Modeling and forecasting intraday load curves using high dimensional methods

Mathilde Mougeot

Joint work with
D. Picard (UPD) & V. Lefieux, L. Teyssier-Maillard (RTE)

SESO, June 26^{th} 2015

Electrical Consumption Time series

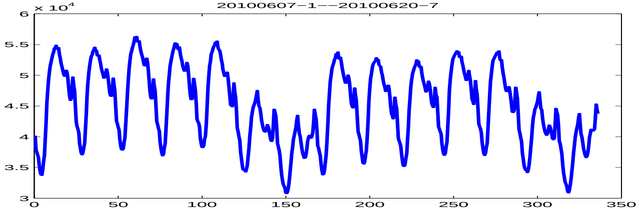


Figure: Two weeks of The French National electrical consumption

Electrical Consumption Time series

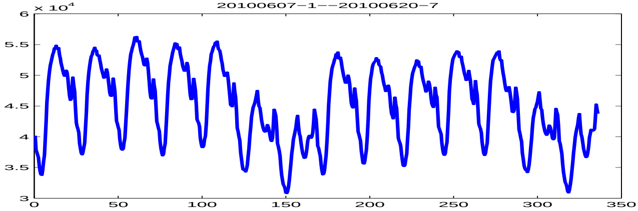
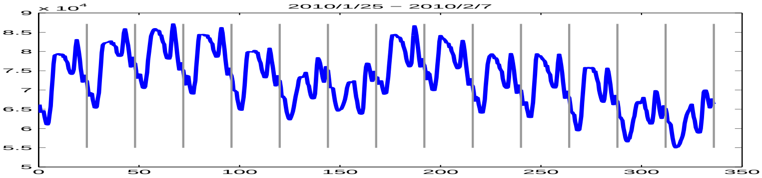


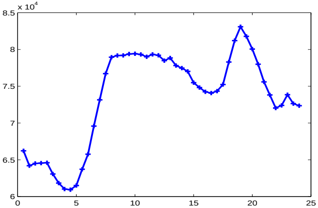
Figure: Two weeks of The French National electrical consumption

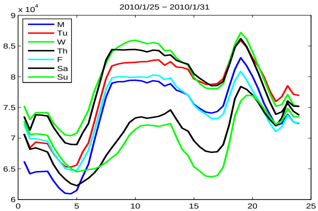
RTE requirement:

"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?"

Intraday load curve:Fonctional data







Intra day load curve, 30' sampling (48 pts),

$Y \in \mathbb{R}^{n=48}$ (Y_t $1 \leq t \leq 2800$)

Outline.

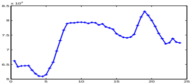
From Sparse Approximation towards Forecast:

- I. High dimensional Regression
  - Theoretical framework
- II. Application to the intra day load curves: sparse Approximation.
  - Generic Dictionary for knowledge discovery
  - Specific dictionary composed of Climate functional variables
- III. Towards Forecast
  - Strategies using Expert
  - Aggregation of Experts

→ *Scientific collaboration with RTE "Réseau Transport électrique" who wants to revised its Forecasting model based on time serie*

Modeling each intra day signal as a function

We investigate the problem in a supervised learning setting:



- We consider each time unit signal:

$$Z_i = (Y_i, U_i),\ i = 1, \ldots, n = 48$$

- 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:

$$Y_i,\ i = 1, \ldots, n$$

  - 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 :

$$Y = X\beta + \epsilon \qquad \textit{with} \;\; X(n \times p)$$

putting $\epsilon_i = h(U_i) + \epsilon_i^{'}$ and $G_{i\ell} = g_\ell(U_i)$.

High dimensional framework

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

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

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & \dots & x_{1p} \\ \\ x_{n1} & & \dots & x_{np} \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \dots \\ \\ \beta_p \end{bmatrix} + \epsilon$$

"Fat matrix"

→ Infinity of $\hat{\beta}$ solutions.

→ Need more assumptions on β to solve the problem

→ Ex: Lasso (ℓ_1 penalization), Ridge (ℓ_2)...

Theoretical background: Learning Out of Leaders

- $Y = X\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, β unknown
- $\hat{\beta} = Argmin||Y - X\beta||^2$, 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, $\sum_j X_j^2/n = 1$):
- Concistency results

step		compute	size
1. SELECTION (threshold)	Find b Leaders $b < n << p$	X_b	(n, b)
2. REGRESSION	on Leaders	$\tilde{\beta} = (X_b^T X_b)^{-1} X_b^T Y$	$(1, b)$
3. THRESHOLD	the coefficients	$\hat{\beta}$	$(1, \hat{S})$

(*) *MM, D. Picard, K. Tribouley, JRSS B 2012,B Stat. Methodol. vol 74*

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:** $p \leq \exp(\square n)$,
  3. **Coherence:** $\tau_n \leq \square \sqrt{\frac{\log p}{n}}$

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:** $p \leq \exp(\square n)$,
  3. **Coherence:** $\tau_n \leq \square\sqrt{\frac{\log p}{n}}$
- Choose: **the thresholds λ_1, λ_2**
  $\lambda_1 = \square\sqrt{\frac{\log p}{n}}, \ \lambda_2 = \square\sqrt{\frac{\log p}{n}}$

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:** $p \leq \exp(\square n)$,
  3. **Coherence:** $\tau_n \leq \square\sqrt{\frac{\log p}{n}}$
- Choose: **the thresholds λ_1, λ_2**
  $\lambda_1 = \square\sqrt{\frac{\log p}{n}},\ \lambda_2 = \square\sqrt{\frac{\log p}{n}}$
- Approximation, Concentration results:
  - Prediction loss: $\frac{1}{n}\sum_{i=1}^n(\widehat{Y}_i - \mathbb{E}Y_i)^2 = d(\hat{\beta}^*,\beta)^2$

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

(*) *MM, D. Picard, K. Tribouley, JRSS B 2012,B Stat. Methodol. vol 74*

Outline.

From Sparse Approximation towards Forecast:

- I. High dimensional Regression
  - Theoretical framework
- II. Sparse Approximation. Application to the intra day load curves
  - Generic Dictionary, knowledge discovery
  - Specific dictionary composed of Climate functional variables
- III. Towards Forecast
  - Strategies using Expert
  - Aggregation of Experts

Generic Dictionary

- Each day t, $\boxed{Y_t = X\beta_t + \epsilon_t}$
- with Dictionary of p functions $\mathcal{D} = \{g_1, \ldots g_p\}$ $G_{i\ell} = g_\ell(U_i)$
- For daily load curves:
  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: $p = 7$, $E_{MAPE} = 1.4\%$

November 18^{th} 2007

$S = 12,\ MAPE = 0.0057 = 0,57\%.$
$MAPE = \frac{1}{n}\sum_{i}^{n=48} |Y_i - \hat{Y}_i| / Y_i$

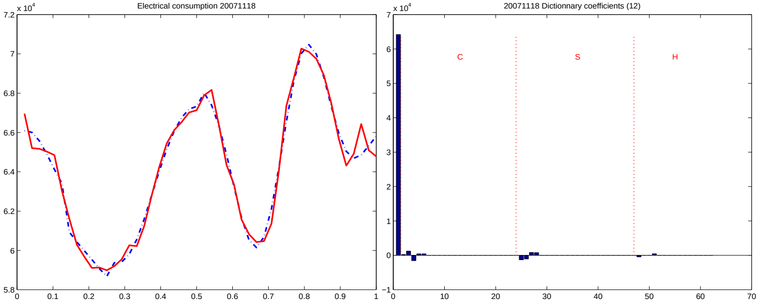


Figure: 2007 11 18

left: observed signal - red line, approximated signal -blue line
right: S coefficients on the dictionary

$S = 5,\ MAPE = 0.0147$

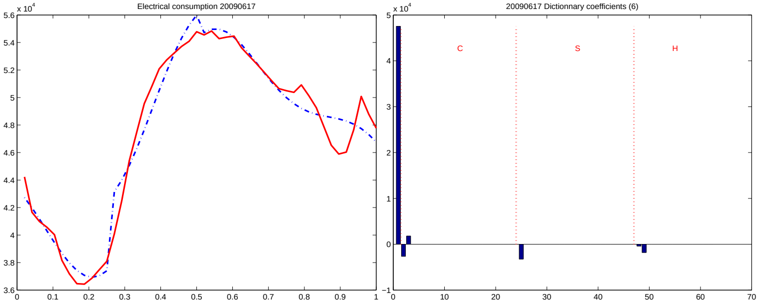


Figure: 2003 04 30

left: observed signal - red line, approximated signal -blue line

right: S coefficients on the dictionary

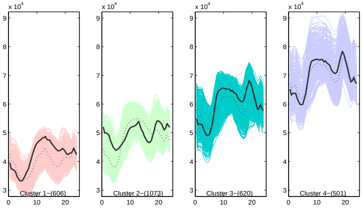
Segmentation of the intra-day load curves using Sparse Approximation on a Generic Dictionary

- Using the sparse approximation (same support, $S = 8$)
- using a clustering algorithm in 2 steps (k-means algorithm)
- Segmentation of the daily signals in clusters
- ...
- From Cluster to groups using calendar interpretation

8 years of data: $T = 2800$ intra day load curves ($n = 48$)

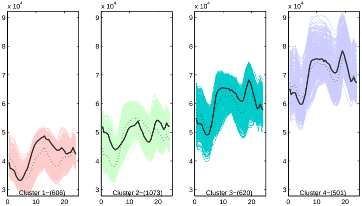
Mining the cluters... to Groups

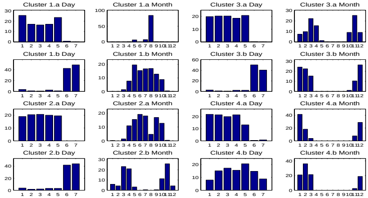
From clusters:



Mining the cluters... to Groups

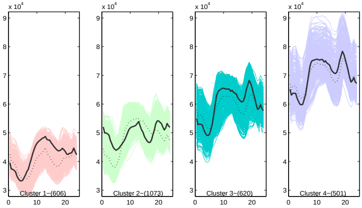
From clusters:

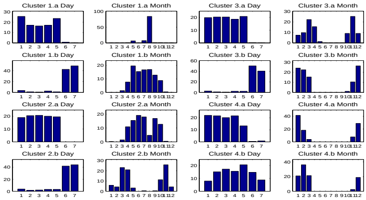




Mining the cluters... to Groups

From clusters:





To groups: calendar interpretation of the clusters

	Months											
Days	1	2	3	4	5	6	7	8	9	10	11	12
1	7	8	5	3	3	3	3	1	3	3	5	7
2	7	8	5	3	3	3	3	1	3	3	5	7
3	7	8	5	3	3	3	3	1	3	3	5	7
4	7	8	5	3	3	3	3	1	3	3	5	7
5	7	8	5	3	3	3	3	1	3	3	5	7
6	6	8	4	4	2	2	2	2	2	2	4	6
7	6	6	4	4	2	2	2	2	2	2	4	6

Outline.

From Sparse Approximation towards Forecast:

- I. High dimensional Regression
  - Theoretical framework
- II. Sparse Approximation. Application to the intra day load curves
  - Generic Dictionary, knowledge discovery
  - Specific dictionary composed of Climate functional variables
- III. Towards Forecast
  - Strategies using Expert
  - Aggregation of Experts

Spot of Temperatures, Cloud Cover and Wind information

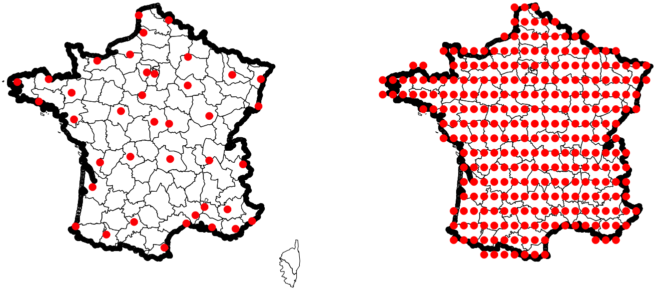


Figure: Temp., Cloud Cover spots (♯39) and wind data (♯293)

Intraday Specific Dictionary

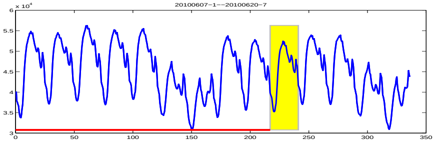
- Each day t, $\boxed{Y_t = X_t \beta_t + \epsilon_t}$
- with Dictionary of p functions $\mathcal{D}_t = \{g_1^t, \dots g_p^t\}$
  Final model, $p = 10$ ($p = 14$)
  1. **2**, Shape fonctions (group centroid, previous week day Y_{t-7})
  2. **8**, Climate fonctions (Temperature and Cloud Cover Indicators 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
    - $S = 7$
    - $\bar{E}_{MAPE} = 1.7\%$

Outline.

From Sparse Approximation towards Forecast:

- I. High dimensional Regression
  - Theoretical framework
- II. Sparse Approximation. Application to the intra day load curves
  - Generic Dictionary, knowledge discovery
  - Specific dictionary composed of Climate functional variables
- III. Towards Forecast
  - Strategies using Expert
  - Aggregation of Experts

From Sparse approximation to Forecast



Each day t:

- $Y_t = \hat{Y}_t + \hat{\epsilon}_t$
- Model: $\hat{Y}_t = \sum_{j=1}^{p} \hat{\beta}_j^t g_j^t$
- Forecast: $\tilde{Y}_t = \sum_{j=1}^{p} \tilde{\beta}_j^t g_j^t + \delta_{\mathbf{t}}$

Looking for a good candidate of coefficients in the past:

- Plug in estimated coefficients
- $\tilde{\beta}_t = \hat{\beta}_{\mathcal{M}(t)}$ with $\mathcal{M}(t) << t$
- $\mathcal{M}$ "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)

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 $\mathcal{M}$.

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

$\hat{\beta}^j_{\mathcal{M}(d)}$, $j = 1, \ldots, p$ 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^{st} September 2009 to $31th$ August 2010.

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

Aggregation of predictors: Exponential weights

Aggregated forecast:

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

with

$$w_d^{\mathcal{M}} = exp(-|\hat{Y}_{d^*_{\mathcal{M}}} - Y_{d^*_{\mathcal{M}}}|_2^2/\theta)$$

θ is a parameter, calibrated by cross-validation.

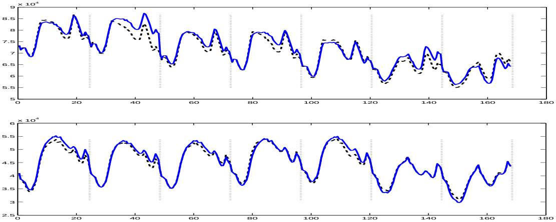
Performances after aggregation

Mape performances for aggregated methods computed for one year

mean	med	min	max
0.0230	0.0197	0.0052	0.0695

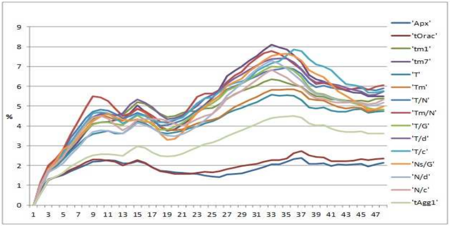
Mape performances for Oracle computed for one year

mean	med	min	max
0.0144	-	-	0.074



RTE feedbacks

- Evaluation of the "forecast software" in RTE.
- 30' relative ℓ_1 errors



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
  - Aggregation using exponential weights,
- Actually continued in the FOREWER ANR
  - prediction for renewable energy
- work in progress for improvement