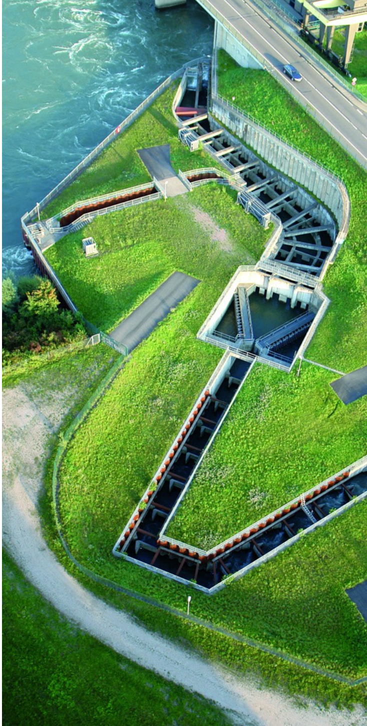
Forecasting the electricity consumption by aggregating specialized experts

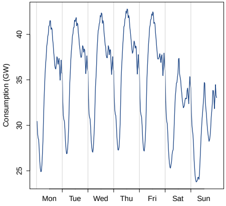
Pierre Gaillard (EDF R&D, ENS Paris)

with Yannig Goude (EDF R&D)
Gilles Stoltz (CNRS, ENS Paris, HEC Paris)

June 2013 – WIPFOR



Goal



Goal

Short-term (one-day-ahead) forecasting of the French electricity consumption

- Many models developed by EDF R&D: parametric, semi-parametric, and non-parametric
- Evolution of the electrical scene in France $\Rightarrow$ existing models get questionable

} Adaptive methods of models aggregation

EDF ENS

Setting – Sequential prediction with expert advice

Each instance t

- Each expert suggests a prediction $x_{i,t}$ of the consumption y_t
- We assign weight to each expert and we predict

$$\widehat{y}_t = \widehat{\boldsymbol{p}}_t \cdot \boldsymbol{x}_t \quad \left(= \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t} \right)$$

Our goal is to minimize our cumulative loss

$$\underbrace{\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2}_{\text{Our loss}} = \underbrace{\min_{i=1,\dots,N} \sum_{t=1}^{T} (x_{i,t} - y_t)^2}_{\text{Loss of the best expert}} + \underbrace{R_T}_{\text{Estimation error}}$$

Good set of experts

Good aggregating algorithm

Setting – Sequential prediction with expert advice

Each instance t

- Each expert suggests a prediction $x_{i,t}$ of the consumption y_t
- We assign weight to each expert and we predict

$$\widehat{y}_t = \widehat{\boldsymbol{p}}_t \cdot \boldsymbol{x}_t \quad \left(= \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t} \right)$$

Our goal is to minimize our cumulative loss

$$\underbrace{\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2}_{\text{Our loss}} = \underbrace{\min_{\boldsymbol{q} \in \Delta_N} \sum_{t=1}^{T} (\boldsymbol{q} \cdot \boldsymbol{x}_t - y_t)^2}_{\text{Loss of the best convex combination}} + \underbrace{R_T}_{\text{Estimation error}}$$

Good set of experts
As varied as possible

Good aggregating algorithm

Minimizing both approximation and estimation error

$$\underbrace{\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2}_{\text{Our loss}} = \underbrace{\min_{\boldsymbol{q} \in \Delta_N} \sum_{t=1}^{T} (\boldsymbol{q} \cdot \boldsymbol{x}_t - y_t)^2}_{\text{Approximation error}} + \underbrace{R_T}_{\text{Estimation error}}$$

Approximation error

⇒ good heterogeneous set of experts

Ex: specializing the experts, bagging, boosting, . . .

Estimation error

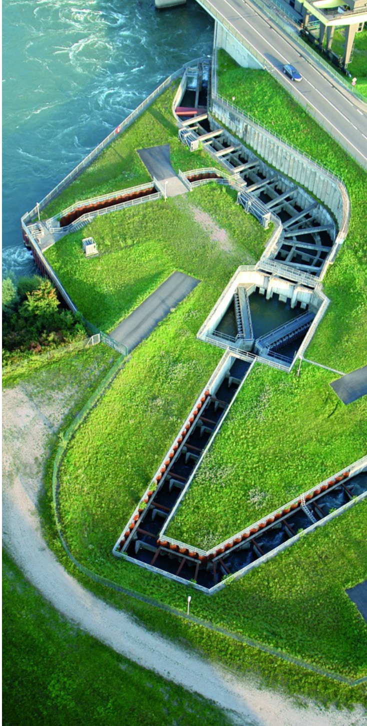
⇒ efficient algorithm for aggregating specialized experts

Ex: Exponentially weighted average, Exponentiated Gradient, Ridge, . . .

Prediction Learning and Games, Cesa-Bianchi and Lugosi, 2006

I. Aggregating algorithms

Prediction Learning and Games, Cesa-Bianchi and Lugosi, 2006



Exponentially weighted average forecaster (EWA)

Each instance t

- Each expert suggests a prediction $x_{i,t}$ of the consumption y_t
- We assign to expert i the weight

$$\widehat{p}_{i,t} = \frac{exp\left(-\eta \sum_{s=1}^{t-1}(x_{i,s} - y_s)^2\right)}{\sum_{j=1}^{N} exp\left(-\eta \sum_{s=1}^{t-1}(x_{j,s} - y_s)^2\right)}$$

- and we predict $\widehat{y}_t = \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t}$

Our cumulated loss is upper bounded by

$$\underbrace{\sum_{t=1}^{T}(\widehat{y}_t - y_t)^2}_{\text{Our loss}} \leqslant \underbrace{\min_{i=1,\dots,d} \sum_{t=1}^{T}(x_{i,t} - y_t)^2}_{\text{Loss of the best expert}} + \underbrace{\square\sqrt{T \log N}}_{\text{Estimation error}}$$

EDF ENS

Exponentially weighted average forecaster (EWA)

Each instance t

- Each expert suggests a prediction $x_{i,t}$ of the consumption y_t
- We assign to expert i the weight

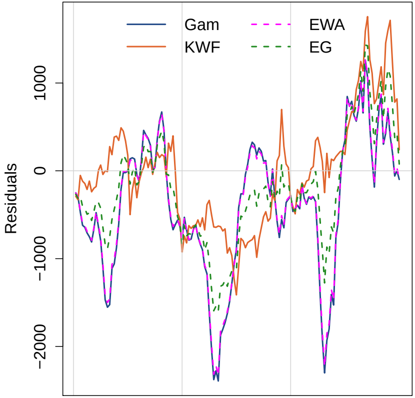
$$\widehat{p}_{i,t} = \frac{exp\left(-\eta \sum_{s=1}^{t-1} (x_{i,s} - y_s)^2\right)}{\sum_{j=1}^{N} exp\left(-\eta \sum_{s=1}^{t-1} (x_{j,s} - y_s)^2\right)}$$

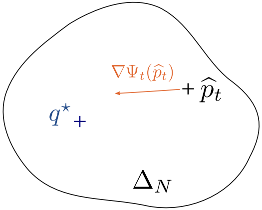
- and we predict $\widehat{y}_t = \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t}$

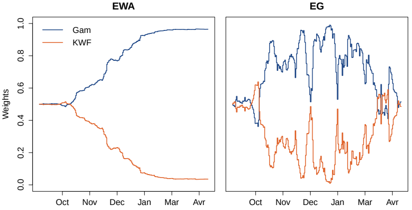
Our cumulated loss is upper bounded by

$$\underbrace{\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2}_{\text{Our loss}} \leqslant \underbrace{\min_{\boldsymbol{q} \in \Delta_N} \sum_{t=1}^{T} (q \cdot x_t - y_t)^2}_{\text{Loss of the best convex combination}} + \underbrace{?}_{\text{Estimation error}}$$

Motivation of convex combinations







Exponentiated gradient forecaster (EG)

Each instance t

- Each expert suggests a prediction $x_{i,t}$ of the consumption y_t
- We assign to expert i the weight

$$\widehat{p}_{i,t} \propto exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i,s}\right) \qquad \text{where } \ell_{i,s} = 2(\widehat{y}_s - y_s)x_{i,s}$$

- and we predict $\widehat{y}_t = \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t}$

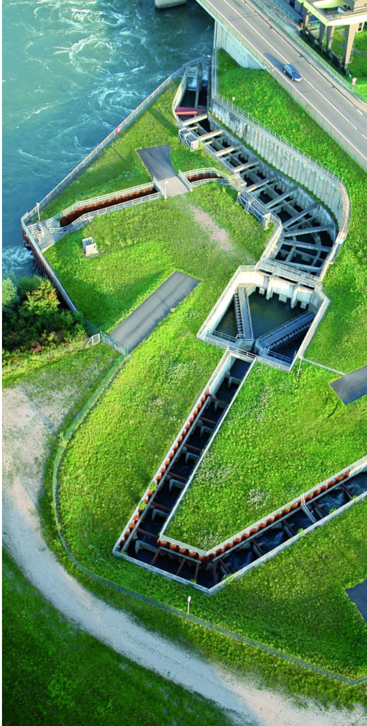
Our cumulated loss is then bounded as follow

$$\underbrace{\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2}_{\text{Our loss}} \leqslant \underbrace{\min_{q \in \Delta_N} \sum_{t=1}^{T} (\boldsymbol{q} \cdot \boldsymbol{x}_t - y_t)^2}_{\text{Loss of the best convexe combination}} + \underbrace{\square \sqrt{T \log N}}_{\text{Estimation error}}$$

Idea of proof

$$\begin{aligned} \sum_{t=1}^{T} (\widehat{y}_t - y_t)^2 - (\boldsymbol{q}^\star \cdot \boldsymbol{x}_t - y_t)^2 &\leqslant \sum_{t=1}^{T} \underbrace{2(\widehat{\boldsymbol{p}}_t \cdot \boldsymbol{x}_t - y_t)\boldsymbol{x}_t}\cdot(\widehat{\boldsymbol{p}}_t - \boldsymbol{q}^\star) \\ &= \sum_{t=1}^{T} \boldsymbol{\ell}_t \cdot (\widehat{\boldsymbol{p}}_t - \boldsymbol{q}^\star) \\ &\leqslant \sum_{t=1}^{T} \widehat{\boldsymbol{p}}_t \cdot \boldsymbol{\ell}_t - \min_i \sum_{t=1}^{T} \ell_{i,t} \end{aligned}$$

II. A good set of experts



Consider as heterogeneous experts as possible

Some ideas to get more variety inside the set of experts

- Consider heterogeneous prediction methods
  - Gam: semi-parametric method
    Generalized Additive Models, Wood, 2006
  - KWF: functional method based on similarity between days
    Clustering functional data using Wavelets, Antoniadis and al, 2013
- Create new experts from the same method thanks to boosting, bagging
- Vary the considered covariate: weather, calendar, . . .
- Specializing the experts: focus on specific situation (cloudy days,. . .) during the training

The dataset

The dataset includes 1 696 days from January 1, 2008 to June 15, 2012

- The electricity consumption of EDF customers
- Side information
  - weather: temperature, nebulosity, wind
  - temporal: date, EJP
  - loss of clients

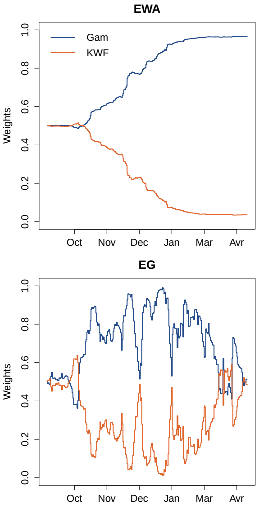
We remove uncommon days (public holidays ± 2) i.e., 55 days each year.

We split the dataset in two subsets

- Jan. 2008 – Aug. 2011: training set to build the experts
- Sept. 2011 – Jun. 2012: testing set

Performance of the forecasting methods and of the aggregating algorithms

Method	RMSE (MW)
Gam	847
KWF	1287
EWA	813
EG	778



Specializing the experts to diversify

Idea

Focus on specific scenarios during the training of the methods

Meteorological scenarios

- High / low temperature
- High / low variation of the temperature (since the last day, during the day)

Other scenarios

- High / low consumption
- Winter / summer

Such specialized experts suggest prediction only the days corresponding to their scenario

Specializing a method in cold days

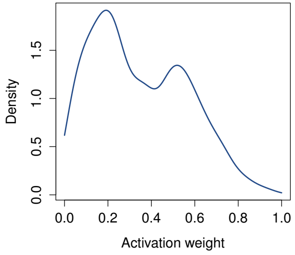
At day t, we consider

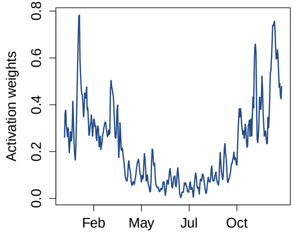
$$T_t = \text{average temperature of the day}$$

We normalize T_t on $[0, 1]$ and we choose for each day the weight

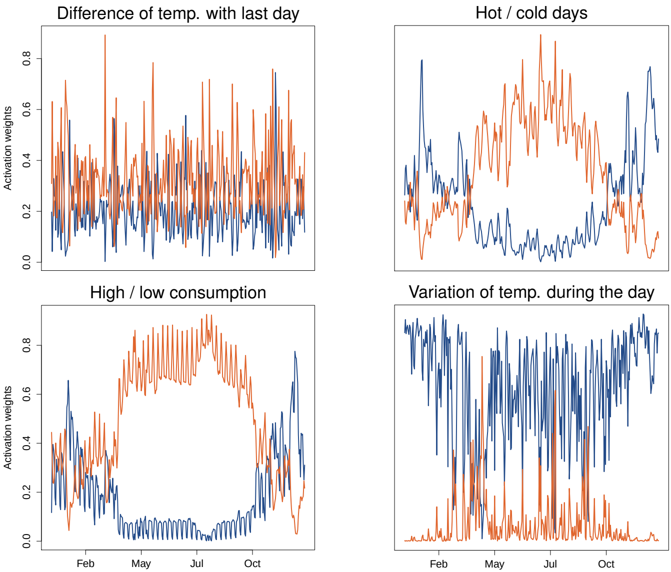
$$w_t = (1 - T_t)^2$$

We then train our forecasting method using the prior weights w_t on the training days





Weights given in 2008 for several specializing scenarios



Aggregating experts that specialize

Setting

Each day some of the experts are active and output predictions (according to their specialization) while other experts do not

When the expert i is non active, we do not have access to its prediction

A solution is to assume that non active experts output the same prediction $\widehat{y}_t$ as we do and solve the fixed-point equation

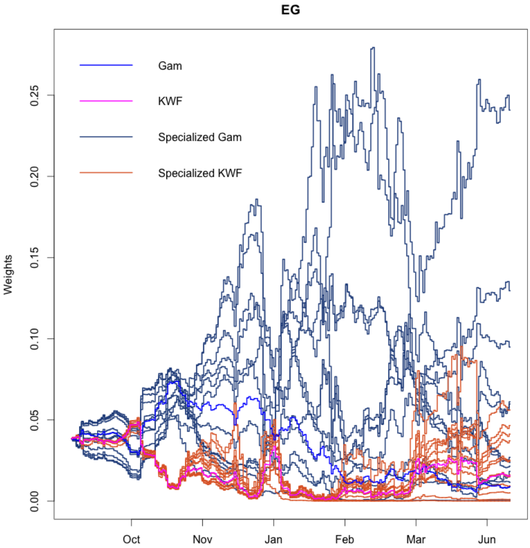
$$\widehat{y}_t = \sum_{j \text{ active}} \widehat{p}_{j,t}\, x_{j,t} + \sum_{i \text{ non active}} \widehat{p}_{i,t}\, \widehat{y}_t$$

Can be extended to activation functions of the experts $\in [0, 1]$

Forecasting the electricity consumption by aggregating specialized experts, Devaine and al., 2013

Performance of algorithms with specialized experts

Méthode	RMSE (MW)
Gam	847
KWF	1287
EWA	813
EG	778
Spec + EWA	765
Spec + EG	714



Performance of algorithms with specialized experts