From Sparse Approximation to Forecast of Intraday Load Curves

Mathilde Mougeot

Joint work with
D. Picard, K. Tribouley (P7)&
V. Lefieux, L. Teyssier-Maillard (RTE)

Electrical Consumption Time series

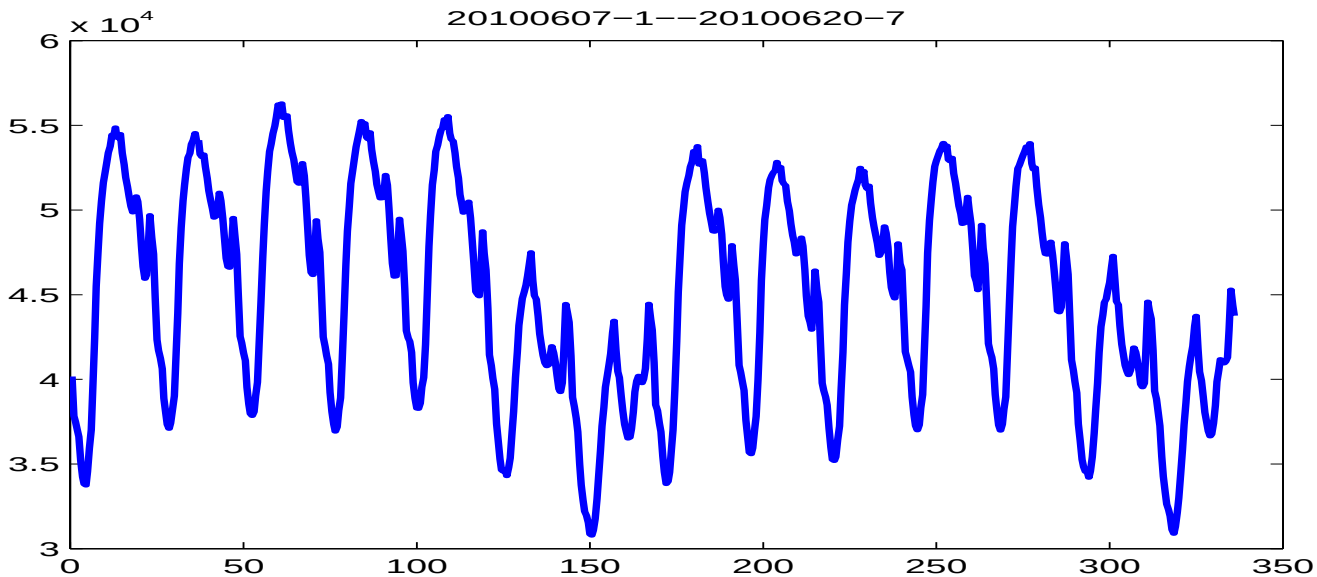


Figure : Two weeks of electrical consumption

Intraday load curves

From Sparse Approximation towards Forecast:

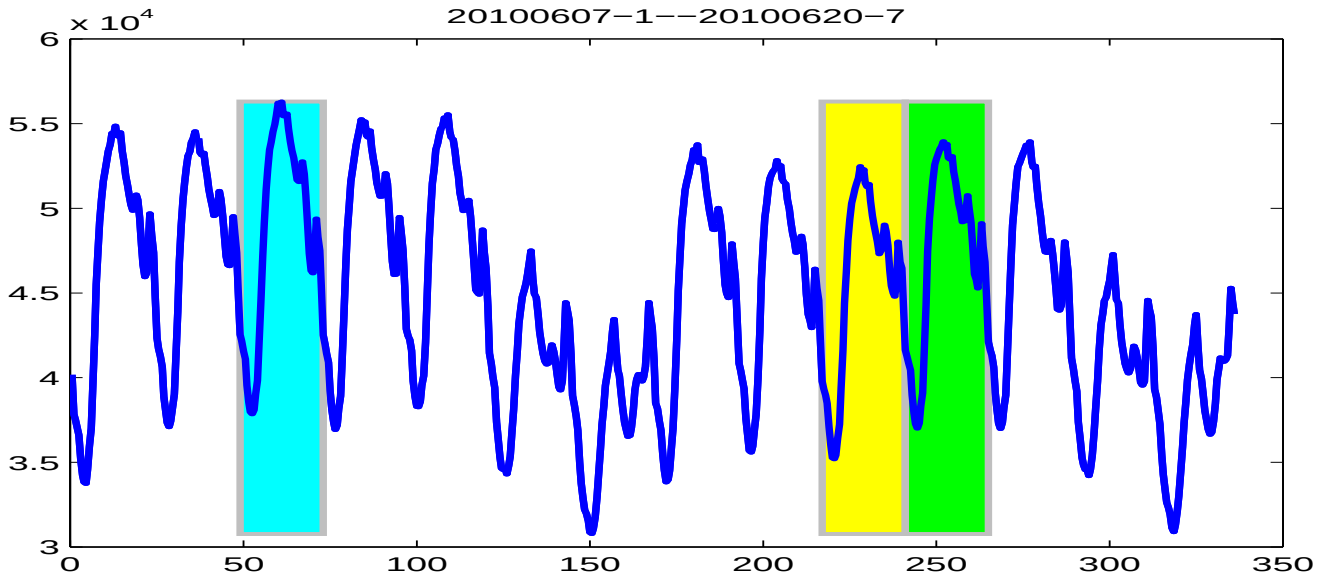
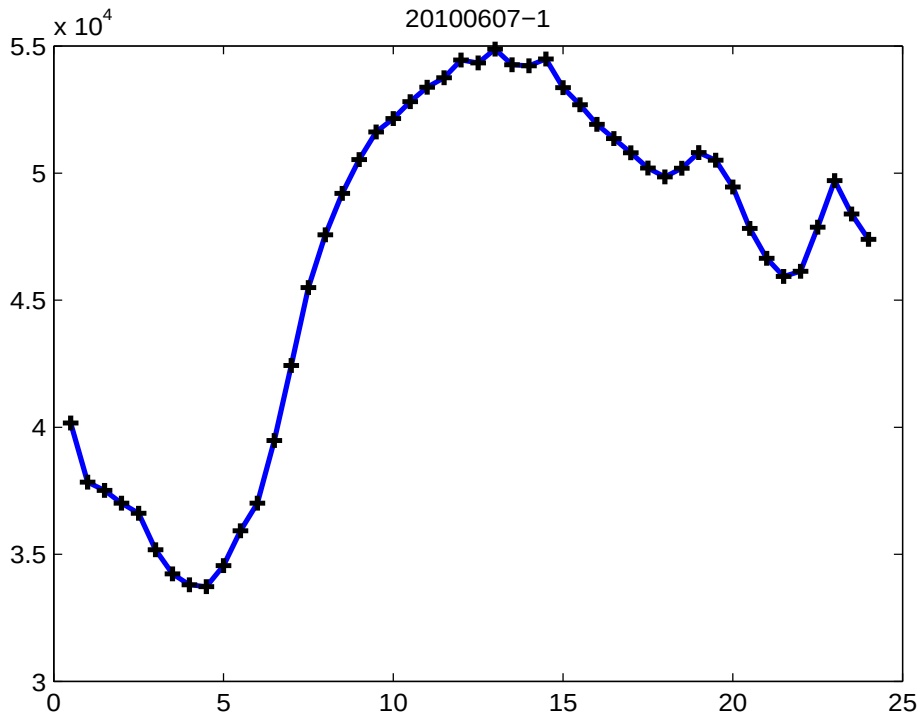


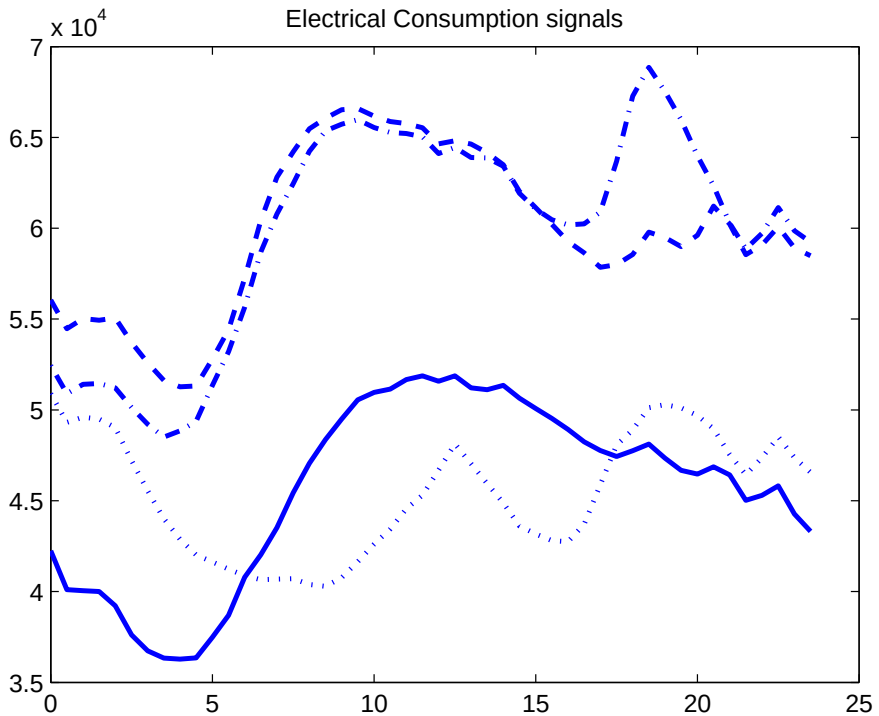
Figure : Functional data, Intra day load curves

Intraday load curve



Intra day load curve, 30' sampling (48 pts),
$Y \in \mathbb{R}^{n=48}$ (Y_t $1 \leq t \leq 2800$)

Intra-day load curves



Intraday load curves for some days.
2003-10-27: dashed dot line, 2003-08-28: solid line, 2003-01-01: dot line, 2003-04-10: dashed line.

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

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

Modeling each signal as a function

We investigate the problem in a supervised learning setting.

- We consider each time unit signal

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

- The generic consumption signal observed on the time unit:

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

- The design (here fixed equi distributed):

$$U_i = \frac{i}{n}$$

- We want to identify f (for each signal) in such a way that the model

$$Y_i = f(U_i) + \epsilon_i^{'}.$$

makes sense (has 'small' errors ϵ_i's).

Using a dictionary

Consider a dictionary $\mathcal{D}$ of functions $\mathcal{D} = \{g_1, \dots g_p\}$ 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

$$Y_i = \sum_{\ell=1}^{p} \beta_\ell g_\ell(U_i) + h(U_i) + \epsilon_i^{'}, \; i = 1, \dots, n$$

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 $G_{i\ell} = g_\ell(U_i)$.

Classical framework

- more observations than variables $n > p$
- and weak collinearity between co variables, $X^T X$ invertible

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ \\ y_n \end{bmatrix} = \underbrace{\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ & & \\ & & \\ & & \\ x_{n1} & \cdots & x_{np} \end{bmatrix}}_{\text{"Thin matrix"}} * \begin{bmatrix} \beta_1 \\ \beta_2 \\ \\ \\ \beta_p \end{bmatrix} + \epsilon$$

$\rightarrow$ Unique Solution: $\hat{\beta} = (X^T X)^{-1} X^T Y$

High dimensional framework

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

- More variables than observations $n << p$

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

"Fat matrix"

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

$\rightarrow$ Need more assumptions on β to solve the problem

$\rightarrow$ Ex: Lasso (ℓ_1 penalization), Ridge (ℓ_2)...

Alternative procedure

- Learning Out of Leaders*: based on 2 Thresholding steps,
- weak complexity, sparse solution,
- Algorithm in 3 steps (X column normalized):

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 - step1

1. SELECTION	Leaders among p	$X \to X_b$	(n, b)

based on a (Y, X_ℓ) "correlation" search and thresholding:

$$\mathcal{K}_\ell = |\frac{1}{n}\sum_{i=1}^{n} X_{i\ell} Y_i| \qquad \forall \ell,\ 1 \le \ell \le p$$

- Find the set $B = \{\ell,\ \mathcal{K}_l \ge \lambda_1\}$.
- Theoretical Threshold, $\lambda_1 = T_1\sqrt{\frac{\log p}{n}}$,
  T_1: constant $(\sigma, \nu, \mathcal{M}, c_0)$
- Data driven choice of λ_1 for practical applications (LOLA)

LOL - step 2

$Y = X\beta + \epsilon$,
Y $(n \times 1)$, X $(n \times p)$, S non zero coefficients β

SELECTION	Find b Leaders	X_b
2. REGRESSION	on Leaders	$\tilde{\beta} = (X_b^t X_b)^{-1} X_b^t Y$
THRESHOLD	the coefficients	$\hat{\beta}$

LOL, step 3: Threshold

- Threshold (again like step 1) the estimated coefficients to obtain the final predictor

$$\hat{\beta}^*_\ell = \widehat{\beta}_\ell \, I\{|\widehat{\beta}_\ell| \geq \lambda_2\}$$

- Threshold $\lambda_2 = T_2 \sqrt{\frac{\log p}{n}}$,
- For some constant $T_2 > 0$, $T_2(\sigma, \nu, \mathcal{M}, c_0)$
- To have:
  - Estimation: $\hat{\beta}^*_\ell$
  - Selection: $\hat{\beta}^*_\ell \neq 0$, **sparse solution**
  - Prediction:$X\hat{\beta}$
- Data driven choice of λ_2 for pratical applications (LOLA)

LOL assumptions, theoretical 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: $\frac{1}{n}\sum_{i=1}^n (\widehat{Y}_i - \mathbb{E}Y_i)^2 = d(\hat{\beta}^*, \beta)^2$

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

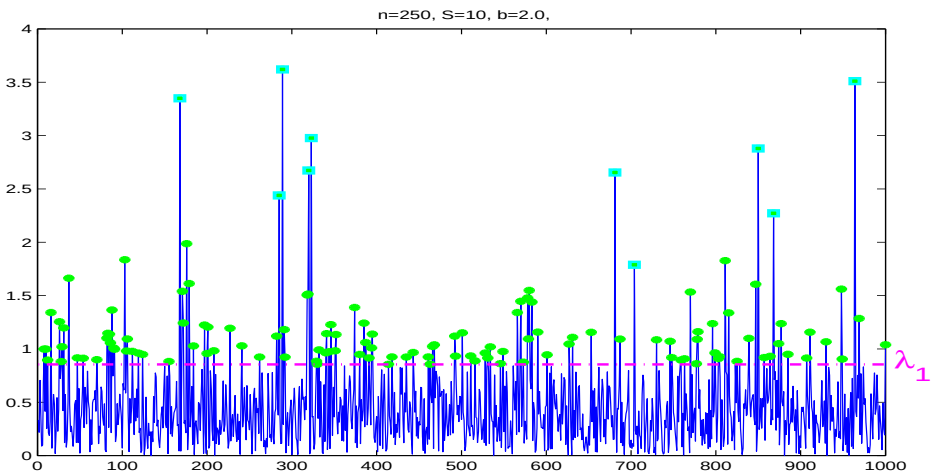
(*) *MM, D. Picard, K. Tribouley, JRSS B 2012,B Stat. Methodol. vol 74*

Illustration on simulated data, LOL step 1

X i.i.d. $\mathcal{N}(0,1)$, $\beta^* \sim \mathcal{N}(2,1)$, $S = 10$

The leaders are:

$B = \{\ell,\ \mathcal{K}_l \geq \lambda_1\}$, with $\lambda_1 = T_1\sqrt{\frac{\log p}{n}}$ Applications: adaptive λ_1

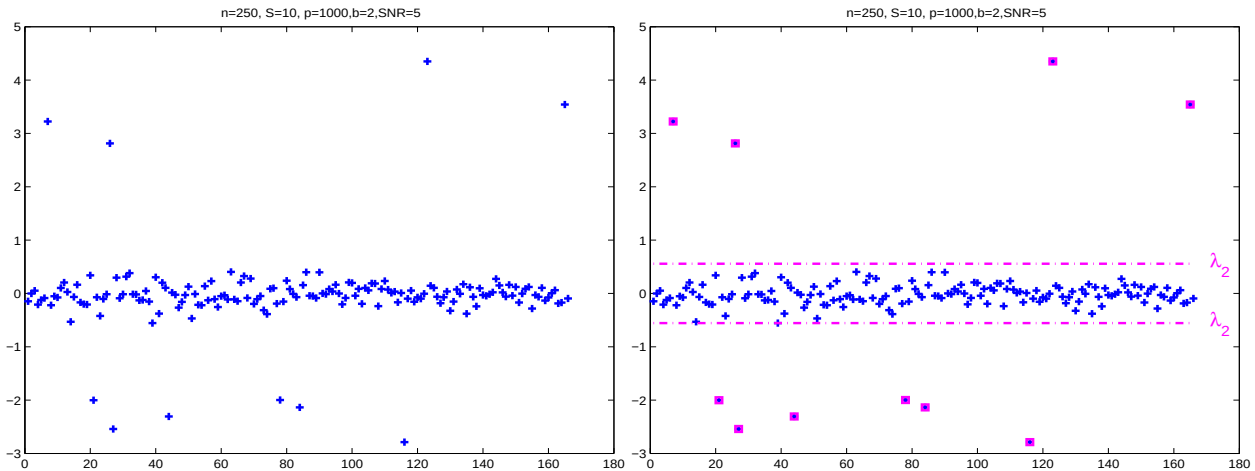


$n = 250$, $p = 1000$ $\quad \rightarrow \rho = \frac{S}{n} = 0.025$, $\delta = 1 - \frac{n}{p} = 0.75$

$card(B) = 170 >> S$

Illustration on simulated data, LOL step2,3

Ex1: X i.i.d. $\mathcal{N}(0,1)$, $\beta^* = \mathcal{N}(2,1)$,$S = 10$, (Leaders $b = 170$)

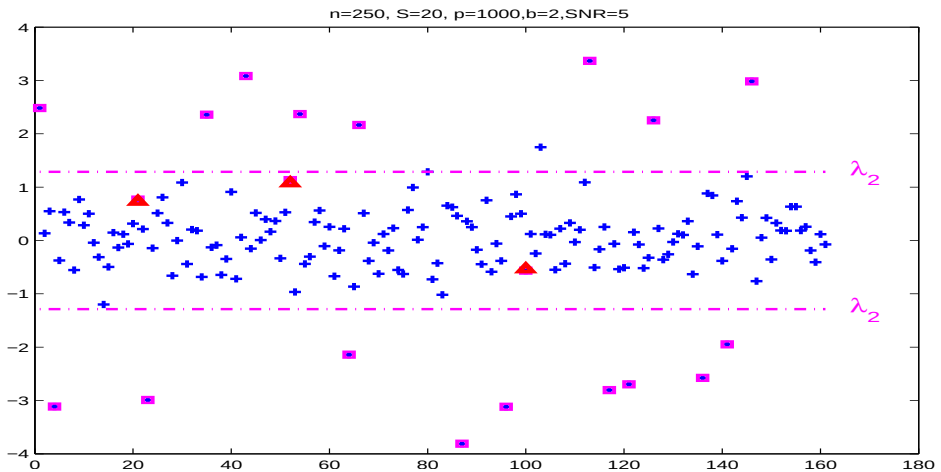


$n = 250$,$p = 1000$ $\rightarrow \rho = 0.04$, $\delta = 0.75$

	$p - \hat{S}$	$\hat{S}$
$p - S$	1000	0
$S = 20$	0	10

Illustration on simulated data, LOL step2,3

Another example Ex2: X i.i.d. $\mathcal{N}(0,1)$, , $S = 20$, $\rho = 0.08$, $\delta = 0.75$



	$p - \hat{S}$	$\hat{S}$
$p - S$	999	3
$S = 20$	1	17

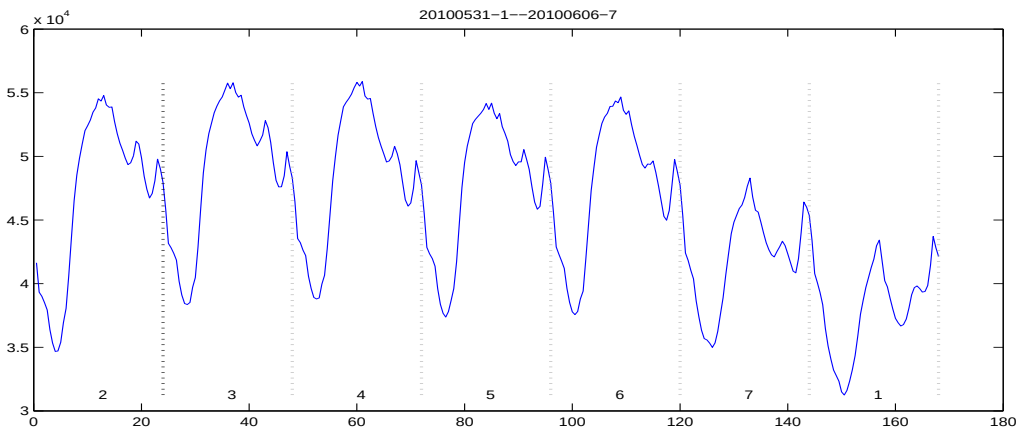
Outline.

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

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

8 years of data:
from January 1^{st} 2003 to August 31^{th} 2010
$T = 2800$ intra day load curves ($n = 48$)



Approximation using a 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: $p = 7$, $E_{MAPE} = 1.4\%$

Apx: November 18^{th} 2007

$S = 12$, $MAPE = 0.0057 = 0.57\%$.
$MAPE = \frac{1}{n} \sum_i^{n=48} |Y_i - \hat{Y}_i| / Y_i$

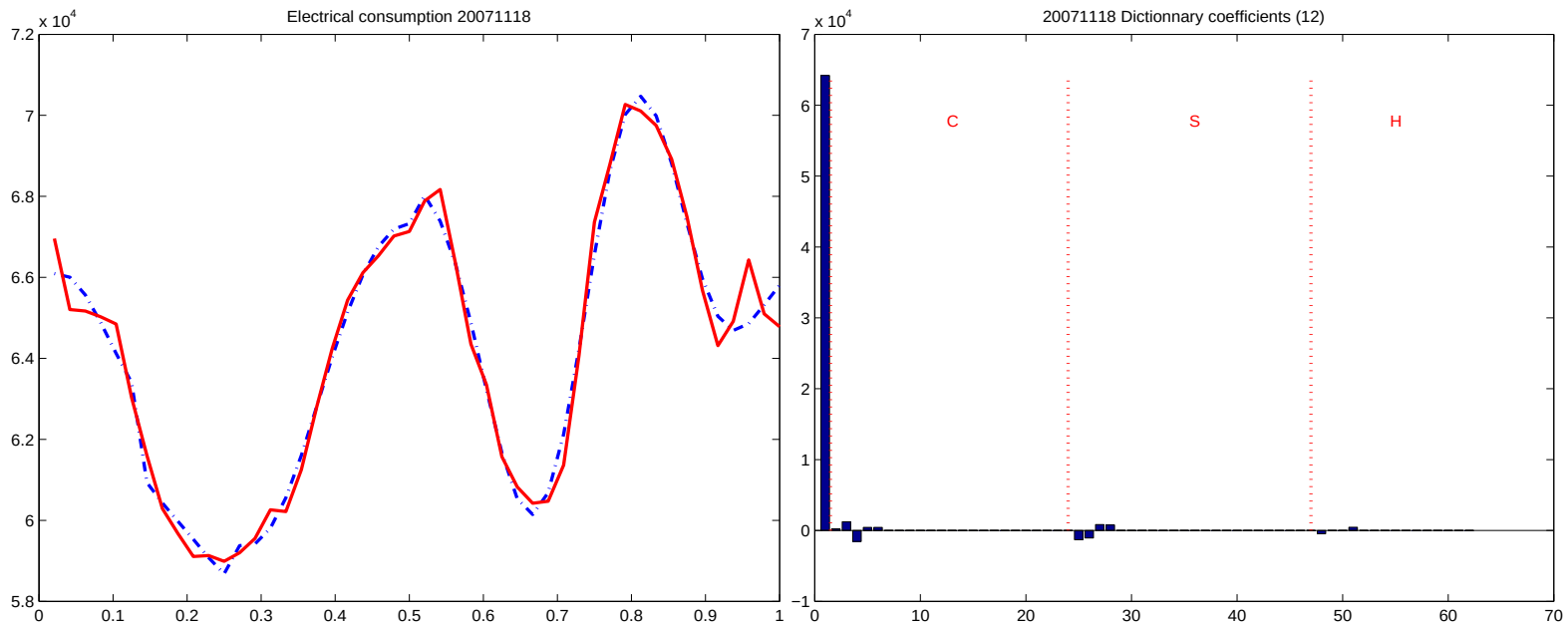


Figure : 2007 11 18

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

Apx: June 17^{th}, 2009

$S = 5$, $MAPE = 0.0147$

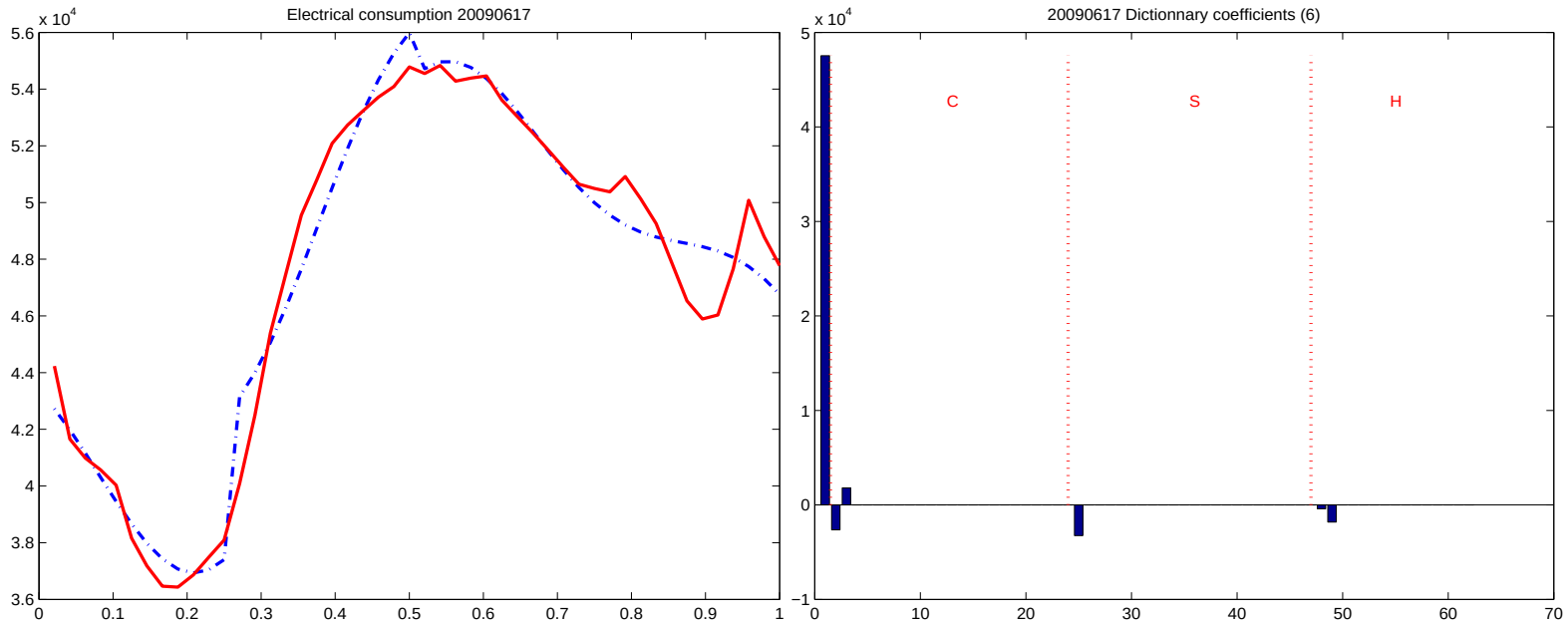


Figure : 2003 04 30

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

right: S coefficients on the dictionary

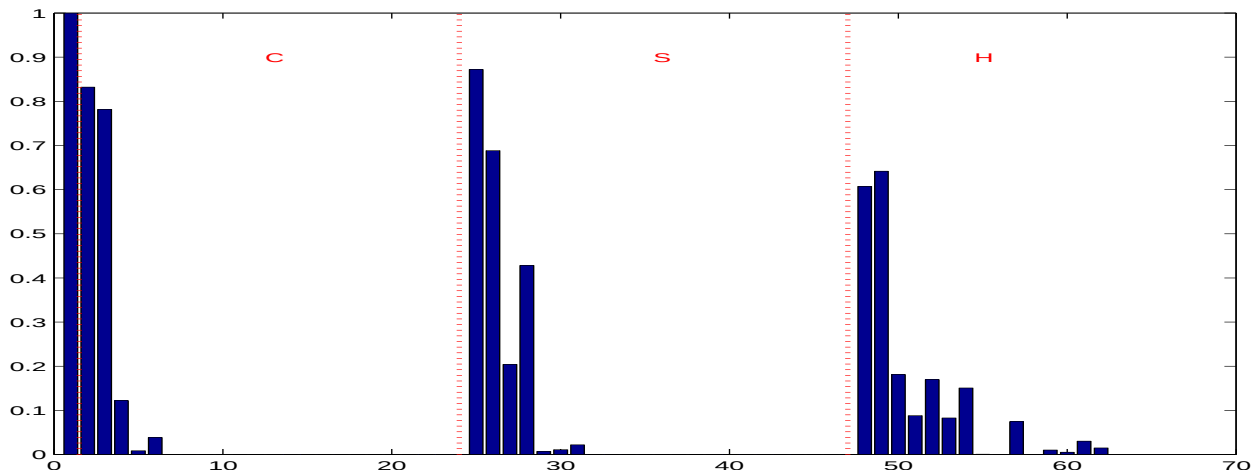
Performances

Comparison between various dictionaries:
MAPE and sparsity average

Dictionary	$\overline{E}$ (%)	$\overline{MAPE}$ (%)	$\overline{S}$
Haar	0.218	3.66	8
Fourier	0.041	1.60	6
DB7	0.192	2.6	9
Mixed	0.034	1.43	7

Support for all coefficients

♯20 using the mixte dictionary (Fourier, Haar)



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

8 years of data:
from January 1^{st} 2003 to August 13^{th} 2010
$T = 2800$ intra day load curves ($n = 48$)

- Sparse approximation of the intra day load curves($\bar{S} = 7$, same support)
- using a clustering algorithm in 2 steps (k-means algorithm)
- Segmentation of the daily signals in clusters
- ...
- From Cluster to Groups using calendar interpretation
- Patterns defined by Group Centroids

Two step Clustering results

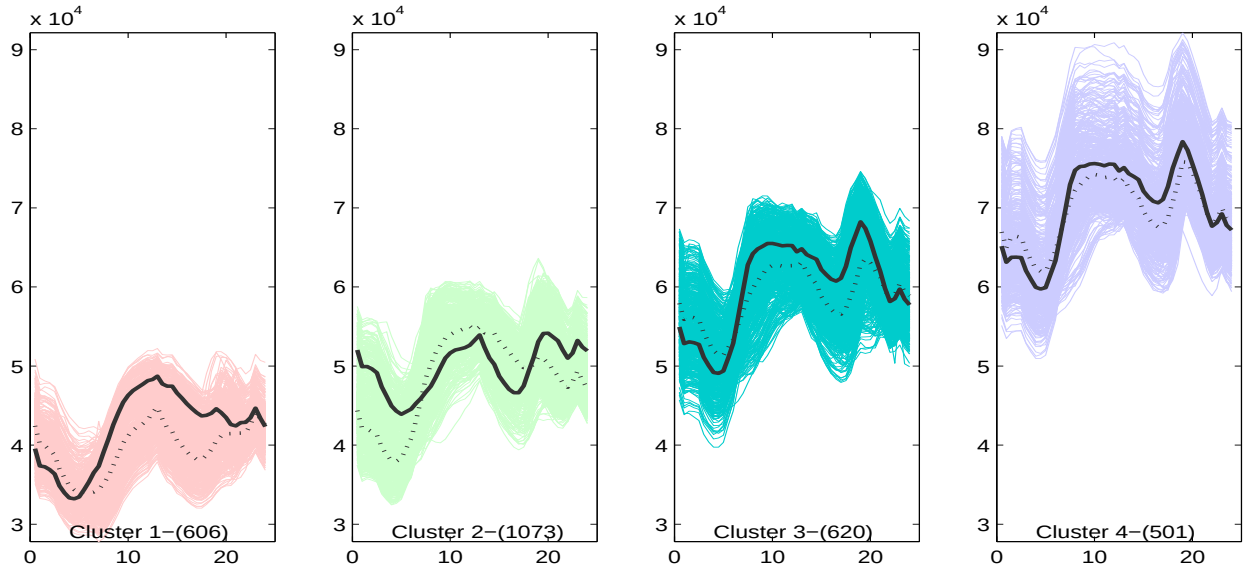
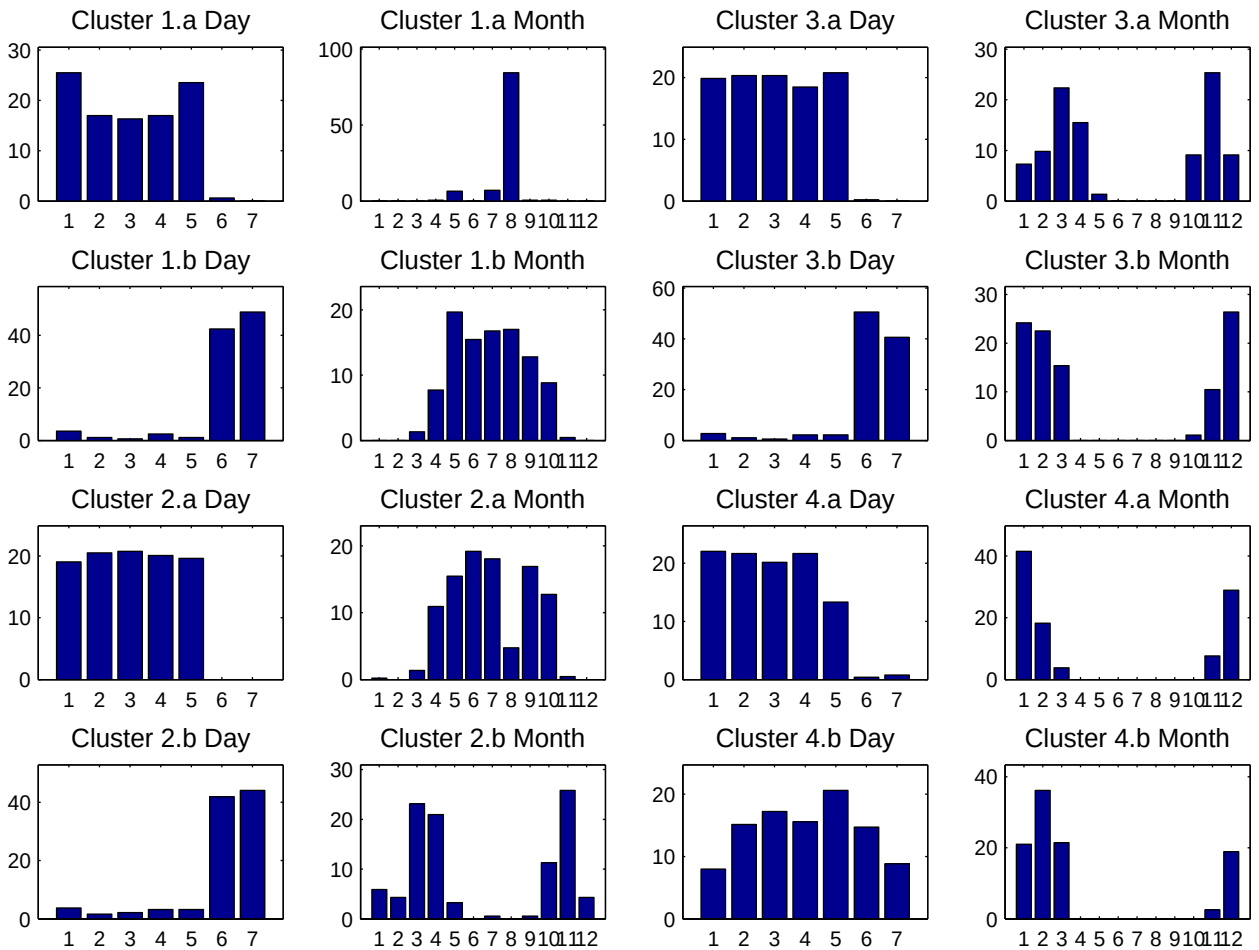


Figure : $T = 2800$ intra day load curves of size $n = 48$ (clustering using $S = 7$ approximated coefficients)

Remarque: stability study for the 4 main clusters

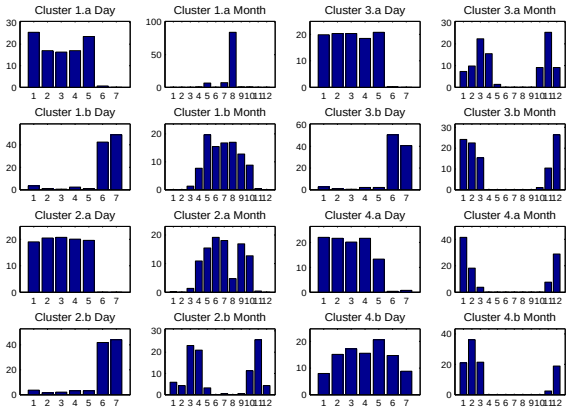
Mining the clusters ...

over days (1...7) and months (1..12) exhibits specific consumption periods



... 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

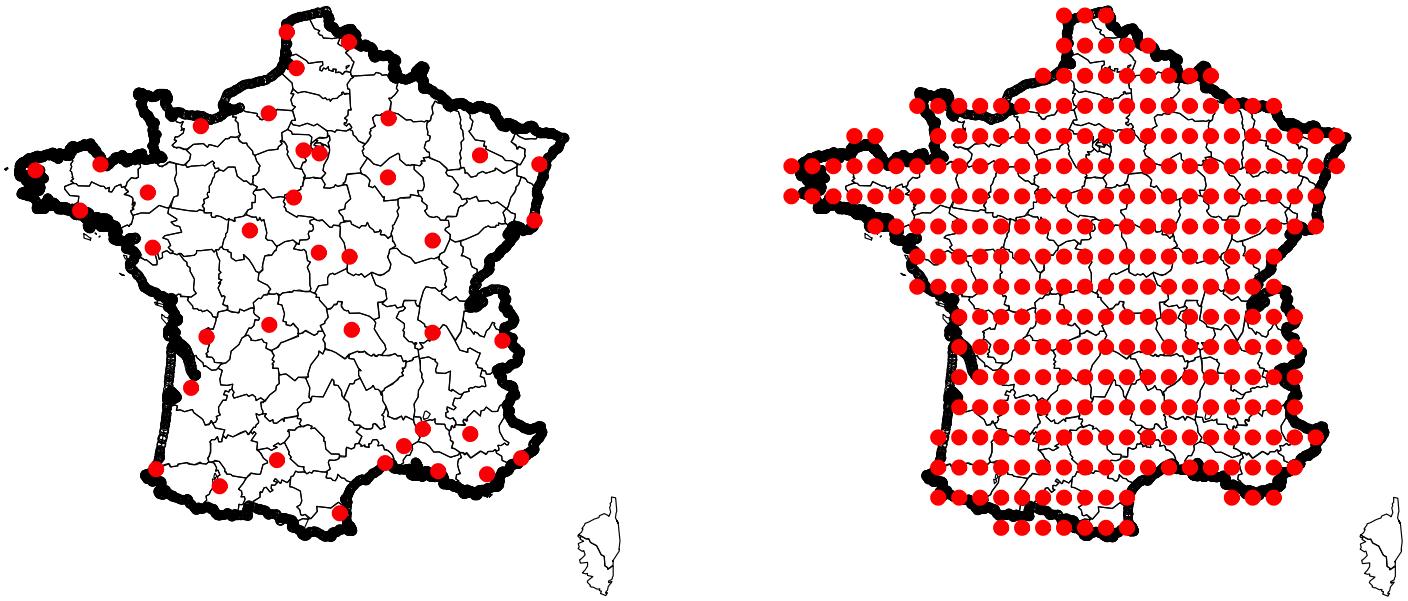


Figure : Temp., Cloud Cover spots (♯39) and wind data (♯293)

Intraday Specific Dictionary: high dimensional model

- Each day t, $\boxed{Y_t = X_t\beta_t + \epsilon_t}$ with $X_t = [P_t\ M_t]$
- First model: $p = 94$
  - $\sharp 2$, $P_t = [C_t B_t]$ (C_t: group centroid, $B_t = Y_{t-7}$)
  - $\sharp 92$, M_t, Meteorogical information (Temperature, Cloud Cover, wind Indicators (min, max, med, std) computed over the 39 meteorological spots and 95% PCA (80: T:5/N:25/W:50).
- Approximation performance:
  - LOL adaptive
    - $S = 13$,
    - $\bar{E}_{MAPE} = 0.34\%$

To remind: Mixte Generic dictionnary: MAPE=1.34%, $\bar{S} = 7$

Intraday Specific Dictionary: low dimensional model

- Each day t, $\boxed{Y_t = X_t\beta_t + \epsilon_t}$ $\quad X_t = [P_t\ M_t]$
  Selected model, $p = 14$.
  1. $P_t = [C_t\ B_t]$ Patterns: $\sharp 2$ group centroid, $B_t = Y_{t-7}$
  2. M_t Meteorological data $\sharp 12$ (Temperature, Cloud Cover, Wind)

- Approximation performance:
  - LOL adaptive
    - $S = 2.5$ [2;8],
    - $\bar{E}_{MAPE} = 1.5\%$ [min 0.002; max0.05]
  - LOL fixed sparsity
    - $S = 5$ [5;5],
    - $\bar{E}_{MAPE} = 0.8\%$ [min 0.0017; max 0.05]

To remind: Mixte Generic dictionnary: MAPE=1.34%, $\bar{S} = 7$

$\rightarrow$: Nice MAPE and Sparsity for approximation

Outline.

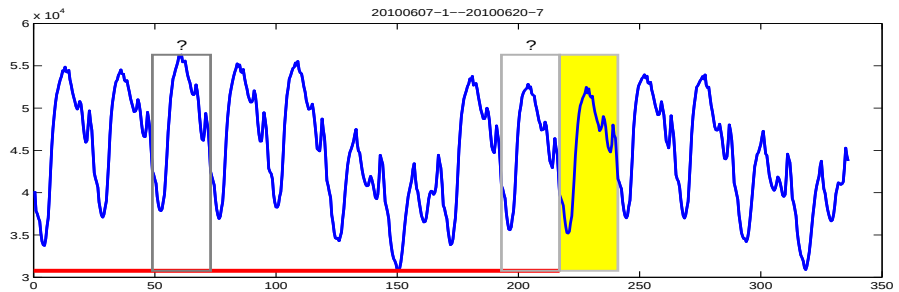
From Sparse Approximation towards Forecast:

- 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
  - Experts dedicated to strategies
  - Aggregation of Experts

From Sparse approximation to Forecast:

$\hat{Y}_t$ Approximation for day t:

- $\hat{Y}_t = X_t\hat{\beta}_t + \delta_{\mathbf{t}}$ with $\quad X_t = [P_t; M_t] \quad P_t = [C_t; B_t]$
- with $n = 48$, $p = 14 = 2 + 3 * 4$



Forecast Expert: $\boxed{\tilde{Y}_t = X_t\ \tilde{\beta}_t}$

- $X_t = [P_t; M_t]$ is supposed to be known
- $\tilde{\beta}_t = \hat{\beta}_{t^*}$
- Plug in estimated coefficients at time $t^* = \mathcal{S}(t) << t$,
- with Strategy $\mathcal{S}$

Specialized Experts focus on

Different Strategies

1. Time depending (t-1, t-7) (2)

—

2. (Meteorological configuration of the day (Temperature) 2)

—

3. Constrained meteorological configuration of the day (Temperature/Cloud Covering)
4. Group constraint meteorological configuration of the day (Temperature/group)
5. Meteorological configuration of the day constrained by the type of the day (Temperature/day)
6. Meteorological configuration of the day constrained by a calendar group (Temperature/calendar)

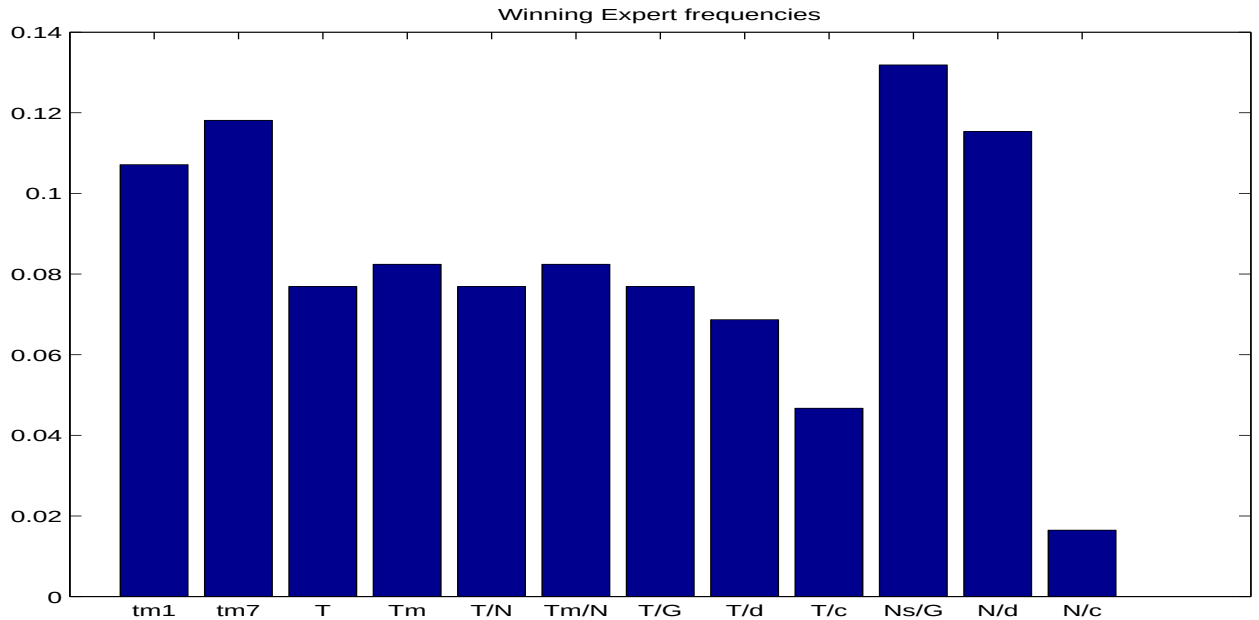
—

7. Meteorological configuration of the day (Cloud cover)
8. group constraint meteorological configuration of the day (Cloud Cover/group)
9. Meteorological configuration of the day constrained by the type of the day (Cloud Cover/day)
10. Meteorological configuration of the day constrained by a calendar group (Cloud Cover/calendar)

$$K = 12$$

Experts hits

Frequencies for the expert to perform at best:



data: form September 1^{st} 2010 to August 31^{th} 2010

$\rightarrow$ The expert are daily competitive

Experts performances

data: form September 1^{st} 2010 to August 31^{th} 2010

Names	mean	median	std
Naive	0.0634	0.0415	0.0514
Oracle	0.0183	0.0151	0.0151
tY	0.0323	0.0262	0.0262
tW	0.0303	0.0239	0.0239
T	0.0305	0.0242	0.0242
Tm	0.0321	0.0264	0.0264
T/N	0.0328	0.0258	0.0258
Tm/N	0.0321	0.0248	0.0248
T/G	0.0337	0.0247	0.0247
T/d	0.0330	0.0257	0.0257
T/c	0.0314	0.0249	0.0249
N/G	0.0297	0.0230	0.0230
N/d	0.0281	0.0219	0.0219
N/c	0.0288	0.0224	0.0224

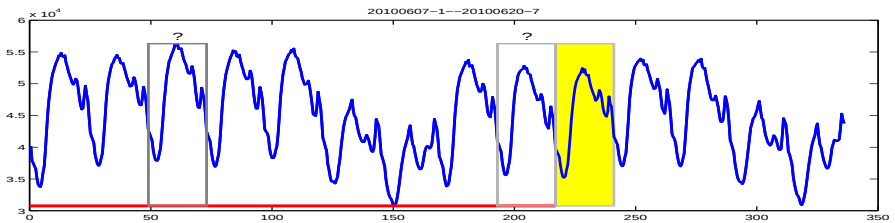
Aggregation and performances

All experts participate to forecast modulated by their performance of approximation given the strategy at $\mathcal{S} = t^*$.

$$\hat{Y}_t = \frac{\sum_{s \in \mathcal{M}} w_t^s \tilde{Y}_t^s}{\sum_{s \in \mathcal{M}} w_t^s}$$

with

- $w_t^s = exp^{-|Y_{t_s^*} - \hat{Y}_{t_s^*}^s|_2^2/\theta}$
- $t_s^* = \mathcal{S}_s(t)$

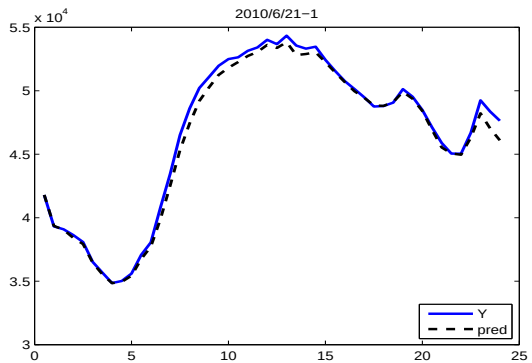


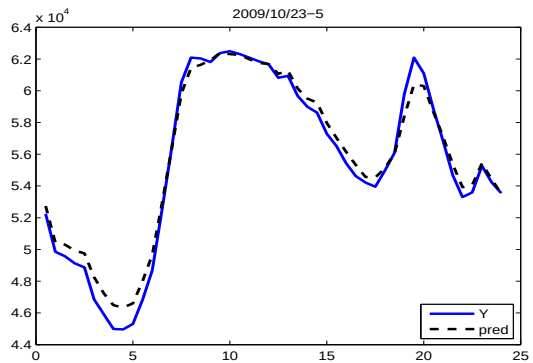
Aggregation performance

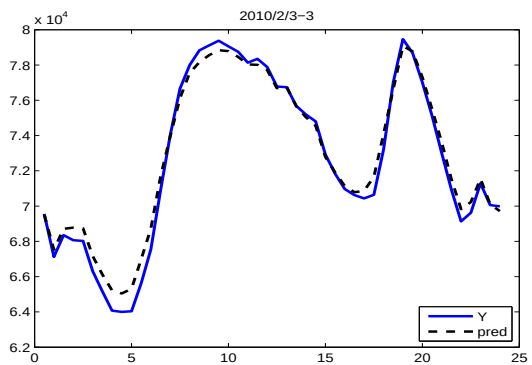
Names	mean	median	std
Naive	0.0634	0.0415	0.0514
Oracle	0.0183	0.0151	0.0151
tY	0.0323	0.0262	0.0262
tW	0.0303	0.0239	0.0239
T	0.0305	0.0242	0.0242
Tm	0.0321	0.0264	0.0264
T/N	0.0328	0.0258	0.0258
Tm/N	0.0321	0.0248	0.0248
T/G	0.0337	0.0247	0.0247
T/d	0.0330	0.0257	0.0257
T/c	0.0314	0.0249	0.0249
N/G	0.0297	0.0230	0.0230
N/d	0.0281	0.0219	0.0219
N/c	0.0288	0.0224	0.0224
AGG	0.0230	0.0197	0.0122

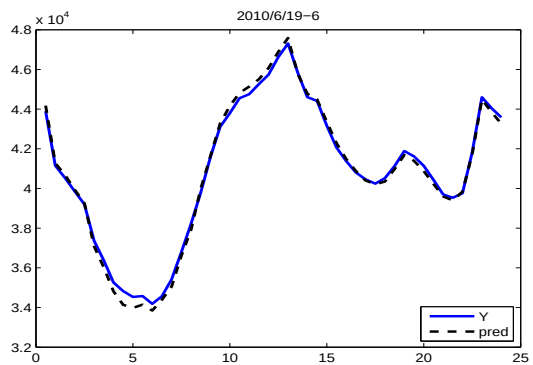
Forecasting some intraday load curves

Some nice intra day forecasting for different periods:









Conclusion and perspectives

- Universal approach for functional data (with "intra day" pattern)
- Sparse approximation using
  - a Generic dictionary for compression and pattern extraction
  - Intra day specific dictionaries for approximation and prediction
- Forecasting
  - Various experts for prediction
  - Agregation using exponential weights
- Competitive approach compared to usual time serie analysis with much less parameters.

Thanks for your attention!

More information on:
sites.google.com/site/mougeotmathilde/research