State-space models for time series forecasting. Application to the electricity markets.

Joseph de Vilmarest

PhD Defense: June 22, 2022

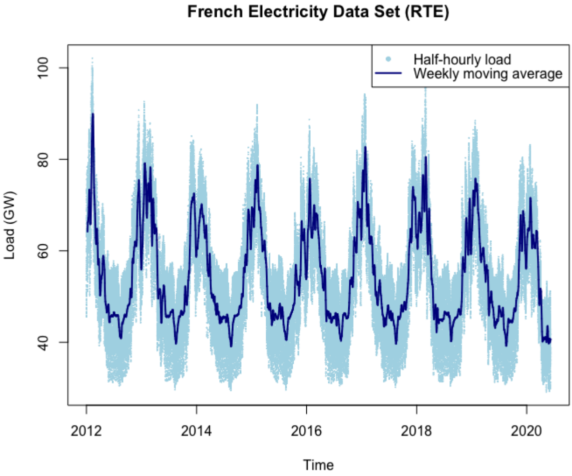
PhD Supervisor: Olivier Wintenberger
Industrial Advisors: Yannig Goude, Thi Thu Huong Hoang

SORBONNE UNIVERSITÉ

EDF

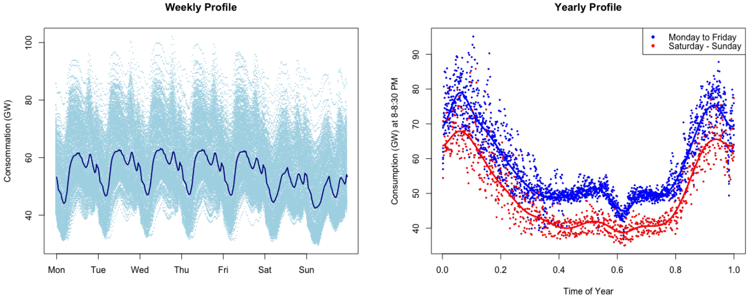
Time Series Forecasting

We aim at forecasting $y_t \in \mathbb{R}$.
Main application of the PhD: electricity load.

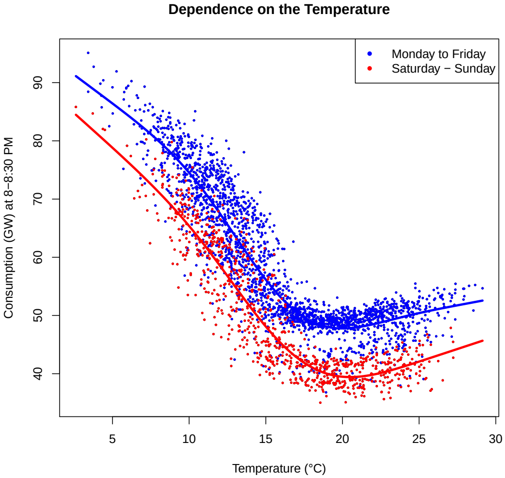


Explanatory Variables: Calendar

Explanatory variables: $x_t \in \mathbb{R}^d$.



Explanatory Variables: Temperature



Forecasting Objective

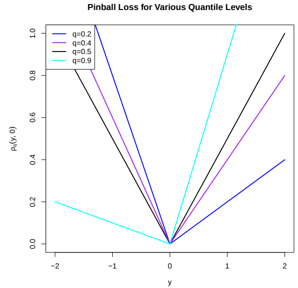
The objective is to forecast y_t given x_t. In what sense ?

- Mean forecasting: estimation of $\mathbb{E}[y_t \mid x_t]$.
  It is the minimum of $\mathbb{E}[(y_t - \hat{y}_t)^2 \mid x_t]$.

Forecasting Objective

The objective is to forecast y_t given x_t. In what sense ?

- Mean forecasting: estimation of $\mathbb{E}[y_t \mid x_t]$.
  It is the minimum of $\mathbb{E}[(y_t - \hat{y}_t)^2 \mid x_t]$.
- Probabilistic forecasting: estimation of $\mathcal{L}(y_t \mid x_t)$.
  For a certain quantile level q we forecast $\hat{y}_{t,q}$ such that $\mathbb{P}(y_t \leq \hat{y}_{t,q} \mid x_t) = q$.
  It is equivalent to minimize $\mathbb{E}[\rho_q(y_t, \hat{y}_t) \mid x_t]$:



Offline vs Online

- **Offline**: $\hat{y}_t = f_{\hat{\theta}}(x_t)$.
  Example: Empirical Risk Minimizer

$$\hat{\theta} \in \arg\min \sum_{t \in \mathcal{T}} \ell(y_t, f_{\hat{\theta}}(x_t))$$

Offline vs Online

- **Offline**: $\hat{y}_t = f_{\hat{\theta}}(x_t)$.
  Example: Empirical Risk Minimizer

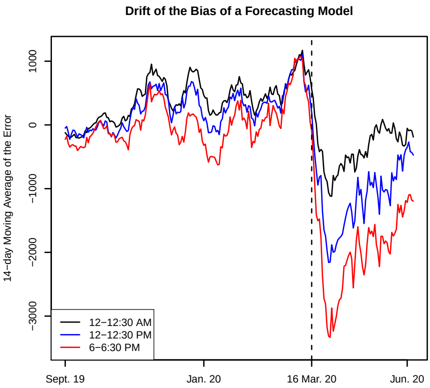
$$\hat{\theta} \in \arg\min \sum_{t \in \mathcal{T}} \ell(y_t, f_{\hat{\theta}}(x_t))$$

- **Online / Adaptive**: $\hat{y}_t = f_{\hat{\theta}_t}(x_t)$ with $\hat{\theta}_{t+1} = \Phi(\hat{\theta}_t, x_t, y_t)$.
  Example: Online Gradient Descent

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \gamma_t \frac{\partial \ell(y_t, f_\theta(x_t))}{\partial \theta}\Big|_{\hat{\theta}_t}$$

Drift of Offline Models

Train set: from January 2012 to September 2019.
Test set: from September 2019 to June 2020.



Tracking State-Space Model

State: $\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,,$

Space: $y_t \sim p_{\theta_t}(\cdot \mid x_t)\,.$

Tracking State-Space Model

$$\text{State:} \qquad \theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,,$$
$$\text{Space:} \qquad y_t \sim p_{\theta_t}(\cdot \mid x_t)\,.$$

Two main models in the PhD:

- Linear Gaussian: $y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma_t^2)$.
- Logistic Regression: $y_t \mid x_t \sim \mathcal{B}\left(\frac{1}{1+e^{-\theta_t^\top x_t}}\right)$.

Tracking State-Space Model

$$\text{State:} \qquad \theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,,$$
$$\text{Space:} \qquad y_t \sim p_{\theta_t}(\cdot \mid x_t)\,.$$

Two main models in the PhD:

- Linear Gaussian: $y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma_t^2)$.
- Logistic Regression: $y_t \mid x_t \sim \mathcal{B}\left(\frac{1}{1+e^{-\theta_t^\top x_t}}\right)$.

Bayesian approach, starting from $\theta_1 \sim \mathcal{N}(\hat\theta_1, P_1)$:

$$\hat\theta_t = \hat\theta_{t|t-1} = \mathbb{E}[\theta_t \mid x_1, y_1, \dots, x_{t-1}, y_{t-1}]\,,$$
$$P_t = P_{t|t-1} = \mathbb{E}[(\theta_t - \hat\theta_{t|t-1})(\theta_t - \hat\theta_{t|t-1})^\top \mid x_1, y_1, \dots, x_{t-1}, y_{t-1}]\,.$$

Linear Gaussian State-space Model

State: $\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,,$

Space: $y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma_t^2)\,.$

Theorem (R. Kalman and R. Bucy, 1961)

Under the state-space assumption with known variances, and if $\theta_1 \sim \mathcal{N}(\hat\theta_1, P_1)$, *it holds* $\theta_{t+1} \mid (x_s, y_s)_{s\le t} \sim \mathcal{N}(\hat\theta_{t+1}, P_{t+1})$ *with*

$$P_{t|t} = P_t - \frac{P_t x_t x_t^\top P_t}{x_t^\top P_t x_t + \sigma_t^2}\,, \qquad P_{t+1} = P_{t|t} + Q_{t+1}\,,$$

$$\hat\theta_{t+1} = \hat\theta_t - \frac{P_{t|t}}{\sigma_t^2}\Big(x_t(\hat\theta_t^\top x_t - y_t)\Big)\,.$$

Summary of the PhD

Gradient interpretation of Bayesian algorithms in state-space models:

$$\hat{\theta}_{t+1} = \hat{\theta}_t - P_{t|t} \frac{\partial \ell(y_t, f_\theta(x_t))}{\partial \theta}\Big|_{\hat{\theta}_t},$$

where $\ell(y, \theta^\top x) = -\log p_\theta(y \mid x)$.

Summary of the PhD

Gradient interpretation of Bayesian algorithms in state-space models:

$$\hat{\theta}_{t+1} = \hat{\theta}_t - P_{t|t} \frac{\partial \ell(y_t, f_\theta(x_t))}{\partial \theta}\Big|_{\hat{\theta}_t},$$

where $\ell(y, \theta^\top x) = -\log p_\theta(y \mid x)$.

- Part I. Analysis of the static setting $(\theta_t = \theta_{t-1})$. *Publication in Journal of Machine Learning Research*.
- Part II, Chapter 5. Choice of the time-invariant covariance matrix Q in $\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q)$.
- Part II, Chapter 6. *Variational Bayesian Variance Tracking*: adaptive estimation of Q_t in $\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)$. *Submitted*.

Part III. Application to electricity load forecasting.
Publications in IEEE Journal of Power Systems and IEEE Open Access Journal of Power and Energy.

State-Space for Generalized Linear Models (GLM)[1]

State:	$\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,,$
Space:	$y_t \sim p_{\theta_t}(\cdot \mid x_t)\,,$

The distributions are in a subclass of the exponential family:

$$p_\theta(y \mid x) = h(y)\exp\left(\frac{y\theta^\top x - b(\theta^\top x)}{a}\right),$$

with $a > 0$ and b, h univariate functions.

[1] P. McCullagh and J. A. Nelder, 1989

State-Space for Generalized Linear Models (GLM)[1]

State: $\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,,$

Space: $y_t \sim p_{\theta_t}(\cdot \mid x_t)\,,$

The distributions are in a subclass of the exponential family:

$$p_\theta(y \mid x) = h(y) \exp\left(\frac{y\theta^\top x - b(\theta^\top x)}{a}\right),$$

with $a > 0$ and b, h univariate functions.

Example (Logistic Regression)

$y \in \{-1, 1\}$ and

$$p_\theta(y \mid x) = \frac{1}{1 + e^{-y\theta^\top x}} = \exp\left(\frac{y\theta^\top x - (2\log(1 + e^{\theta^\top x}) - \theta^\top x)}{2}\right)$$

[1] P. McCullagh and J. A. Nelder, 1989

Analytical Form of the First Two Moments

Proposition

GLM distributions satisfy:

$$\mathbb{E}_\theta[y \mid x] = b'(\theta^\top x)\,, \quad Var_\theta[y \mid x] = ab''(\theta^\top x)\,.$$

Analytical Form of the First Two Moments

Proposition

GLM distributions satisfy:

$$\mathbb{E}_\theta[y \mid x] = b'(\theta^\top x)\,, \quad Var_\theta[y \mid x] = ab''(\theta^\top x)\,.$$

Weaker state-space model:

$$\begin{aligned} &\text{State:} && \theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,, \\ &\text{Space:} && y_t = b'(\theta_t^\top x_t) + \varepsilon_t\,, \end{aligned}$$

where ε_t is a centered noise of variance $ab''(\theta_t^\top x_t)$.

Analytical Form of the First Two Moments

Proposition

GLM distributions satisfy:

$$\mathbb{E}_\theta[y \mid x] = b'(\theta^\top x)\,, \quad Var_\theta[y \mid x] = ab''(\theta^\top x)\,.$$

Weaker state-space model:

$$\begin{aligned} &\text{State:} && \theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,, \\ &\text{Space:} && y_t = b'(\theta_t^\top x_t) + \varepsilon_t\,, \end{aligned}$$

where ε_t is a centered noise of variance $ab''(\theta_t^\top x_t)$.

Linear approximation of the space equation:

$$\begin{aligned} y_t &= b'(\theta_t^\top x_t) + \varepsilon_t \\ &\approx b'(\hat{\theta}_t^\top x_t) + b''(\hat{\theta}_t^\top x_t)(\theta_t - \hat{\theta}_t)^\top x_t + \varepsilon_t\,. \end{aligned}$$

Static Extended Kalman Filter

Proposition (Extended Kalman Filter as a Gradient Descent)

The EKF is equivalent to the following recursion:

$$P_{t|t}^{-1} = P_t^{-1} + \ell''(y_t, \hat{\theta}_t^\top x_t) x_t x_t^\top \,,$$

$$\hat{\theta}_{t+1} = \hat{\theta}_t - P_{t|t} \left(\ell'(y_t, \hat{\theta}_t^\top x_t) x_t \right) \,,$$

$$P_{t+1} = P_{t|t} + Q_{t+1} \,,$$

where $\ell(y, \theta^\top x) = -\log p_\theta(y \mid x)$.

In the static setting ($Q_{t+1} = 0$):

- Correspondence established by Y. Ollivier (2018).
- Also referred to as Stochastic Newton (B. Bercu et al., 2019).
- $P_{t|t} \approx H^{\star -1}/t$.

Misspecified Static Setting

The model $y_t \sim p_\theta(\cdot \mid x_t)$ allows to derive the EKF. However, in our analysis we don't assume that the data-generating process is the GLM.

Two standard assumptions on the data:

- (x_t, y_t) is i.i.d.
- We define $L(\theta) = \mathbb{E}[\ell(y, \theta^\top x)]$.
  There exists $\theta^\star$ such that $L(\theta^\star) = \inf_\theta L(\theta)$.
  $H^\star$ is the hessian matrix of the risk at the optimum.

1. Parallel with Online Newton Step (ONS)

The ONS is defined, for Θ and γ, by

$$
\begin{aligned}
P_{t+1}^{-1} &= P_t^{-1} + \ell'(y_t, \hat{\theta}_t^\top x_t)^2 x_t x_t^\top \,, \\
\hat{\theta}_{t+1} &= \Pi_\Theta\Big(\hat{\theta}_t - \gamma P_{t+1}\left(\ell'(y_t, \hat{\theta}_t^\top x_t) x_t\right)\Big) \,.
\end{aligned}
$$

1. Parallel with Online Newton Step (ONS)

The ONS is defined, for Θ and γ, by

$$P_{t+1}^{-1} = P_t^{-1} + \ell'(y_t, \hat{\theta}_t^\top x_t)^2 x_t x_t^\top \,,$$
$$\hat{\theta}_{t+1} = \Pi_\Theta\Big(\hat{\theta}_t - \gamma P_{t+1}\left(\ell'(y_t, \hat{\theta}_t^\top x_t) x_t\right)\Big)\,.$$

Theorem (M. Mahdavi, L. Zhang and R. Jin, 2015)

If (x_t, y_t) is i.i.d., $\theta^\star \in \Theta$ and ℓ is $1/\kappa$-exp-concave in Θ ($\ell'' \geq (1/\kappa)\ell'^2$), for any $\delta > 0$ it holds with probability $1-\delta$ that simultaneously for $n \geq 1$:

$$\sum_{t=1}^{n} (L(\hat{\theta}_t) - L(\theta^\star)) = O\Big(\kappa(d \log n + \log \delta^{-1})\Big)\,.$$

Logistic setting: $\kappa = \exp\left(\max_{\theta\in\Theta,t}(\theta^\top x_t)\right)$.

Our objective: $\exp\left(\max_t(\theta^{\star\top} x_t)\right)$ while removing the projection step.

2. Asymptotic Result for Logistic Regression (Truncated)

We consider the following modification of the algorithm for $0 < \beta < \frac{1}{2}$:

$$P_{t+1}^{-1} = P_t^{-1} + \max\left(\ell''(y_t, \hat{\theta}_t^\top x_t), \frac{1}{t^\beta}\right) x_t x_t^\top ,$$
$$\hat{\theta}_{t+1} = \hat{\theta}_t - P_{t+1}\left(\ell'(y_t, \hat{\theta}_t^\top x_t) x_t\right) .$$

Theorem (B. Bercu, A. Godichon and B. Portier, 2019)

Under the previous assumptions, in the logistic setting, we have

$$\|\frac{1}{t} P_t^{-1} - H^\star\|^2 = O\left(\frac{1}{t^{2\beta}}\right) \text{ a.s.}$$
$$\|\hat{\theta}_t - \theta^\star\|^2 = O\left(\frac{\log t}{t}\right) \text{ a.s.}$$

($H^\star$ is the hessian matrix of the risk at the optimum).

Structure of the Analysis

1. Localized Analysis. Tight bound on the cumulative excess risk under a strong convergence assumption. Similar as the analysis of the ONS.
2. Proof of the convergence in the logistic setting, using the truncated algorithm of B. Bercu et al. (2019).

1. Localized Analysis. Assumptions

Assumption (Localized Assumption)

We set $\varepsilon > 0$. For any $\delta > 0$, there exists $T(\varepsilon, \delta) \in \mathbb{N}$ such that with probability $1 - \delta$,

$$\forall t > T(\varepsilon, \delta), \quad \|\hat{\theta}_t - \theta^\star\| \leq \varepsilon.$$

1. Localized Analysis. Assumptions

Assumption (Localized Assumption)

We set $\varepsilon > 0$. For any $\delta > 0$, there exists $T(\varepsilon, \delta) \in \mathbb{N}$ such that with probability $1 - \delta$,

$$\forall t > T(\varepsilon, \delta), \quad \|\hat{\theta}_t - \theta^\star\| \leq \varepsilon.$$

We assume that for some $\varepsilon > 0$ and $\theta, \theta_0 \in \mathcal{B}^{\varepsilon}_{\theta^\star}$,

- $\ell'(y, \theta^\top x)^2 \leq \kappa_\varepsilon \ell''(y, \theta^\top x)$ a.s. for some $\kappa_\varepsilon > 0$.
- $0 \leq \ell''(y, \theta^\top x) \leq h_\varepsilon$ a.s. for some $h_\varepsilon > 0$.
- $\ell''(y, \theta^\top x) \geq \rho_\varepsilon \ell''(y, \theta_0^\top x)$ a.s. for some $\rho_\varepsilon > 0.95$.

1. Localized Analysis. Assumptions

Assumption (Localized Assumption)

We set $\varepsilon > 0$*. For any* $\delta > 0$*, there exists* $T(\varepsilon, \delta) \in \mathbb{N}$ *such that with probability* $1 - \delta$*,*

$$\forall t > T(\varepsilon, \delta), \quad \|\hat{\theta}_t - \theta^\star\| \leq \varepsilon.$$

We assume that for some $\varepsilon > 0$ and $\theta, \theta_0 \in \mathcal{B}^{\varepsilon}_{\theta^\star}$,

- $\ell'(y, \theta^\top x)^2 \leq \kappa_\varepsilon \ell''(y, \theta^\top x)$ a.s. for some $\kappa_\varepsilon > 0$.
- $0 \leq \ell''(y, \theta^\top x) \leq h_\varepsilon$ a.s. for some $h_\varepsilon > 0$.
- $\ell''(y, \theta^\top x) \geq \rho_\varepsilon \ell''(y, \theta_0^\top x)$ a.s. for some $\rho_\varepsilon > 0.95$.

Example (Logistic Regression)

In the logistic setting, it holds with $\kappa_\varepsilon = e^{D_X(\|\theta^\star\|+\varepsilon)}$, $h_\varepsilon = \frac{1}{4}$, $\rho_\varepsilon = e^{-\varepsilon D_X}$.

Remark. We handle the quadratic loss with specific assumptions.

1. Localized Analysis. Result

Theorem

Under the previous assumptions, for any $\delta > 0$, it holds with probability at least $1 - 3\delta$ that simultaneously for any $n \geq 1$

$$\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} (L(\hat{\theta}_t) - L(\theta^\star)) = O\Big(\kappa_\varepsilon(d \ln n + \ln \delta^{-1})\Big) \, .$$

We obtain the upper-bound on the ONS with $\Theta = \mathcal{B}^{\varepsilon}_{\theta^\star}$ and optimal exp-concavity constant.

1. Localized Analysis. Sketch of Proof

- Adversarial analysis close to E. Hazan et al. (2007): for any $n \in \mathbb{N}$,

$$\sum_{t=1}^{n} \left(\left(\ell'(y_t, \hat{\theta}_t^\top x_t) x_t \right)^\top (\hat{\theta}_t - \theta^\star) - \frac{1}{2} (\hat{\theta}_t - \theta^\star)^\top \left(\ell''(y_t, \hat{\theta}_t^\top x_t) x_t x_t^\top \right) (\hat{\theta}_t - \theta^\star) \right)$$
$$= O\Big(\kappa_\varepsilon d \ln n \Big) .$$

1. Localized Analysis. Sketch of Proof

- Adversarial analysis close to E. Hazan et al. (2007): for any $n \in \mathbb{N}$,

$$\sum_{t=1}^{n} \left(\left(\ell'(y_t, \hat{\theta}_t^\top x_t) x_t \right)^\top (\hat{\theta}_t - \theta^\star) - \frac{1}{2} (\hat{\theta}_t - \theta^\star)^\top \left(\ell''(y_t, \hat{\theta}_t^\top x_t) x_t x_t^\top \right) (\hat{\theta}_t - \theta^\star) \right) = O\Big(\kappa_\varepsilon d \ln n \Big) .$$

- For any $\theta \in \mathcal{B}_{\theta^\star}^{\varepsilon}$ and $0 < c < \rho_\varepsilon$, it holds

$$L(\theta) - L(\theta^\star) \leq \frac{\rho_\varepsilon}{\rho_\varepsilon - c} \left(\frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star) - c(\theta - \theta^\star)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_\theta (\theta - \theta^\star) \right) .$$

1. Localized Analysis. Sketch of Proof

- Adversarial analysis close to E. Hazan et al. (2007): for any $n \in \mathbb{N}$,

$$\sum_{t=1}^{n} \left(\left(\ell'(y_t, \hat{\theta}_t^\top x_t) x_t \right)^\top (\hat{\theta}_t - \theta^\star) - \frac{1}{2} (\hat{\theta}_t - \theta^\star)^\top \left(\ell''(y_t, \hat{\theta}_t^\top x_t) x_t x_t^\top \right) (\hat{\theta}_t - \theta^\star) \right) = O\Big(\kappa_\varepsilon d \ln n \Big) .$$

- For any $\theta \in \mathcal{B}_{\theta^\star}^{\varepsilon}$ and $0 < c < \rho_\varepsilon$, it holds

$$L(\theta) - L(\theta^\star) \leq \frac{\rho_\varepsilon}{\rho_\varepsilon - c} \left(\frac{\partial L}{\partial \theta}\Big|_\theta^\top (\theta - \theta^\star) - c(\theta - \theta^\star)^\top \frac{\partial^2 L}{\partial \theta^2}\Big|_\theta (\theta - \theta^\star) \right) .$$

- Martingale analysis relying on B. Bercu and A. Touati (2008) and D. Freedman (1975).

2. Logistic Regression. Truncated Algorithm

We remind that $y \in \{-1, 1\}$ and $p_\theta(y \mid x) = \frac{1}{1+e^{-y\theta^\top x}}$.
The truncated algorithm for $0 < \beta < \frac{1}{2}$ (B. Bercu et al., 2019) is the following:

$$P_{t+1}^{-1} = P_t^{-1} + \max\left(\frac{1}{(1+e^{\hat{\theta}_t^\top x_t})(1+e^{-\hat{\theta}_t^\top x_t})}, \frac{1}{t^\beta}\right) x_t x_t^\top ,$$

$$\hat{\theta}_{t+1} = \hat{\theta}_t - P_{t+1}\left(\frac{-y_t x_t}{1+e^{y_t \hat{\theta}_t^\top x_t}}\right) .$$

2. Logistic Regression. Convergence Result

One last assumption: $\mathbb{E}[xx^\top]$ is invertible.

2. Logistic Regression. Convergence Result

One last assumption: $\mathbb{E}[xx^\top]$ is invertible.

Theorem

Under the previous assumptions, it holds

$$\forall t > T(\varepsilon, \delta), \qquad \|\hat{\theta}_t - \theta^\star\| \leq \varepsilon\,, \quad \frac{1}{t^\beta} \leq \frac{1}{(1 + e^{\hat{\theta}_t^\top x_t})(1 + e^{-\hat{\theta}_t^\top x_t})}\,,$$

with probability at least $1 - \delta$, where $T(\varepsilon, \delta) \in \mathbb{N}$ is explicitly defined.

2. Logistic Regression. Sketch of Proof

- Thanks to the truncation:

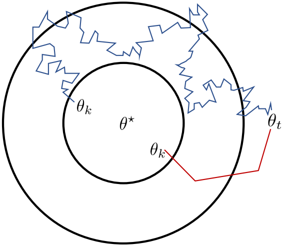
$$\underbrace{\frac{c_1}{t} I \preccurlyeq}_{a.s.} P_t \underbrace{\preccurlyeq \frac{c_2}{t^{1-\beta}} I}_{w.h.p.} .$$

2. Logistic Regression. Sketch of Proof

- Thanks to the truncation:

$$\underbrace{\frac{c_1}{t} I \preccurlyeq}_{a.s.} P_t \underbrace{\preccurlyeq \frac{c_2}{t^{1-\beta}} I}_{w.h.p.} .$$

- Analysis seen as a non-asymptotic Robbins-Siegmund theorem.



2. Logistic Regression. Global Result

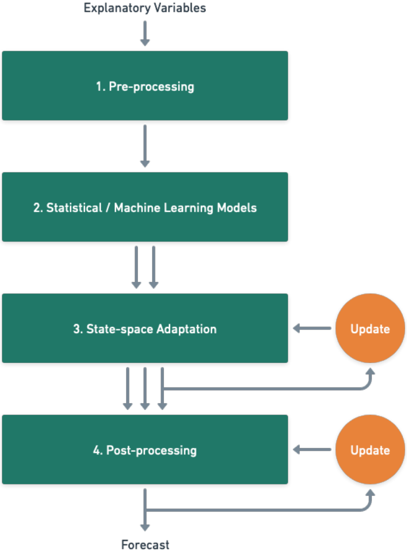
Corollary

Under the previous assumptions, for any $\varepsilon, \delta > 0$, it holds with probability at least $1 - 4\delta$ that simultaneously for any $n \geq 1$:

$$\sum_{t=1}^{n} (L(\hat{\theta}_t) - L(\theta^\star)) = O\Big(\kappa_\varepsilon (d \ln n + \ln \delta^{-1})\Big) + \sum_{t=1}^{T(\varepsilon,\delta)} (L(\hat{\theta}_t) - L(\theta^\star)) \,.$$

Applications

- Confidential data at EDF.
- Chapter 7 (joint work with D. Obst). French national load.
- Chapter 8. Competition at a city level. 1st place.
- Chapter 9. Competition at a building level. 1st place.
- Chapter 10 (ongoing work with J. Browell and M. Fasiolo). Probabilistic forecast. Electricity *net*-load in Great-Britain and load in big US cities.
- M6 Financial Forecasting Competition (with N. Werge). Probabilistic ranking. 2nd place in forecasting in the 1st quarter.

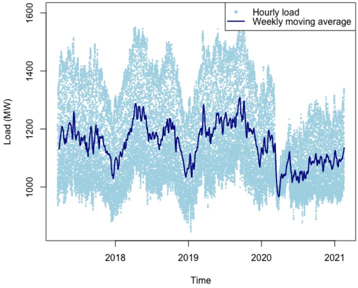


Competition: Load Forecasting at a City-Wide Level

Day-Ahead Electricity Demand Forecasting: Post-COVID Paradigm[2]

y_t: electricity load.

x_t: meteorological forecasts, calendar variables ...



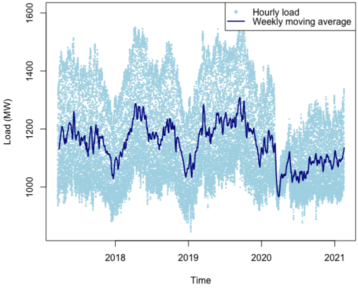
[2] M. Farrokhabadi, J. Browell, S. Makonin, W. Su and H. Zareipour, 2022

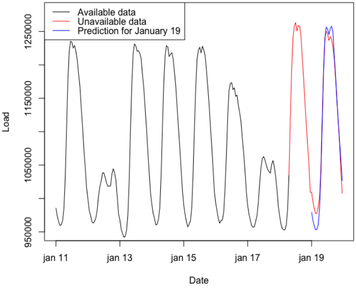
Competition: Load Forecasting at a City-Wide Level

Day-Ahead Electricity Demand Forecasting: Post-COVID Paradigm[2]

y_t: electricity load.

x_t: meteorological forecasts, calendar variables ...

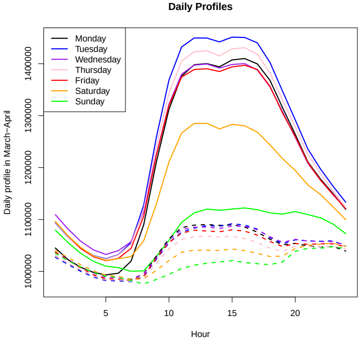


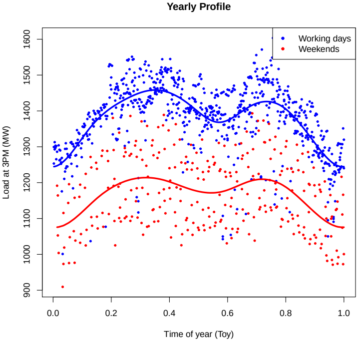


30 consecutive days: forecast the hourly load of next day.

[2] M. Farrokhabadi, J. Browell, S. Makonin, W. Su and H. Zareipour, 2022

Dependence on Calendar Variables

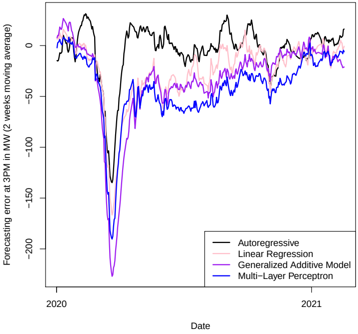




Offline Methods

We define forecasting models by hour of the day.

- Seasonal Auto-Regressive Model: $y_t = \sum_{l \in \mathcal{L}} \alpha_l y_{t-l} + \varepsilon_t$.
- Linear Regression: $y_t = \theta^\top x_t + \varepsilon_t$.
- Generalized Additive Model: $y_t = \sum_{j=1}^{d} f_j(x_{t,j}) + \varepsilon_t$ where the effects f_j are decomposed on spline bases.
- Small Multi-Layer Perceptron (2 hidden layers of 15 and 10 neurons).



State-Space Model with Time-Invariant Variances

State: $\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q)\,,$

Space: $y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma^2)\,.$

[2] Chapter 5

State-Space Model with Time-Invariant Variances

$$\text{State:} \qquad \theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q)\,,$$
$$\text{Space:} \qquad y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma^2)\,.$$

We optimize the log-likelihood with respect to $\Theta = (\hat{\theta}_1, P_1, \sigma^2, Q)$:

$$\ln p(x_{1:n}, y_{1:n} \mid \Theta) = \sum_{t=1}^{n} \ln p(x_t, y_t \mid x_{1:(t-1)}, y_{1:(t-1)}, \Theta)\,.$$

[2] Chapter 5

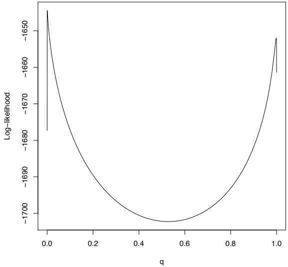
State-Space Model with Time-Invariant Variances

$$\text{State:} \quad \theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q)\,,$$
$$\text{Space:} \quad y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma^2)\,.$$

We optimize the log-likelihood with respect to $\Theta = (\hat{\theta}_1, P_1, \sigma^2, Q)$:

$$\ln p(x_{1:n}, y_{1:n} \mid \Theta) = \sum_{t=1}^{n} \ln p(x_t, y_t \mid x_{1:(t-1)}, y_{1:(t-1)}, \Theta)\,.$$

- Non-convex log-likelihood. No guarantee of global optimality.
- We restrict to a diagonal Q. Coefficient optimized using an *iterative grid search*[2].



[2] Chapter 5

Definition of x_t

The vector x_t is defined by the model we need to adapt:

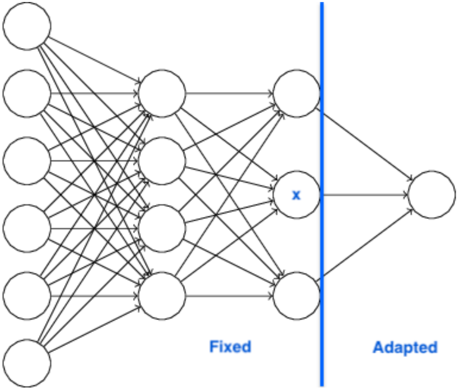
- Linear Regression: x_t is the covariate vector.
- SAR: x_t is composed of the different lags of the AR model.
- Generalized Additive Model:

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

Adaptive GAM:

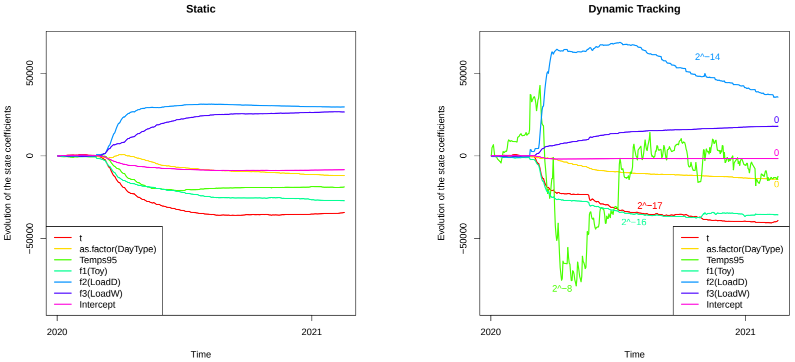
$$\begin{aligned} y_t &= \theta_t^{(1)} f_1(z_t^{(1)}) + \theta_t^{(2)} f_2(z_t^{(2)}) + \ldots + \varepsilon_t \\ &= \theta_t^\top \underbrace{f(z_t)}_{x_t} + \varepsilon_t \,. \end{aligned}$$

Multi-Layer Perceptron



- Deepest layers are fixed,
- We adapt only the last (linear) layer.

Kalman Adaptation of GAM: Static vs Dynamic



Static: $Q = 0$ and "*gradient step* $= O(1/t)$".
Dynamic Tracking: $Q \succcurlyeq 0$ and "*gradient step* $= O(1)$".

Time-Varying Variances

State: $\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t)\,,$

Space: $y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma_t^2)\,.$

[3] V. Smidl and A. Quinn, 2006

Time-Varying Variances

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

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

$$
\sigma_t^2 = \exp(a_t)\,, \qquad Q_t = diag(\phi(b_t))\,,
$$

$$
a_t - a_{t-1} \sim \mathcal{N}(0, \rho_a)\,, \qquad b_t - b_{t-1} \sim \mathcal{N}(0, \rho_b I)\,.
$$

[3] V. Smidl and A. Quinn, 2006

Time-Varying Variances

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

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

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

Inference relies on the variational Bayes approach[3]. 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] V. Smidl and A. Quinn, 2006

Comparison to Kalman Filter

Theorem

Given all the other parameters, the minimum of the KL is achieved with the following[4]*:*

Viking

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

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

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

Kalman

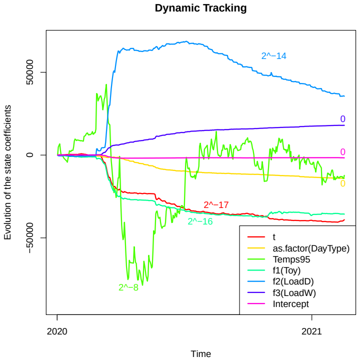
$$P_t = P_{t-1|t-1} + Q_t\,,$$

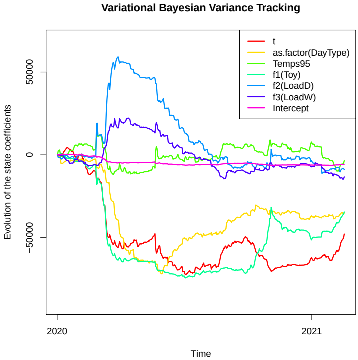
$$\square = \square - \frac{\square}{\square + \sigma_t^2},$$

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

[4] Chapter 6

Kalman Dynamic vs Viking





Conclusion

- Inference algorithms for state-space models (Kalman filter, Viking) are similar to gradient algorithms.
- The estimation of the variances is still a challenging issue where the best method depends on the application considered.
- State-space models capture well the evolution of the electricity load in various countries, scales, and tasks.

Our result on the final iterate of an averaged SGD (annealing step size):

$$L(\overline{\theta}_n) - L(\theta^\star) \leq \frac{16 g^2 \ln \delta^{-1}}{\mu_\varepsilon n} + \frac{1}{n} \underbrace{\sum_{t=1}^{k} (L(\theta_t) - L(\theta^\star))}_{O((g^8 (\ln \delta^{-1})^2)/(\mu_\varepsilon^2 \varepsilon^6))} .$$

Related work:

- Optimal bound for the Empirical Risk Minimizer[5]:

$$L(\hat{\theta}_n) - L(\theta^\star) = O\Big(\frac{tr(G^\star H^{\star -1}) \ln \delta^{-1}}{n}\Big) .$$

- Result in expectation[6]:

$$\mathbb{E}[\|\overline{\theta}_n - \theta^\star\|^2] \leq \frac{tr(\Sigma^\star)}{n} + \frac{C}{n^{5/4}} .$$

  Also results in higher orders.

---

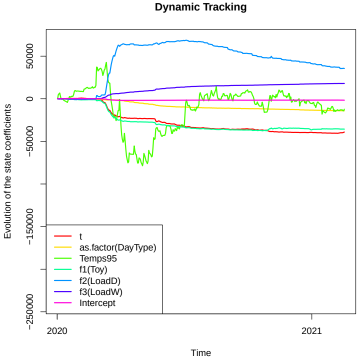
[5] D. Ostrovskii and F. Bach, 2021

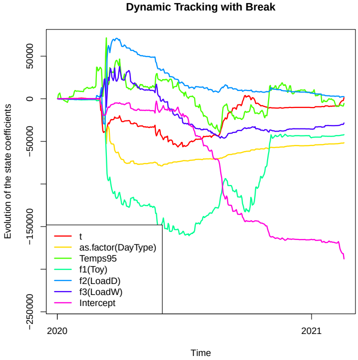
[6] S. Gadat and F. Panloup, 2017

Leads for Variance Estimation

- **Time-invariant** (chapter 5):
  Better non-convex optimization algorithm.
  Structure of Q (diagonal in *iterative grid search*). $Q = UDU^\top$?
- **Time-varying** (chapter 6):
  Structure of Q_t with sparsity.
  Q_t, σ_t^2 dependent. For instance Q_t/σ_t^2 and σ_t^2 independent.

Break





Data set: city-wide competition.
Break: $Q_t = Q$ except $Q_T \gg Q$.

Time-Varying Variances

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

We estimate $p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_{t-1})$. We assume

$$p(\theta_t, \sigma_t^2, Q_t \mid \theta_{t-1}, \sigma_{t-1}^2, Q_{t-1}) = \mathcal{N}(\theta_t - \theta_{t-1} \mid 0, Q_t)p(\sigma_t^2, Q_t \mid \sigma_{t-1}^2, Q_{t-1})\,.$$

Bayesian approach: at each step,

- **Prior**: $p(\theta_{t-1}, \sigma_{t-1}^2, Q_{t-1} \mid \mathcal{F}_{t-1})$,
- **Prediction**: $p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_{t-1})$,
- **Filtering** (Bayes rule):

$$p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_t) \propto p(x_t, y_t \mid \theta_t, \sigma_t^2, Q_t)p(\theta_t, \sigma_t^2, Q_t \mid \mathcal{F}_{t-1})\,.$$

We propagate:

$$p(\theta_{t-1}, \sigma^2_{t-1}, Q_{t-1} \mid \mathcal{F}_{t-1}) \\ = \mathcal{N}(\theta_{t-1} \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1}) p_{\Phi_{t-1|t-1}}(\sigma^2_{t-1}) p_{\Psi_{t-1|t-1}}(Q_{t-1}),$$

where $\Phi_{t-1|t-1}, \Psi_{t-1|t-1}$ parametrize distributions for σ^2_{t-1}, Q_{t-1}. With the appropriate transition $p(\sigma^2_t, Q_t \mid \sigma^2_{t-1}, Q_{t-1})$ we obtain:

$$p(\theta_t, \sigma^2_t, Q_t \mid \mathcal{F}_{t-1}) \approx \mathcal{N}(\theta_t \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1} + Q_t) \\ p_{\Phi_{t|t-1}}(\sigma^2_t) p_{\Psi_{t|t-1}}(Q_t).$$

A posteriori distribution:

$$p(\theta_t, \sigma^2_t, Q_t \mid \mathcal{F}_t) = \frac{p(x_t, \mathcal{F}_{t-1})}{p(\mathcal{F}_t)} \mathcal{N}(y_t \mid {\theta_t}^\top x_t, \sigma^2_t) \\ \mathcal{N}(\theta_t \mid \hat{\theta}_{t-1|t-1}, P_{t-1|t-1} + Q_t) p_{\Phi_{t|t-1}}(\sigma^2_t) p_{\Psi_{t|t-1}}(Q_t).$$

Variance Tracking

Auxiliary latent variables a_t, b_t such that $\sigma_t^2 = \exp(a_t), Q_t = f(b_t)$.

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

A posteriori distribution estimated by the minimum of

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

Kullback-Leibler Divergence

There exists 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

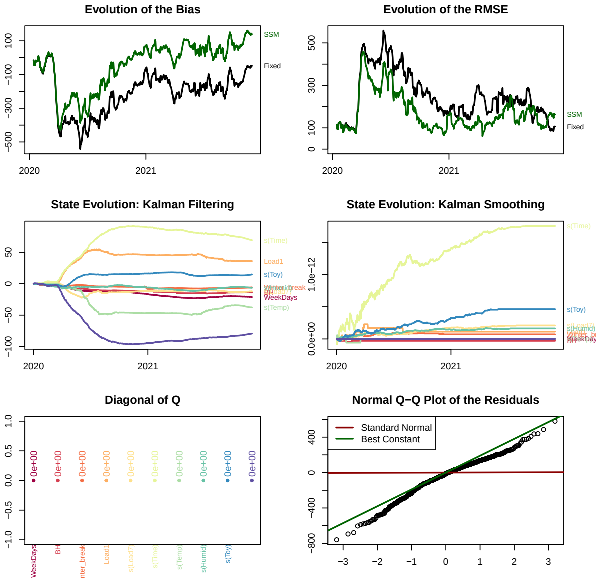
$$KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t}) \times \mathcal{N}(\hat{a}_{t|t}, s_{t|t}) \times \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; P_{\mathcal{F}_t}\Big) = -\frac{1}{2}\log\det P_{t|t} - \frac{1}{2}\log s_{t|t}$$
$$- \frac{1}{2}\log\det \Sigma_{t|t} + \frac{1}{2}((y_t - \hat{\theta}_{t|t}{}^\top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

with

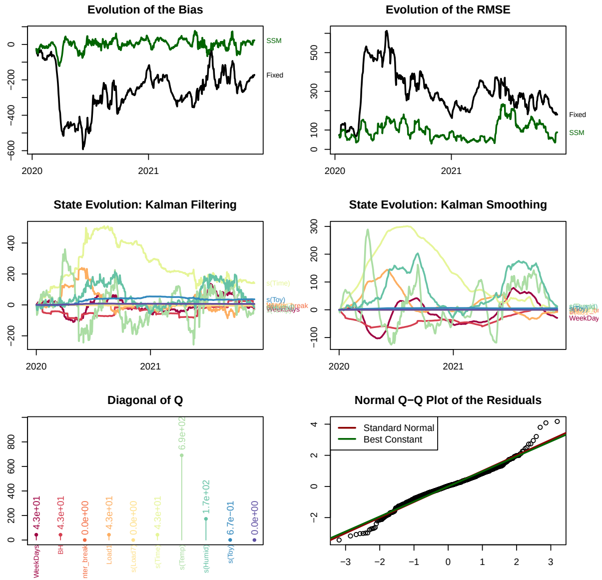
$$\psi_t(b_t) = \log\det(P_{t-1|t-1} + f(b_t))$$
$$+ Tr\Big((P_{t|t} + (\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})(\hat{\theta}_{t|t} - \hat{\theta}_{t-1|t-1})^\top)(P_{t-1|t-1} + f(b_t))^{-1}\Big)\,.$$

Evidence Lower Bound for $\hat{a}_{t|t}, s_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$.

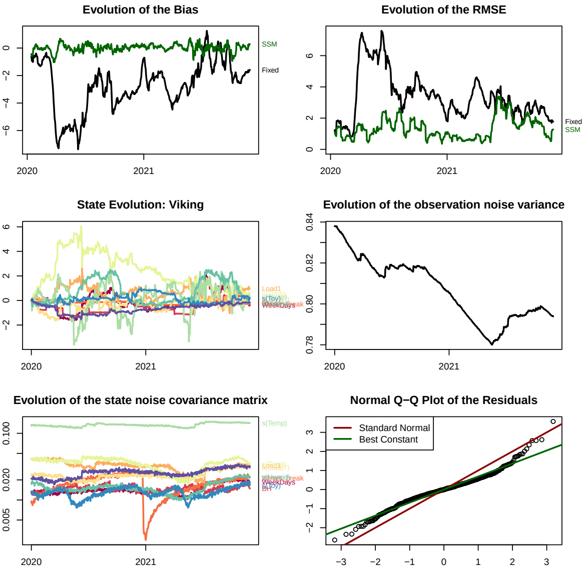
Package Viking: Static



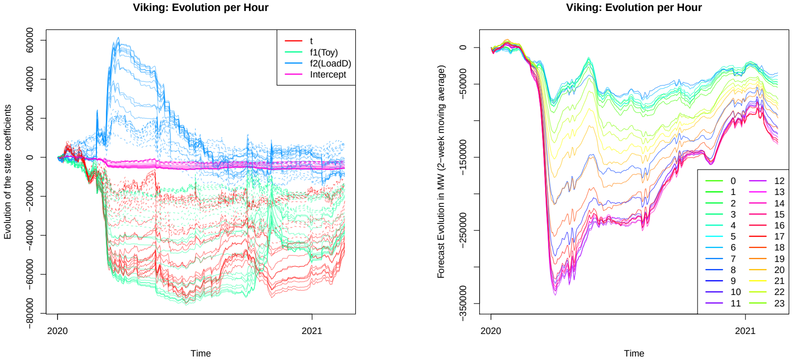
Package Viking: Dynamic



Package Viking: Viking Estimation



Different Evolution of the 24 Models



Data set: city-wide competition.