Statistical and machine learning methods to model and forecast Energy

Mathilde Mougeot

ENSIIE & LPMA

NUS-USPC
November 2017
Electricity Framework

<table>
<thead>
<tr>
<th>Production</th>
<th>Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear power plants</td>
<td>industrial plants</td>
</tr>
<tr>
<td>coal fired power plants</td>
<td>home and heating</td>
</tr>
<tr>
<td>wind farms</td>
<td>building heating</td>
</tr>
<tr>
<td>photovoltaic farms</td>
<td>...</td>
</tr>
</tbody>
</table>

→ Electricity can hardly be stored. There is a need to:
Balance between electrical production and consumption
Forecast consumption and production
Statistical and machine learning methods to model and forecast Energy

We have studied models for Energy in several directions:

- **Consumption**
  High dimensional regression models to Forecast the French National Consumption. with D. Picard, K. Tribouley, V. Lefieux (RTE), JRSSB, AADA

- **Production**
  Modeling the Wind Farm production using machine learning tools with A. Fischer, L. Montuelle, D. Picard, Wind Energy 2017

- **Savings**
  Over consumption monitoring and diagnostic for industrial equipments : aggregation with O. Cadet (Air Liquide)
Electricity Production

Motivation: balance between electrical production and consumption
"Is it possible to built forecast models in the electricity consumption field which would rely on very few parameters and would be easy to calibrate without the need of human expertise - and which at the same time, would show good performances?"
Today, RTE provides day-ahead load forecasting using

- A statistical forecast model METEHORE (decision-making tool)
- An expert human expertise ("The forecasters")

Remark: The metehore model provides accurate forecast. However, it depends on a very large number of parameters and is not enough adaptive, whereas the electricity demand is evolving in a changing context.
The context today:

- The evolution of electrical uses, the energy-demand management, the smartgrids context... induce changes in consumers’ behavior

- New electric heating systems (such as heat pumps) induces variations in the weather-sensitivity of consumption

- The development of decentralized production (wind, solar) has an indirect impact on the consumption signal as the electrical transport network nodes

→ Forecast models of consumption have to adapt to quick changes in the load curve and to better control large errors
Recent works and existing models

- Time series analysis : exponential weights (Taylor 2012)
- Regression tree (Y. Goude et al. 2013,..)
- Aggregation (P. Gaillard & Y. Goude 2014)
- Model-based clustering (E. Devijver 2014)
Intraday load curve : Functional data

\[ Y \in \mathbb{R}^{n=48} (Y_t \ 1 \leq t \leq 2800) \]
Modeling each intra day signal as a function

We investigate the problem in a supervised learning setting:

- We consider each time unit signal:
  \[ Z_i = (Y_i, U_i), \quad i = 1, \ldots, n = 48 \]

- For each signal, we want to identify \( f \), an unknown function such that:
  \[ Y_i = f(U_i) + \epsilon_i. \]

where:

- The generic consumption signal observed on the time unit:
  \[ Y_i, \quad i = 1, \ldots, n \]

- The design (here fixed equi distributed):
  \[ U_i = \frac{i}{n} \]
Using a dictionary

Consider a dictionary $\mathcal{D}$ of functions $\mathcal{D} = \{g_1, \ldots, g_p\}$ and Assume that $f$ can be well fitted by this dictionary

$$f = \sum_{\ell=1}^{p} \beta_\ell \ g_\ell + h$$

where $h$ is a 'small' function (in absolute value).

The model is

$$Y_i = \sum_{\ell=1}^{p} \beta_\ell g_\ell(U_i) + h(U_i) + \epsilon'_i, \ i = 1, \ldots, n$$

which coincides with the linear model:

$$Y = X \beta + \epsilon \quad \text{with} \quad X(n \times p) \quad \text{putting} \quad \epsilon_i = h(U_i) + \epsilon'_i \ \text{and} \ G_{i\ell} = g_\ell(U_i).$$
High dimensional framework

Solution: $\hat{\beta} = \text{Argmin} ||Y - X\beta||^2$

- More variables (functions) than observations $n << p$

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
\end{bmatrix} = 
\begin{bmatrix}
  x_{11} & \cdots & \cdots & x_{1p} \\
  x_{n1} & \cdots & \cdots & x_{np}
\end{bmatrix} \ast 
\begin{bmatrix}
  \beta_1 \\
  \beta_2 \\
  \vdots \\
  \beta_p
\end{bmatrix} + \epsilon
\]

"Fat matrix"

→ Infinity of $\hat{\beta}$ solutions.
→ Need more assumptions on $\beta$ to solve the problem
→ Ex: Lasso ($\ell_1$ penalization), Ridge ($\ell_2$)…
Statistical framework

Theoretical background : Learning Out of Leaders

- \( Y = X\beta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2), \beta \) unknown
- \( \hat{\beta} = \text{Argmin} ||Y - X\beta||^2 \), OLS

Sparse approximation using Thresholding : Learning Out of Leaders* :

- based on 2 Thresholding steps,
- weak complexity, sparse and non linear solution,
- Algorithm in 3 steps (\( X \) column normalized, \( \sum_j X_j^2 / n = 1 \)) :
- Consistency results

<table>
<thead>
<tr>
<th>step</th>
<th>compute</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. SELECTION (threshold ( \lambda_1 ))</td>
<td>Find ( b ) Leaders ( b &lt; n &lt;&lt; p ) on Leaders the coefficients ( X_b ) ( \tilde{\beta} = (X_b^T X_b)^{-1} X_b^T Y )</td>
<td>((n, b)) ((1, b)) ((1, \hat{S}))</td>
</tr>
<tr>
<td>2. REGRESSION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. THRESHOLD ( \lambda_2 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \text{Picard, K. Tribouley, JRSS B 2012,B Stat. Methodol. vol 74} \)

Mathilde Mougeot (ENSIIE&LPMA)  NUS-USPC 2017
LOL assumptions and thresholds

- **When:**
  1. **Sparsity:**
     
     \[ B_0(S, M) := \{ \beta \in \mathbb{R}^p, \sum_{j=1}^{p} I\{ |\beta_j| \neq 0 \} \leq S, \| \beta \|_{l_1(p)} \leq M \}. \]
  2. **Dimension:** \( p \leq \exp(\Box n) \),
  3. **Coherence:** \( \tau_n \leq \Box \sqrt{\frac{\log p}{n}} \) ("max of correlation between columns")

- **Choose:** the thresholds \( \lambda_1, \lambda_2 \)
  
  \[ \lambda_1 = \Box \sqrt{\frac{\log p}{n}}, \lambda_2 = \Box \sqrt{\frac{\log p}{n}} \]

- **Approximation, Concentration results:**
  - Prediction loss:
    
    \[ \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - E Y_i)^2 = d(\hat{\beta}^*, \beta)^2 \]

\[
\sup_{\beta \in B_0(S, M)} \mathbb{P}\left( d(\hat{\beta}^*, \beta) > \eta \right) \leq \begin{cases} 
4e^{-\gamma \eta^2} & \text{for } \eta^2 \geq DS\left[\sqrt{\frac{\log p}{n}} \vee \tau_n\right]^2 \\
1 & \text{for } \eta^2 \leq DS\left[\sqrt{\frac{\log p}{n}} \vee \tau_n\right]^2
\end{cases}
\]


Mathilde Mougeot  (ENSIIE&LPMA)  
NUS-USPC 2017
Approximation

Generic Dictionary

• Each day $t$, $Y_t = X \beta_t + \epsilon_t$

• with Dictionary of $p$ functions $\mathcal{D} = \{g_1, \ldots g_p\}$ $G_{i\ell} = g_\ell(U_i)$

• For daily load curves ($dim(Y_t = 48)$): a good choice happened finally to be a mixture of the Fourier basis and the Haar basis, $p = 62$.

  1. (1 :1) constant function (1)
  2. (2 :24) cosine functions (with increasing frequencies) (23)
  3. (25 :47) sine functions (with increasing frequencies) (23)
  4. (48 :62) Haar functions (with increasing frequencies) (15)

• Approximation : $p = 7$, $E_{MAPE} = 1.4\%$
June 17th, 2009

$S = 5$, $MAPE = 0.0147$

Figure – 2003 04 30

left: observed signal - red line, approximated signal - blue line
right: $S$ coefficients on the dictionary
November 18\textsuperscript{th} 2007

\[ S = 12, \text{ MAPE} = 0.0057 = 0.57\%. \quad \text{MAPE} = \frac{1}{n} \sum_{i=1}^{n=48} |Y_i - \hat{Y}_i| / Y_i \]

\textbf{Figure – 2007 11 18}

left: observed signal - red line, approximated signal - blue line
right: \( S \) coefficients on the dictionary
Spot of Temperatures, Cloud Cover and Wind information

Figure – Temp., Cloud Cover spots (#39) and wind data (#293)
### Intraday Specific Dictionary

- Each day $t$, \[ Y_t = X_t \beta_t + \epsilon_t \]

- with Dictionary of $p$ functions $\mathcal{D}_t = \{g^t_1, \ldots g^t_p\}$

  Final model, $p = 10$ ($p = 14$)

  1. Shape functions (group centroid, previous week day $Y_{t-7}$)
  2. Climate functions (Temperature and Cloud Cover Indicators computed over the 39 meteorological spots. (and Wind...(+4))

- Approximation performance:
  - LOL adaptive using shape and meteorological variables
    - $S = 2.35 [2;6]$,
    - $\bar{E}_{MAPE} = 1.5\%$
  - LOL adaptive using a generic dictionary
    - Trigonometric-Fourier
    - $S = 7$
    - $\bar{E}_{MAPE} = 1.7\%$
Illustration: intraday load curve model/prediction (Winter Monday, Thursday)

\[ Y_t = X_t\beta_t + \epsilon_t \]
Each day \( t \):

- \( Y_t = \hat{Y}_t + \hat{\epsilon}_t \) with model: \( \hat{Y}_t = \sum_{j=1}^{p} \hat{\beta}_j g_j^t \)

- Forecast \((d > t)\) \( \tilde{Y}_d = \sum_{j=1}^{p} \tilde{\beta}_j^d g_j^d + \delta_d \)

Looking for a good candidate of coefficients in the past:

- Plug in estimated coefficients
- \( \tilde{\beta}_d = \hat{\beta}_{\mathcal{M}(d)} \) with \( \mathcal{M}(d) << d \)
- \( \mathcal{M} "Expert" \)
Expert $\mathcal{M}$ to forecast

**Strategy** Let $\mathcal{M}$ be a function (strategy), from $\mathbb{N}$ to $\mathbb{N}$ such that for any $t \in \mathbb{N}$, $\mathcal{M}(t) < t$. (data dependent or not)

**Plug-in** To the strategy $\mathcal{M}$ we associate the expert $\tilde{Y}_d^{\mathcal{M}}$: the forecast of the signal of day $d$ using prediction strategy $\mathcal{M}$.

$$\tilde{Y}_d^{\mathcal{M}} = \sum_{j=1}^{p} \hat{\beta}_j^{\mathcal{M}(d)} g_j^{d} + \delta_d$$

$\hat{\beta}_j^{\mathcal{M}(d)}$, $j = 1, \ldots, p$ are the estimated coefficients computed with LOL algorithm at day $\mathcal{M}(d)$. 
Specialized Experts focus on

Nearest neighbor strategies based on different variables and metrics:

1. (2) Time depending (t-1, t-7)
2. (2) climatic configuration of the day (Temperature)
3. (2) constrained climatic configuration of the day (Temperature/Cloud Covering)
4. group constraint climatic configuration of the day (Temperature/group)
5. climatic configuration of the day constrained by the type of the day (Temperature/day)
6. climatic configuration of the day constrained by a calendar group (Temperature/calendar)
7. climatic configuration of the day (Cloud cover)
8. group constraint climatic configuration of the day (Cloud Covering/group)
9. climatic configuration of the day constrained by the type of the day (Cloud Covering/day)
10. climatic configuration of the day constrained by a calendar group (Cloud Covering/calendar)
11. Wind ...
MAPE Forecast performances

Forecast results are computed using one year of data from 1\textsuperscript{st} September 2009 to 31\textsuperscript{th} August 2010.

<table>
<thead>
<tr>
<th>M</th>
<th>mean</th>
<th>med</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.0634</td>
<td>0.0415</td>
<td>0.0046</td>
<td>0.1982</td>
</tr>
<tr>
<td>Apx</td>
<td>0.0183</td>
<td>0.0151</td>
<td>0.0035</td>
<td>0.0862</td>
</tr>
<tr>
<td>tm1</td>
<td>0.0323</td>
<td>0.0262</td>
<td>0.0050</td>
<td>0.1412</td>
</tr>
<tr>
<td>tm7</td>
<td>0.0303</td>
<td>0.0239</td>
<td>0.0056</td>
<td>0.1920</td>
</tr>
<tr>
<td>T</td>
<td>0.0305</td>
<td>0.0242</td>
<td>0.0065</td>
<td>0.2232</td>
</tr>
<tr>
<td>Tm</td>
<td>0.0321</td>
<td>0.0264</td>
<td>0.0062</td>
<td>0.2138</td>
</tr>
<tr>
<td>T/N</td>
<td>0.0328</td>
<td>0.0258</td>
<td>0.0043</td>
<td>0.4762</td>
</tr>
<tr>
<td>Tm/N</td>
<td>0.0321</td>
<td>0.0248</td>
<td>0.0057</td>
<td>0.1639</td>
</tr>
<tr>
<td>T/G</td>
<td>0.0337</td>
<td>0.0247</td>
<td>0.0058</td>
<td>0.4762</td>
</tr>
<tr>
<td>T/d</td>
<td>0.0330</td>
<td>0.0257</td>
<td>0.0052</td>
<td>0.3749</td>
</tr>
<tr>
<td>T/c</td>
<td>0.0314</td>
<td>0.0249</td>
<td>0.0054</td>
<td>0.1848</td>
</tr>
<tr>
<td>Cs/G</td>
<td>0.0297</td>
<td>0.0230</td>
<td>0.0047</td>
<td>0.1915</td>
</tr>
<tr>
<td>C/d</td>
<td>0.0281</td>
<td>0.0219</td>
<td>0.0036</td>
<td>0.2722</td>
</tr>
<tr>
<td>C/c</td>
<td>0.0288</td>
<td>0.0224</td>
<td>0.0036</td>
<td>0.2722</td>
</tr>
</tbody>
</table>
Aggregation of predictors : Exponential weights

\[ \hat{Y}_d^{wgt*} = \frac{\sum_{m=1}^{M} w_d^m \tilde{Y}_d^m}{\sum_{m=1}^{M} w_d^m} \]

with

\[ w_d^M = \exp(-|\hat{Y}_{d*}^M - Y_{d*}^M|^2/\theta) \]

\( \theta \) is a parameter, calibrated by cross-validation.
Performances after aggregation

Mape performances for aggregated methods computed for one year

<table>
<thead>
<tr>
<th>mean</th>
<th>med</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0230</td>
<td>0.0197</td>
<td>0.0052</td>
<td>0.0695</td>
</tr>
</tbody>
</table>

Mape performances for Oracle computed for one year

<table>
<thead>
<tr>
<th>mean</th>
<th>med</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0144</td>
<td>-</td>
<td>-</td>
<td>0.074</td>
</tr>
</tbody>
</table>

Mathilde Mougeot (ENSIIE&LPMA) NUS-USPC 2017
Conclusion

- Competitive approach compared to usual time serie analysis with much less parameters.

- Sparse approximation
  - a Generic dictionary for compression and pattern extraction
  - Intra day specific dictionaries for approximation and prediction

- Forecasting
  - Various experts for prediction based on a retrieval information strategy
  - Aggregation using exponential weights,

- FOREWER project research
  - prediction for renewable energy with machine learning methods

- work in progress for improvement