Forecasting intraday-load curve using sparse learning methods

Dominique Picard

LPMA- Université Paris-Diderot-Paris 7

Collaborators : Mathilde Mougeot UPD, Vincent Lefieux RTE, Laurence Maillard RTE

Numerical methods for high dimensional problems

Pre-big data- framework, towards streaming machine learning

Pre-streaming machine learning

- Volume - moderate
- Variety -moderate
- Velocity -small

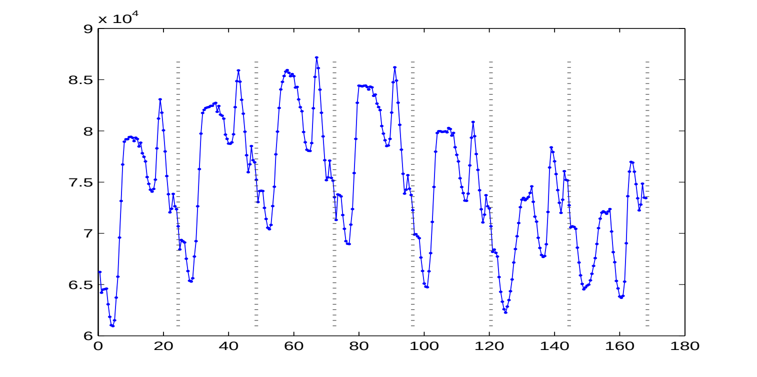
Pre-streaming machine learning

- Volume - moderate
  - smart (data-driven) organisation of the information
  - methods allowing increasing volume of data
- Variety -moderate
  - multidimensional functional data
- Velocity -small

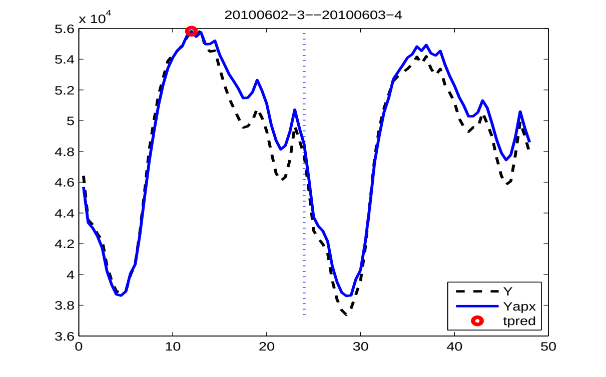
We describe a forecasting pipeline i.e. chain of learning algorithms to achieve a final functional prediction.

Description of the problem

Intraday load curve during a week Monday January 25th to Sunday January 31th



Intraday load curve forecasting -here 48h-



Forecasting pipeline

1. Construction of a ’ smart encyclopedia’ of past scenarios out of a data basis using different learning algorithms.
2. Build a set of prediction experts consulting the encyclopedia.
3. Aggregate the prediction experts

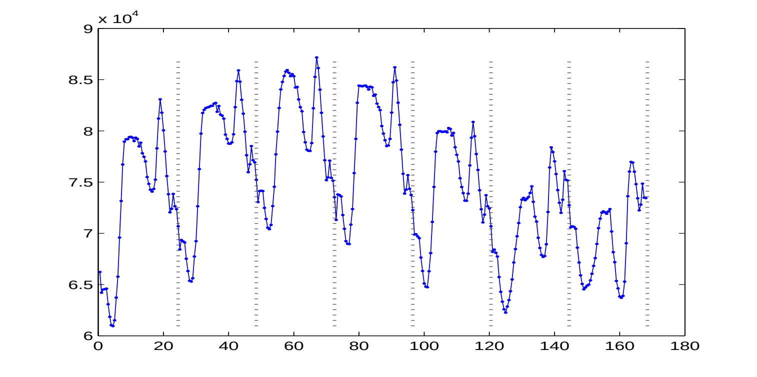
The past data basis

- Electrical consumption of the past
- Other 'shape variables': calendar data, functional bases
- Meteorological input

Electrical consumptions of the past

- Recorded every half hour from January 1^{st}, 2003 to August 31^{th}, 2010.

- For this period of time, the global consumption signal is split into $N = 2800$ sub signals $(Y_1, \ldots, Y_t, \ldots, Y_N)$. $Y_t \in R^n$, defines the intra day load curve for the t^{th} day of size $n = 48$.


x 10^4
9
8.5
8
7.5
7
6.5
6
0
20
40
60
80
100
120
140
160
180

Shape and seasonal effects

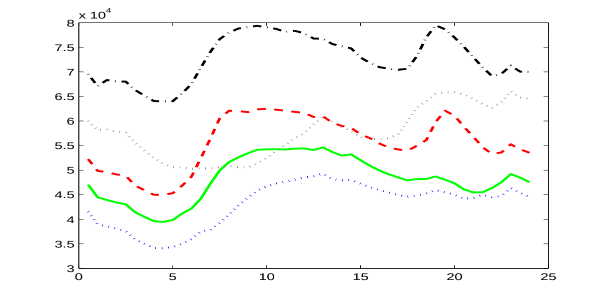


Figure : Intraday load curves for various days. 2010-02-03 winter: black dashed dot line, 2010-05-21 spring: red dashed line, 2009-10-23 autumn: green solid line, 2010-08-19 summer: blue dot line, 2010-01-01 public day: gray dot line.

Calendar and functional effects (endogenous)

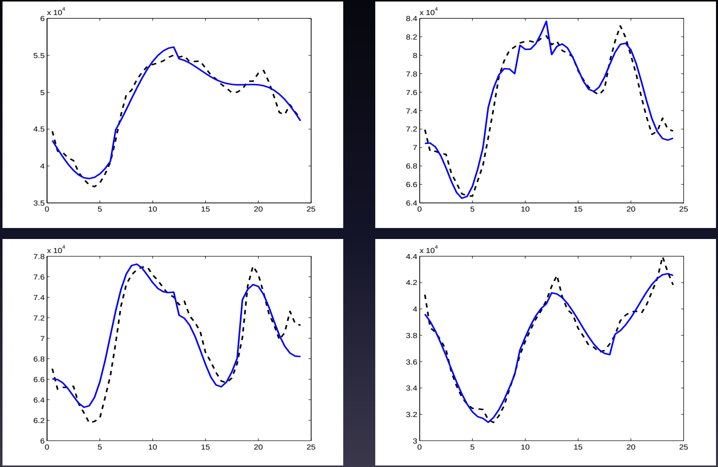


Figure : autumn, winter, spring and summer

Calendar and functional effects (shape description)

- Consumption on day T can be explained by consumption of days $t' < T$ of the past.
- can be explained by calendar values of the day T (monday,..., sunday, months, seasons,...
- Is a function of time and can be expressed in a standard dictionary of functions (wavelets, Fourier,..)

Functional aspect : dictionary

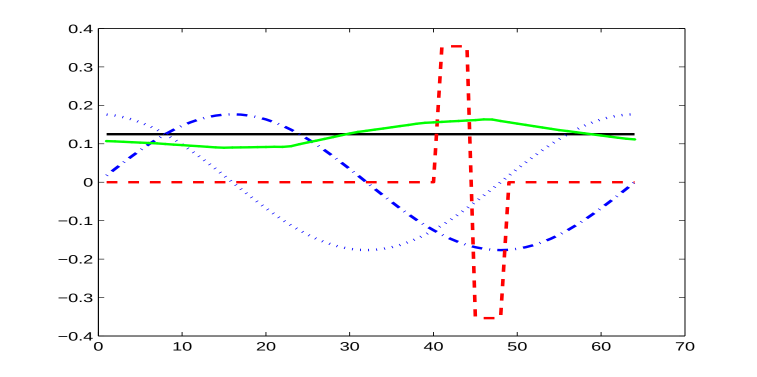


Figure : Functions of the dictionary. Constant (black-solid line), cosine (blue-dotted line), sine (blue-dashdot line), Haar (red-dashed line) and temperature (green-solid line with points) functions.

Meteorological inputs: Exogeneous variables

- A total of 371 (=2x39+293) meteorological variables
- recorded each day half-hourly over the 2800 days of the same period of time.

Temperature:

T^{k} for $\mathrm{k} = 1, \ldots, 39$ measured in 39 weather stations scattered all over the French territory.

Cloud Cover:

N^{k} for $\mathrm{k} = 1, \ldots, 39$ measured in the same 39 weather stations.

Wind:

$\mathrm{W}^{\mathrm{k}'}$ for $\mathrm{k}' = 1, \ldots, 293$ available at 293 network points scattered all over the territory.

Weather stations

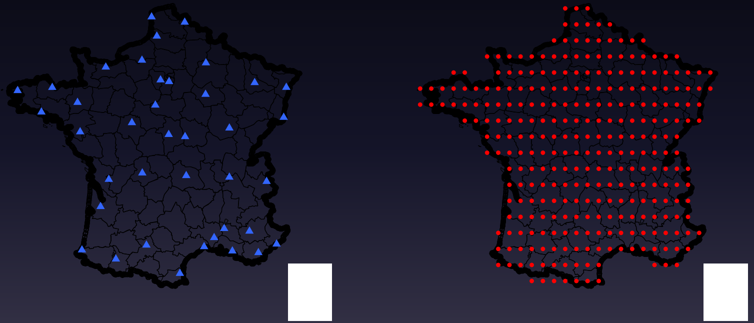
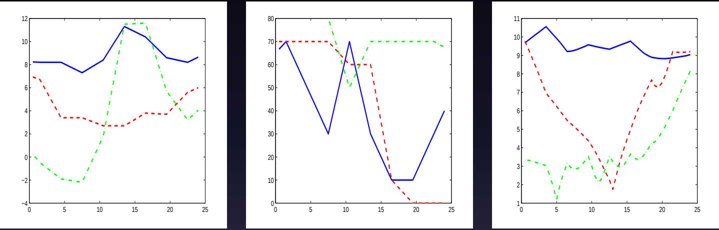


Figure : Temperature and Cloud covering measurement stations. Wind stations

Brest- Lille- Marseille



(a) T (b) CC (c) W

Figure : Brest (blue line), Lille (red line) and Marseille (green).

Main issues

1. Large dimension
2. Prediction requires to explain with a small number of predictable parameters
3. Most of the potentially explanatory variables (load curve, meteo, functions of the dictionary) are highly correlated

Reduced set of explanatory variables

For each t index of the day of interest, we register the daily electrical consumption signal Y_t and

$$Z_t = [[C]_t \; [M]_t]$$

$[C]_t$ is the concatenation of the "calendar,functional, past-consumptions" variables and $[M]_t$ "meteo variables".

Sparse approximation on the learning set

Sparse Approximation of each consumption day on a learning set of days (2003-2010), using the set of potentially explanatory variables.

- For each day t of the learning set, we build an approximation $\hat{Y}_t$ of the (observed) signal Y_t with the help of the set of explanatory variables (Z_t):
- $\hat{Y}_t = G_t(Z_t)$

$$G_t(Z_t) = Z_t\hat{\beta}_t$$

(*) Sparse Approximation and Knowledge Extraction for Electrical Consumption Signals, 2012,

M. Mougeot, D. P., K. Tribouley & V. Lefieux, L. Teyssier-Maillard

High dimensional Linear Models

$$Y = X\beta + \epsilon$$

$\beta \in R^k$ is the unknown parameter (to be estimated)

- $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^*$ is a (non observed) vector of random errors. It is assumed to be variables i.i.d. $N(0, \sigma^2)$
- X is a known matrix $n \times k$.

High dimension : $k >> n^t$

Forecasting procedure

Forecasting using the encyclopedia

- Construction of a set of forecasting experts.
- Aggregation of the experts.

Expert associated to the strategy $\mathcal{M}$

Forecasting experts

- Strategy : $\mathcal{M}$ a function, data dependent or not, from $\mathbb{N}$ to $\mathbb{N}$ such that for any $d \in \mathbb{N}, \mathcal{M}(d) < d$ (purely non anticipative).
- Plug-in To the strategy $\mathcal{M}$ we associate the expert $\tilde{Y}_t^{\mathcal{M}}$: the prediction of the signal of day t using forecasting strategy $\mathcal{M}$,

$$\tilde{Y}_t^{\mathcal{M}} = G_{\mathcal{M}(t)}(Z_t) = Z_t \hat{\beta}_{\mathcal{M}(t)}$$

Examples of strategies : time depending

tm1: Refer to the day before: (The coefficients used for prediction are those calculated the previous day)
$\mathcal{M}(\mathrm{d}) = \mathrm{d} - 1$

$$\tilde{Y}_t^{tm1} = Z_t \hat{\beta}_{t-1}$$

tm7: Refer to one week before:
$\mathcal{M}(\mathrm{d}) = \mathrm{d} - 7$

$$\tilde{Y}_t^{tm7} = Z_t \hat{\beta}_{t-7}$$

Experts introducing meteorological scenarios

- T: Find the day having the closest temperature indicators, regarding the sup distance (over the days, and over the indicators):

  $$\mathcal{M}(\mathrm{d}) = \mathrm{ArgMin}_{\mathrm{t}} \sup_{\mathrm{k}\in\{1,\dots,6\},\, \mathrm{i}\in\{1,\dots,48\}} |\mathrm{T}^{\mathrm{k}}_{\mathrm{d}}(\mathrm{i}) - \mathrm{T}^{\mathrm{k}}_{\mathrm{t}}(\mathrm{i})|$$

- T_m: Find the day having the closest median temperature with the sup distance (over the days):

  $$\mathcal{M}(\mathrm{d}) = \mathrm{ArgMin}_{\mathrm{t}} \sup_{\mathrm{i}\in\{1,\dots,48\}} |\mathrm{T}^{3}_{\mathrm{d}}(\mathrm{i}) - \mathrm{T}^{3}_{\mathrm{t}}(\mathrm{i})|$$

MAPE error

For day t, the prediction MAPE error over the interval $[0, T]$ is defined by:

$$\mathrm{MAPE}(Y, \tilde{Y}_t^{\mathcal{M}})(T) = \frac{1}{T}\sum_{i=1}^{T} \frac{|\tilde{Y}_t^{\mathcal{M}}(i) - Y_t(i)|}{Y_t(i)}$$

$$\mathrm{MISE}(Y, \tilde{Y}_t^{\mathcal{M}})(T) = \frac{1}{T}\sum_{i=1}^{T} |\tilde{Y}_t^{\mathcal{M}}(i) - Y_t(i)|^2$$

Prediction evaluation

M	mean	med	min	max
Naive	0.0634	0.0415	0.0046	0.1982
Apx	0.0129	0.0104	0.0023	0.0786
tm1	0.0340	0.0281	0.0063	0.1490
tm7	0.0327	0.0258	0.0054	0.2297
T	0.0306	0.0263	0.0058	0.1085
Tm	0.0329	0.0275	0.0047	0.2020
T/N	0.0347	0.0293	0.0056	0.1916
Tm/N	0.0358	0.0300	0.0054	0.2156
T/G	0.0323	0.0271	0.0050	0.1916
T/d	0.0351	0.0278	0.0053	0.1916
T/c	0.0340	0.0259	0.0053	0.1937
Ns/G	0.0322	0.0251	0.0049	0.2078
N/d	0.0305	0.0239	0.0042	0.1449
N/c	0.0307	0.0237	0.0042	0.1990

Prediction evaluation-Comparing experts

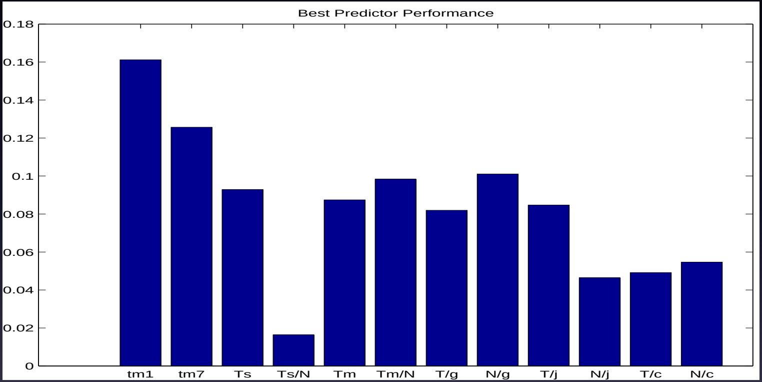


Figure : Percentage of best predictor

Prediction evaluation-Comparing experts on days

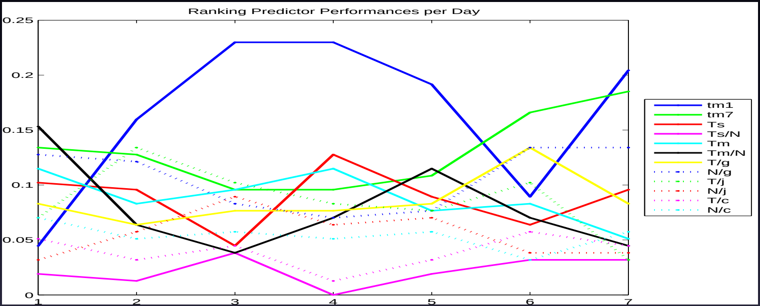


Figure : Percentage of best predictor among days (1:monday, . . . 7:sunday)

Prediction evaluation-Comparing experts

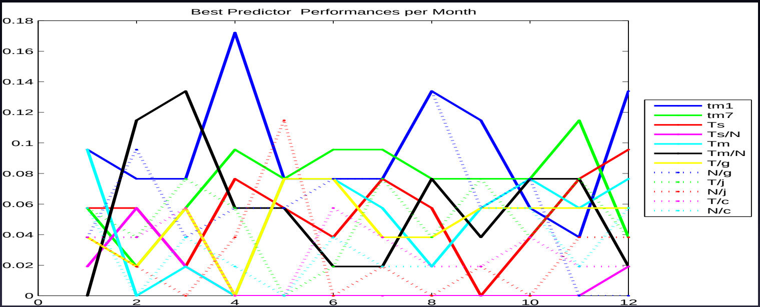


Figure : Percentage of best predictor among month

Aggregation of predictors: Exponential weights

(inspired by various theoretical results -see Lecue, Rigollet, Stolz, Tsybakov,...-)

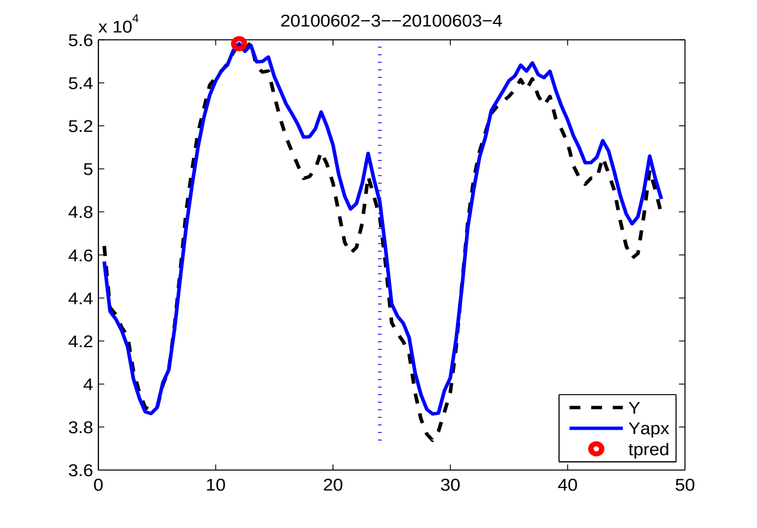
$$\tilde{Y}_d^{wgt*} = \frac{\sum_{m=1}^{M} w_d^m \tilde{Y}_d^m}{\sum_{m=1}^{M} w_d^m}$$

with

$$w_d^{\mathcal{M}} = \exp(-\frac{1}{T\theta}\sum_{i=1}^{T}|\tilde{Y}_d^{\mathcal{M}}(i) - Y_t(i)|^2)$$

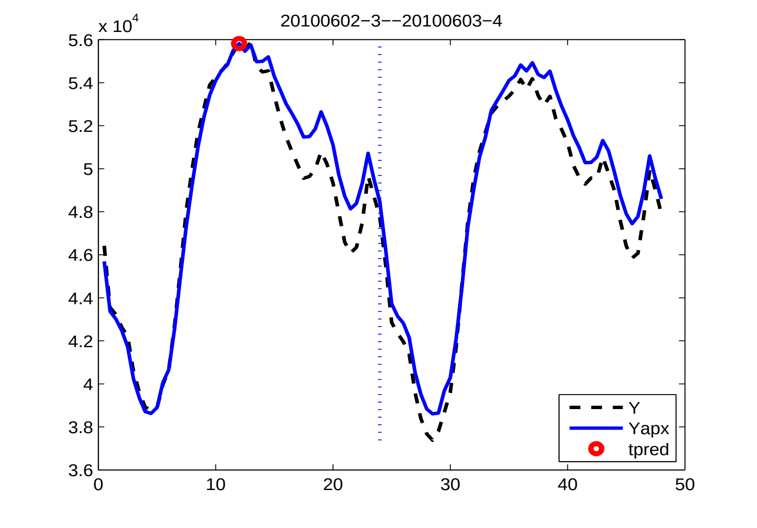
θ is a parameter, (often called temperature in physic applications, see the discussion below) $T = Tpred$.

(mape=0.7%).



20100602−3−−20100603−4
x 10^4
5.6
5.4
5.2
5
4.8
4.6
4.4
4.2
4
3.8
3.6
0
10
20
30
40
50
Y
Yapx
tpred

(mape=0.7%).


20100602−3−−20100603−4
x 10^4
5.6
5.4
5.2
5
4.8
4.6
4.4
4.2
4
3.8
3.6
0
10
20
30
40
50
Y
Yapx
tpred

Winter forecast

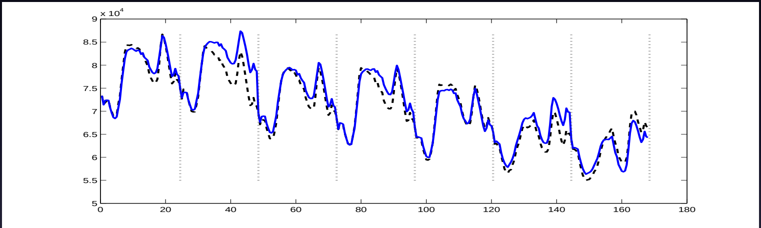


Figure : Forecast (solid blue line) and observed (dashed dark line) electrical consumption for a winter week from Monday February 1st to Sunday January 7th 2010.

Spring forecast

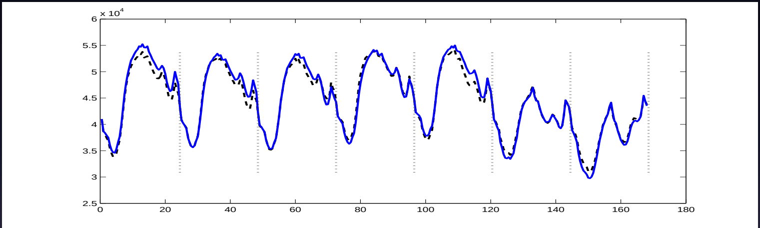


Figure : Forecast (solid blue line) and observed (dashed dark line) electrical consumption for a spring week from Monday June 14th to Sunday June 21th 2010.

Sparse methods- collinearity- structure

High dimensional Linear Models

$$Y = X\beta + \epsilon$$

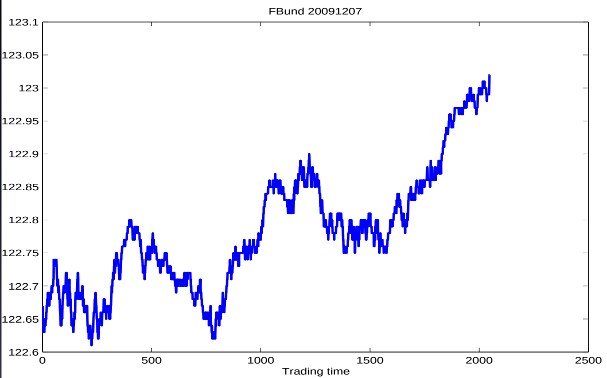
$\beta \in \mathrm{R}^k$ is the unknown parameter (to be estimated)

- $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^*$ is a (non observed) vector of random errors. It is assumed to be variables i.i.d. $N(0, \sigma^2)$
- X is a known matrix $n \times k$.

High dimension : $k >> n^t$

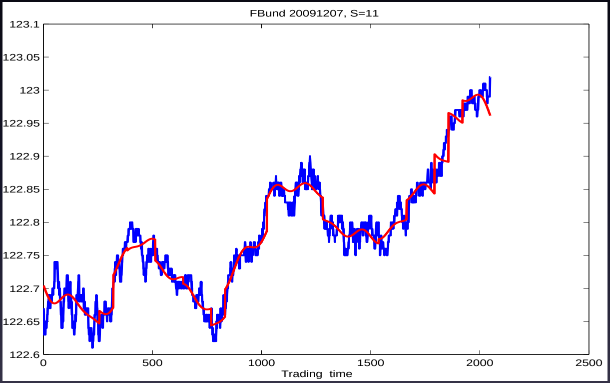
(*) M. Mougeot, D. P., K. Tribouley, JRSS B 2012,B Stat. Methodol. vol 74

FBUND sparse reconstruction



M. Mougeot

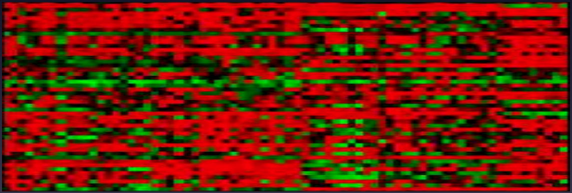
FBUND sparse reconstruction



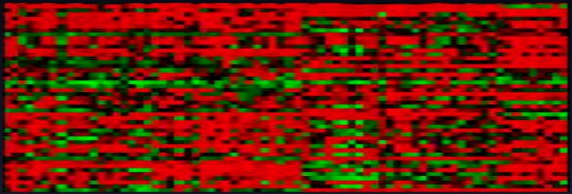
M. Mougeot

Genomic example

$$Y = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \end{pmatrix}$$

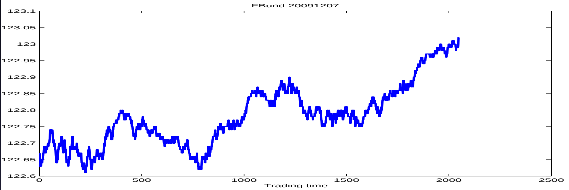
$X =$ 

The matrix X : genomic

$X =$ 

- X : expression of different genes
  behaves like $n \times p$ random variables i.i.d. $N(0, 1)$. (large random matrices)

Signal denoising

Y = 

What is X in this case ?

- Statistical learning, regression estimation

$$Y_i = f(t_i) + \epsilon_i + u_i,\ i = 1 \dots n$$

  - ϵ_i's are i.i.d. $N(0, 1)$.
  - u_i's possibly random, not necessarily random nor iid but 'small'.
  - t_i are observation times ($t_i = \frac{i}{n}$).
  - f is the parameter to be estimated.

Using a dictionary

To estimate f, we consider a dictionary $\mathcal{D}$ of size $\#\mathcal{D} = \mathrm{p}$

$$\mathcal{D} \; = \; \{\mathrm{g}_1, \ldots \mathrm{g}_\mathrm{p}\}$$

and assume that f can be well fitted by this dictionary.

$$\mathrm{f} = \sum_{\ell=1}^{\mathrm{p}} \beta_\ell \, \mathrm{g}_\ell + \mathrm{h} \tag{1}$$

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

Modeling

Which coincide with the following model:

$$Y = X\beta + u + \epsilon$$

if we put $u_i = h(t_i)$ and

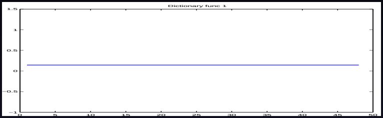
$$X = \begin{pmatrix} g_1(t_1) & \dots & g_p(t_1) \\ . & . & . \\ . & . & . \\ g_1(t_n) & \dots & g_p(t_n) \end{pmatrix}$$

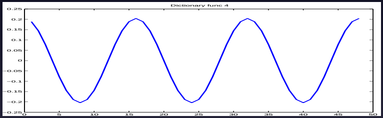
The dictionary problem

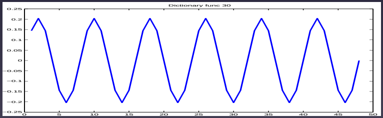
Of course sparsity is linked with the dictionary.

- Fourier Basis
- Wavelet basis
- Needlets
- Combination of 'bases'

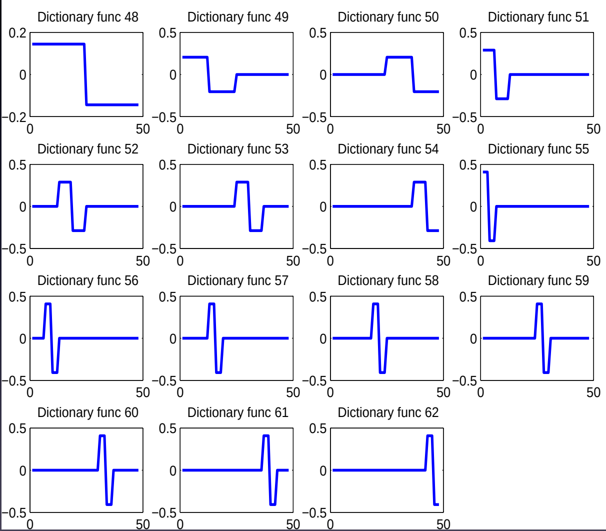
Fourier basis







Haar wavelets



Conditions generally required to solve the problem

- 'Sparsity'. conditions on the vector β
- Conditions on the matrix X (not too high collinearities, RIP...

Restricted identity property

For $\mathcal{C} \subset \{1, \ldots p\}$, denote $X_{\mathcal{C}}$ the matrix X restricted to the raws which are in $\mathcal{C}$ and the associated Gram-matrix

$$M(\mathcal{C}) := \frac{1}{n} X_{\mathcal{C}}^t X_{\mathcal{C}}$$

Restricted identity property. means that $M(\mathcal{C})$ is almost the identity matrix for any $\mathcal{C}$ small enough.

Example 1: RIP

$\mathrm{RIP}(\mathrm{m}_0, \nu)$ assumes that
There exist $0 \leq \nu < 1$ and $\mathrm{m}_0 \geq 1$ such that :

$$\forall \mathrm{x} \in \mathrm{R}^{\mathrm{m}}, \quad \|\mathrm{x}\|^2_{\mathrm{l}_2(\mathrm{m})}(1-\nu) \leq \mathrm{x}^t \mathrm{M}(\mathcal{C})\mathrm{x} \leq \|\mathrm{x}\|^2_{\mathrm{l}_2(\mathrm{m})}(1+\nu),$$

Example 2: Coherence condition

- $$M := \frac{1}{n} X^t X.$$
- $M_{jj} = 1$ for all j.
- Coherence.

$$\tau_n = \sup_{\ell \neq m} |M_{\ell m}| = \sup_{\ell \neq m} |\frac{1}{n} \sum_{i=1}^{n} X_{i\ell} X_{im}|$$

Coherence $\implies$ RIP$(\lfloor \nu/\tau_n \rfloor, \nu)$

Sparsity conditions

Sparsity conditions

$$\#\{\ell \in \{1,\ldots,k\},\ |\beta_\ell| \neq 0\} \ \leq \mathrm{S}$$

$$\sum_\ell |\beta_\ell|^{\mathrm{q}} \leq \mathrm{M}, \quad 0 < \mathrm{q} < 1 \ (\mathrm{B_q}(\mathrm{M}))$$

SMALL NUMBER OF BIG COEFFICIENTS

Penalization for sparsity

Many penalizations introduced historically in the regression framework (to put identification constraints on β)

- Ridge: $\mathrm{E}(\beta, \lambda) = ||\mathrm{Y} - \mathrm{X}\beta||^2 + \lambda \Sigma_j \beta_j^2$
- Lasso: $\mathrm{E}(\beta, \lambda) = ||\mathrm{Y} - \mathrm{X}\beta||^2 + \lambda \Sigma_j |\beta_j|$
- Scad: $\mathrm{E}(\beta, \lambda) = ||\mathrm{Y} - \mathrm{X}\beta||^2 + \lambda \Sigma_j w_j g(\beta_j)$

Solutions based on:
$\rightarrow$ Convex Optimization for l_2, l_1, non convex Opti. for Scad
Candes & Tao (2007), Fan & Lv (2008, 2010), ...
Many others...

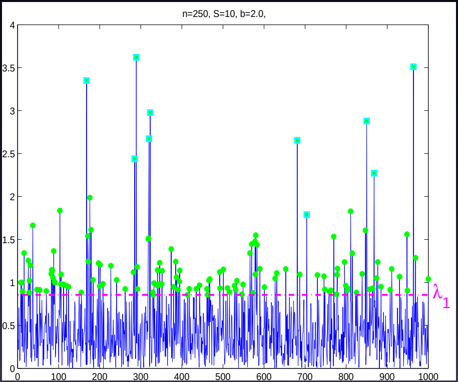
Fast greedy methods : 2-step thresholding procedures

$Y = X\beta + \epsilon \quad Y\ (n \times 1),\ X\ (n \times k)$

steps		compute	size
Step 1=pre-selection	Find b Leaders $b < n << k$	X_b	(n, b)
Least squares	on Leaders	$\tilde{\beta} = (X_b^* X_b)^{-1} X_b^* Y$	$(1, b)$
Step 2=denoising	the coefficients	$\hat{\beta}$	$(1, \hat{S})$

LOL : coefficient-wise : Step1

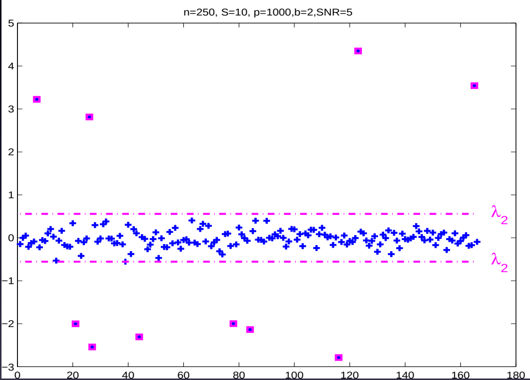
$B = \{\ell,\ \mathcal{K}_\ell \geq \lambda_1\}, \qquad \mathcal{K}_\ell = |\frac{1}{n}\sum_{i=1}^{N} X_{i\ell}Y_i|$



$n = 250,\ p = 1000,\ X$ i.i.d. $\mathcal{N}(0,1),\ S = 10$

$\mathrm{card}(B) = 170 >> S$

LOL: step 3



Structuring

$$Y = X\beta + \epsilon,\ X\ : N \times k$$

We decide to re-arrange the k predictors into p ($p \leq k$) groups of variables

$$X = [X_{\mathcal{G}_1}, \dots, X_{\mathcal{G}_p}]$$

where $\mathcal{G}_1, \dots, \mathcal{G}_p$ is a partition of $\{1, \dots, k\}$.

$$X_\ell = X_{(j,t)},\ X_{\mathcal{G}_j} = [X_{(j,1)}, \dots, X_{(j,|\mathcal{G}_j|)}]$$

- $j \in \{1, \dots, p\}$ is the index of the group $\mathcal{G}_j$
- t is the altitude (height) of ℓ inside the group $\mathcal{G}_j$.

Structured Sparsity

- Structured Sparsity

$$\sum_{j=1}^{p} w_j \|\beta\|_{\mathcal{G}_j,r}^{q} = \sum_{j=1}^{p} w_j [\sum_{t=1}^{T} |\beta_{(j,t)}|^r]^{q/r} \leq (M)^q.$$

$$\text{if } w_j = 1,\ r \geq q,\ \sum_{j=1}^{p} \|\beta\|_{\mathcal{G}_j,r}^{q} = \sum_{j=1}^{p} [\sum_{t=1}^{T} |\beta_{(j,t)}|^r]^{q/r} \leq \sum_{j,t} |\beta_{(j,t)}|^q$$

• Structured sparsity generally less stringent than ordinary one

• Means we require a small number of 'big' groups

Example of structure: Wavelet-grouping

- Block thresholding (global blocks)

$$\beta = (\beta_{jk})$$

$$\mathcal{G}_j = \{(j,k),\ 0 \le k \le 2^j\}, 0 \le j \le p$$

Size of $\mathcal{G}_j = 2^j$,
Sparsity = Besov(s, r, q)(M)

GR-LOL - step1

The columns of X are again normalized

$$\frac{1}{n}\sum_{i=1}^{N} X_{i(j,t)}^2 = 1, \ \forall\,(j,t).$$

"grouped correlation" search and thresholding:

$$\mathcal{K}_{(j,t)} = |\frac{1}{n}\sum_{i=1}^{N} X_{i(j,t)} Y_i| \qquad \forall\,(j,t),\ 1 \le j \le p,\ 1 \le t \le T$$

$$\rho_j^2 = \sum_{t=1,\dots,T} \mathcal{K}_{(j,t)}^2, \qquad T = \max|\mathcal{G}_j|$$

GR-LOL - step1

- $\lambda(1)$ is tuning parameter
- $$\mathcal{B} = \left\{ \mathrm{j} = 1, \ldots, \mathrm{p},\ \rho_{\mathrm{j}}^2 \geq \lambda(1)^2 \right\} \tag{2}$$
- $\mathcal{G}_{\mathcal{B}} = \cup_{\mathrm{j} \in \mathcal{B}} \mathcal{G}_{\mathrm{j}}$.

GR-LOL - step2

OLS on the block-leaders by considering the new pseudo-linear model

$$\mathrm{Y} = \mathrm{X}_{\mathcal{G}_{\mathcal{B}}} \beta_{\mathcal{G}_{\mathcal{B}}} + \mathrm{error}.$$

$$\hat{\beta}_{\mathcal{G}_{\mathcal{B}}} = \hat{\beta}(\mathcal{B}) \quad \text{and} \quad \hat{\beta}_{\mathcal{G}_{\mathcal{B}}^{c}} = 0$$

where

$$\hat{\beta}(\mathcal{B}) = [\mathrm{X}^{t}_{\mathcal{G}_{\mathcal{B}}} \mathrm{X}_{\mathcal{G}_{\mathcal{B}}}]^{-1} \mathrm{X}_{\mathcal{G}_{\mathcal{B}}} \mathrm{Y}.$$

GR-LOL - step3 Block-Thresholding

- $\lambda(2)$ is another tuning parameter.
- We apply the second threshold on the estimated coefficients

$$\forall \ell = (\mathrm{j}, \mathrm{t}) \in \{1, \ldots, \mathrm{k}\}, \quad \hat{\beta}^*_\ell = \hat{\beta}_\ell \, \mathbb{I}\{ \, \|\hat{\beta}\|_{\mathcal{G}_{\mathrm{j}},2} \geq \lambda(2) \, \}$$

- $$\|\hat{\beta}\|^2_{\mathcal{G}_{\mathrm{j}},2} := \sum_{0 \leq \mathrm{t} \leq \mathrm{T}} \hat{\beta}^2_{(\mathrm{j},\mathrm{t})}.$$

Boosting the convergence by grouping

$Y = X\beta + \epsilon \quad Y\ (N \times 1),\ X\ (N \times k) \quad p$ groups

- Calculate the (internal) correlations of the columns of the matrix X as well as their (external) correlation with the target Y.
- Put columns which are highly correlated (internal correlation) in different groups
- Gather the columns with typically close correlation to the target (external correlation)
- Make T (number of groups) as small as possible

Boosting the rates : cut off

- Divide the columns of X into two sets : S_1 : highly correlated, S_2 : weakly correlated.
- Put S_1 as 'group beginners' (each of them has smallest altitude in its group) to separate them.
- Choose the cut off between S_1 and S_2.
- Fill the groups with affinity with the delegate in terms of $\mathcal{K}_l = |\frac{1}{n}\sum_{i=1}^{N} X_{il}Y_i|$: Gathering the columns with typically close correlation with the target

Back to electrical consumption

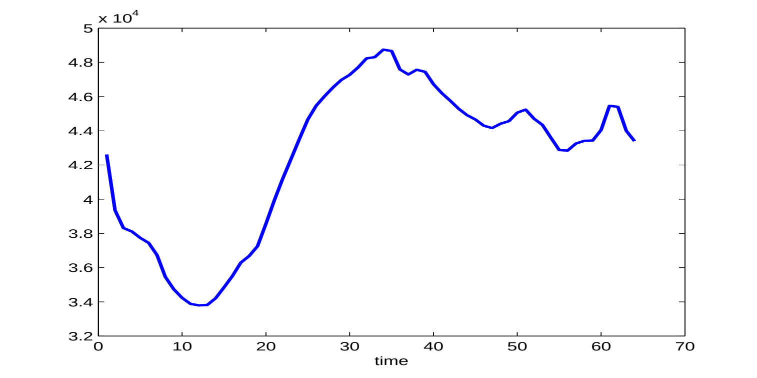


Figure : French consumption

Temperatures

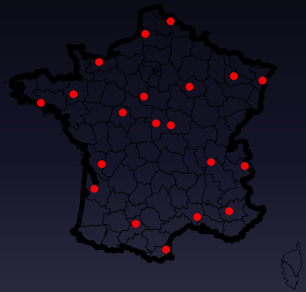


Figure : Temperature spots

Dictionary

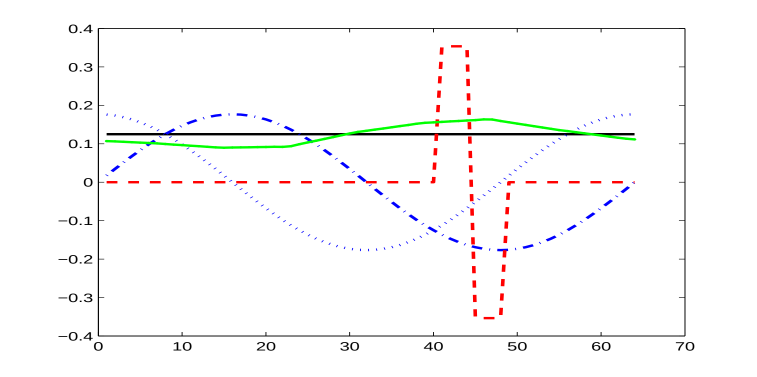


Figure : Functions of the dictionary. Constant (black-solid line), cosine (blue-dotted line), sine (blue-dashdot line), Haar (red-dashed line) and temperature (green-solid line with points) functions.

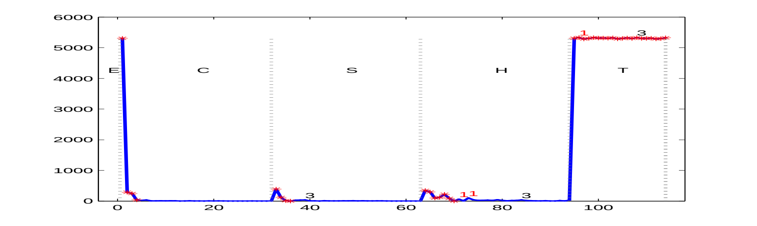


Figure : "Correlation" between the consumption signal and the various dictionary functions. The chosen delegates are tagged with a red star.

- For LOL, E = 1.86% (×24) selected functions: T-T-C-T-H-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-S-C and are meaningful functions (20 : T), (2 : C), (1 : S), (1 : H) .
- Group LOL: E = 0.75%, 24 regressors / 8 groups THS-THH-THH-TCS-TCST-HHTC-STSH. meaningful functions (8 : T); (3 : C); (5 : S); (8 : H).

Approximation

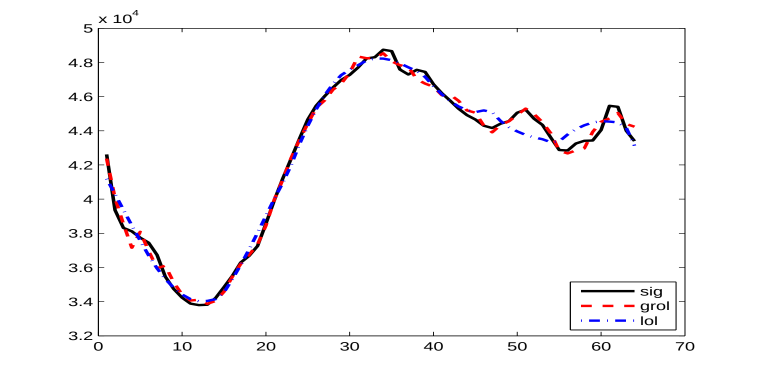


Figure : Model of the consumption signal (black-solid line) using GROL (red-dashed line) and LOL (blue dot dashed line).

Pre-processing the explanatory variables

Reduced dictionary : endogeneous variables : patterns

- Represent sparsely each day on the dictionary (H, S, C).
- Use K-means algorithm to cluster this representation : 8 groups
- Define these groups into calendar boolean variables
- Define in each group the consumption 'pattern' of the group (simply the mean) $\mathrm{mean}_{G(t)}$
- $$Z_t = [[C]_t \; [M]_t]$$
- Put $[C]_t = [\mathrm{mean}_{G(t)}, Y_{t-7}]$

Table : Groups, 1 . . . 8, are defined using a calendar interpretation of clusters from Monday (day 1) to Sunday (day 7) and from January (month 1) to December (month 12) computed form January 1st to August 31th [?].

	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

.

K-means algorithm

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move.

Important features

1. Number of clusters
2. Statibility of the algorithm

Reduced dictionary : Meteo variables

- Linear summary of the variables PCA 90% of the variance. Each variable separately.
- Non-linear summary : for each variable, (Max, Min, Med, Variance)

Convergence results

2-thresholding-step Procedures

$$Y = X\beta + \epsilon \quad Y\ (n \times 1),\ X\ (n \times p)$$

steps		compute	size
Step 1=preselection	Find b Leaders $b < n << p$	X_b	(n, b)
least squares	on Leaders	$\tilde{\beta} = (X_b^* X_b)^{-1} X_b^* Y$	$(1, b)$
Step 2=denoising	the coefficients	$\hat{\beta}$	$(1, \hat{S})$

GR-LOL - step1

The columns of X are normalized

$$\frac{1}{n}\sum_{i=1}^{N} X_{i(j,t)}^2 = 1, \ \forall \ (j,t).$$

"grouped correlation" search and thresholding:

$$\mathcal{K}_{(j,t)} = |\frac{1}{n}\sum_{i=1}^{N} X_{i(j,t)} Y_i| \qquad \forall \ (j,t), \ 1 \leq j \leq p, \ 1 \leq t \leq T$$

$$\rho_j^2 = \sum_{t=1,\dots,T} \mathcal{K}_{(j,t)}^2, \qquad T = \max |\mathcal{G}_j|$$

- $\lambda(1)$ is tuning parameter
- $$\mathcal{B} = \{\mathrm{j} = 1, \ldots, \mathrm{p},\ \rho_{\mathrm{j}}^2 \geq \lambda(1)^2\} \tag{3}$$
- $\mathcal{G}_{\mathcal{B}} = \cup_{\mathrm{j} \in \mathcal{B}} \mathcal{G}_{\mathrm{j}}$.

GR-LOL - step2

OLS on the block-leaders by considering the new pseudo-linear model

$$\mathrm{Y} = \mathrm{X}_{\mathcal{G}_{\mathcal{B}}} \beta_{\mathcal{G}_{\mathcal{B}}} + \text{error}.$$

$$\hat{\beta}_{\mathcal{G}_{\mathcal{B}}} = \hat{\beta}(\mathcal{B}) \quad (\text{ hence } \quad \hat{\beta}_{\mathcal{G}_{\mathcal{B}}^{c}} = 0)$$

where

$$\hat{\beta}(\mathcal{B}) = [\mathrm{X}^{t}_{\mathcal{G}_{\mathcal{B}}} \mathrm{X}_{\mathcal{G}_{\mathcal{B}}}]^{-1} \mathrm{X}_{\mathcal{G}_{\mathcal{B}}} \mathrm{Y}.$$

GR-LOL - step3 Block-Thresholding

- $\lambda(2)$ is another tuning parameter.
- We apply the second threshold on the estimated coefficients

$$\forall \ell = (\mathrm{j}, \mathrm{t}) \in \{1, \ldots, \mathrm{k}\}, \quad \hat{\beta}^*_\ell = \hat{\beta}_\ell \, \mathbb{I}\{ \, \|\hat{\beta}\|_{\mathcal{G}_\mathrm{j},2} \geq \lambda(2) \, \}$$

- $$\|\hat{\beta}\|^2_{\mathcal{G}_\mathrm{j},2} := \sum_{0 \leq \mathrm{t} \leq \mathrm{T}} \hat{\beta}^2_{(\mathrm{j},\mathrm{t})}.$$

GR-LOL - Tuning thresholds

Choose:

- Threshold λ_1 such that

$$\text{GRLOL } \lambda(1) = \kappa_1 [\sqrt{\frac{T \vee \log p}{n}} \vee \tau^*]$$

$$\text{LOL } \lambda_1 = \kappa_1 [\sqrt{\frac{\log p}{n}} \vee \tau^*]$$

- Threshold λ_2 such that

$$\text{GRLOL } \lambda(2) = \kappa_2 [\sqrt{\frac{T \vee \log p}{n}} \vee \tau^*]$$

$$\text{LOL } \lambda_2 = \kappa_2 [\sqrt{\frac{\log p}{n}} \vee \tau^*]$$

$$T := \max_j |\mathcal{G}_j|$$

Concentration results

Loss Function

$$d(\hat{\beta}^*, \beta)^2 = \sum_{l=1}^{k} (\widehat{\beta}_l - \beta_l)^2$$

Assumptions

- Sparsity:

$$\sum_{j=1}^{p} \|\beta\|_{\mathcal{G}_j,1}^q = \sum_{j=1}^{p} [\sum_{t=1}^{T} |\beta_{(j,t)}|]^q \leq (M)^q.$$

$$(\beta \ \in B_{1,q}(M))$$

- Dimension: $p \leq \exp(c'n), \qquad (c' \text{ constant})$

Concentration results

$$\sup_{B_{1,q}(M)} \mathbb{E}d(\hat{\beta}^*, \beta)^2 \leq D[\sqrt{\frac{T \vee \log p}{n}} \vee \tau^*]^{(2-q)}$$

$$\sup_{B_{1,0}(S)} \mathbb{E}d(\hat{\beta}^*, \beta)^2 \leq DS[\sqrt{\frac{T \vee \log p}{n}} \vee \tau^*]^2$$

for some positive constant D

What is τ^* ?

Coherence.

- Let M be the $k \times k$ Gram matrix :

$$M := \frac{1}{n} X^* X.$$

- and the Coherence

$$\tau_n = \sup_{\ell \neq m} |M_{\ell m}| = \sup_{\ell \neq m} |\frac{1}{n} \sum_{i=1}^{N} X_{i\ell} X_{im}|$$

$$= \sup_{(j,t) \neq (j',t')} |M_{(j,t)(j',t')}| = \sup_{(j,t) \neq (j',t')} |\frac{1}{n} \sum_{i=1}^{N} X_{i(j,t)} X_{i(j',t')}|$$

Splitting the coherence : multitask inspiration

We split the coherence τ_n into γ_{BG} and γ_{BA} where

$$\gamma_{BG} := \sup_{t} \sup_{j \neq j'} \left| M_{(j,t)(j',t)} \right| .$$

between groups-given altitude, sup over altitude

$$\gamma_{BA} := \sup_{j,j'} \sup_{t \neq t'} \left| M_{(j,t)(j',t')} \right| \quad \text{(small)}$$

different altitudes, no matter which groups

τ^*

Let us define :

$$\tau^* = \mathrm{T}\,\gamma_{\mathrm{BA}} + \gamma_{\mathrm{BG}}$$

where $\mathrm{T} = \max_{\mathrm{j}=1,\ldots,\mathrm{p}} \#\{\mathcal{G}_{\mathrm{j}}\}$.

Example

ST coefficients, all equal to γ.

$$\gamma_{\mathrm{BA}} = 0, \gamma_{\mathrm{BG}} = \gamma \geq \sqrt{\frac{\log \mathrm{k}}{\mathrm{n}}}$$

	LOL	GRLOL (opt)	GRLOL (worse)
RATES	$\mathrm{ST}[\gamma^2 + \frac{\log \mathrm{k}}{\mathrm{n}}]$	$\mathrm{S}[\gamma^2 + \frac{\mathrm{T}}{\mathrm{n}} + \frac{\log \mathrm{k}/\mathrm{T}}{\mathrm{n}}]$	$\mathrm{ST}[\gamma^2 + \frac{\mathrm{T}}{\mathrm{n}} + \frac{\log \mathrm{k}/\mathrm{T}}{\mathrm{n}}]$

Boosting the rates : strategies for grouping

$Y = X\beta + \epsilon \quad Y\ (N \times 1),\ X\ (N \times k)$

Question : how to group to obtain better rates, when possible ?

$[\sum_{j=1}^{p}[\sum_{t=1}^{T}\|\beta_{(j,t)}\|]^q]$	$[\sqrt{\frac{T \vee \log p}{n}} \vee \{T\ \gamma_{BA} + \gamma_{BG}\}]$
$\downarrow$	$\downarrow$
GATHERING	WORKING on T $\gamma_{BA} + \gamma_{BG}$

Boosting the rates

- Divide the columns of X into two sets : S_1 : highly correlated, S_2 : weakly correlated.
- Put S_1 as 'group beginners' (each of them has smallest altitude in its group) $\longrightarrow \gamma_{BA} << \gamma_{BG} = \gamma_{max}$
- Realize a 'good cut off S_1 and S_2, ensuring :

$$T\gamma_{BA} \leq \gamma_{BG}, \quad \log p/n \leq \gamma_{BG}^2, \quad T/n \leq \gamma_{BG}^2$$

- Fill the groups with affinity with the delegate in terms of $\mathcal{K}_l = |\frac{1}{n}\sum_{i=1}^{N} X_{il}Y_i|$: indication of $\sum_{j=1}^{p}[\sum_{t=1}^{T}|\beta_{(j,t)}|]^q$ as small as possible