{(2)}) + \ldots + \varepsilon_t \,.$$

Adaptive GAM:

$$\begin{aligned} y_t &= \theta_t^{(1)} f_1(z_t^{(1)}) + \theta_t^{(2)} f_2(z_t^{(2)}) + \ldots + \varepsilon_t \\ &= \theta_t^\top f(z_t) + \varepsilon_t \,. \end{aligned}$$

- The effects are fixed (f does not depend on t).
- Adaptation of a linear combination of the effects (θ_t depends on t).

Multi-Layer Perceptron



- Deepest layers fixed,
- Adaptation of the last layer.