Clustering Electricity Consumers using High Dimensional Regression Mixture Models.

Emilie Devijver[1], Yannig Goude[2,3] and Jean-Michel Poggi[2,4]

[1] KU Leuven
[2] Université d'Orsay
[3] EDF R&D
[4] Université Paris-Descartes

August 30, 2016

Context

Goal: perform the prediction of the electricity

Idea: better performance for disaggregated (at the good level) load [1]

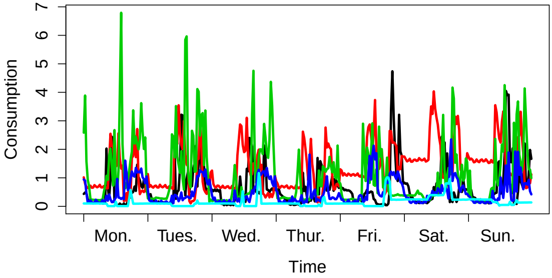
$\rightsquigarrow$ we focus here on the **clustering**

Difficulty: high variability of the individual consumptions

[1] *A model for the effect of aggregation on short term load forecasting*, Sevlian, R.A. and Rajagopal, R., IEEE 2014

Data[2]

- Irish consumption of electricity
- 4225 consumers (residentials or small enterprises)
- Consumption observed every 30 minutes, from January 1st to December 31st 2010
- Access to external information (tariffs, temperature, . . .)



[2] *Electricity smart metering customer behaviour trials findings report*, Commission for energy regulation, Dublin, 2011

1. Method

2. Aggregated consumption

3. Individual consumption

Method [3]

Finite mixture of regression models:

$$s(y|x) = \sum_{k=1}^{K} \pi_k \varphi(\beta_k x, \Sigma_k),$$

Procedure

- Selection of relevant variables (Group-Lasso estimator)
- Refitting by MLE
- Model selection (slope heuristic)
- Clustering (MAP principle)

[3] Model-based clustering for high-dimensional data. Application to functional data, Devijver, ADAC 2016

Method [3]

Finite mixture of regression models:

$$s(y|x) = \sum_{k=1}^{K} \pi_k \varphi(\beta_k^J x, \Sigma_k),$$

Procedure

- Selection of relevant variables (Group-Lasso estimator)
- Refitting by MLE
- Model selection (slope heuristic)
- Clustering (MAP principle)

[3] Model-based clustering for high-dimensional data. Application to functional data, Devijver, ADAC 2016

Method [3]

Finite mixture of regression models:

$$s(y|x) = \sum_{k=1}^{K} \pi_k \varphi(\beta_k^J x, \Sigma_k),$$

Procedure

- Selection of relevant variables (Group-Lasso estimator)
- Refitting by MLE
- Model selection (slope heuristic)
- Clustering (MAP principle)

[3] Model-based clustering for high-dimensional data. Application to functional data, Devijver, ADAC 2016

Method [3]

Finite mixture of regression models:

$$s(y|x) = \sum_{k=1}^{K} \pi_k \varphi(\beta_k^J x, \Sigma_k),$$

Procedure

- Selection of relevant variables (Group-Lasso estimator)
- Refitting by MLE
- Model selection (slope heuristic)
- Clustering (MAP principle)

[3] Model-based clustering for high-dimensional data. Application to functional data, Devijver, ADAC 2016

Method [3]

Finite mixture of regression models:

$$s(y|x) = \sum_{k=1}^{K} \pi_k \varphi(\beta_k^J x, \Sigma_k),$$

Procedure

- Selection of relevant variables (Group-Lasso estimator)
- Refitting by MLE
- Model selection (slope heuristic)
- Clustering (MAP principle)

[3] Model-based clustering for high-dimensional data. Application to functional data, Devijver, ADAC 2016

Wavelets

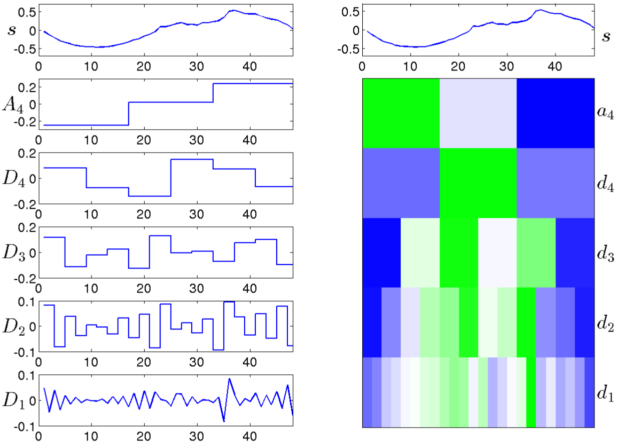
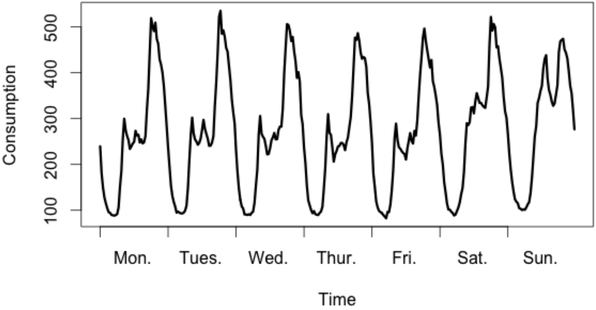


Figure: Decomposition of the signal onto the Haar basis at level 4.

Aggregated dataset

Sample of the considered dataset

- $n = 338$ days
- X: consumption of the day $d - 1$
- Y: consumption of the day d



Our procedure: model with 2 clusters

Aggregated dataset

Model selection: use of the slope heuristic

Penalized likelihood criterion

$$\mathrm{pen}(m) = \kappa D_m$$

with D_m the number of parameters to estimate in the model m.
How to calibrate κ?

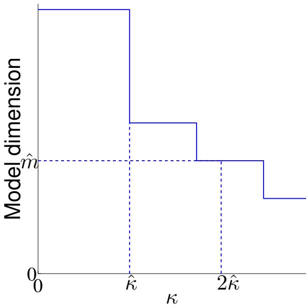
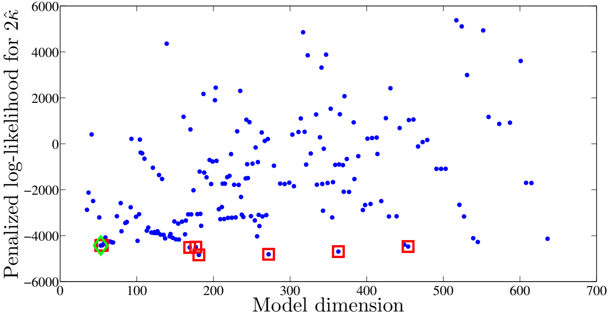


Figure: Use of the dimension jump.

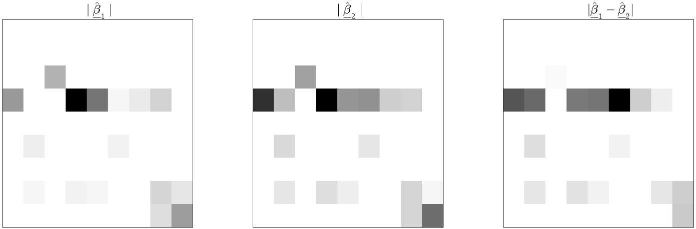
Aggregated dataset

Interesting models



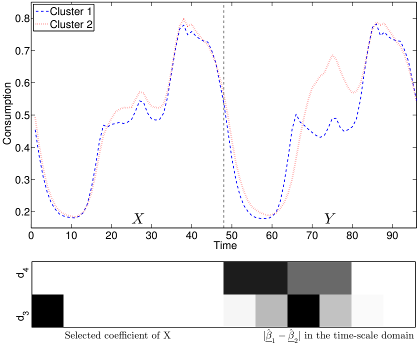
Aggregated dataset

Estimation of the parameters



Aggregated dataset

Interpretation of the clusters



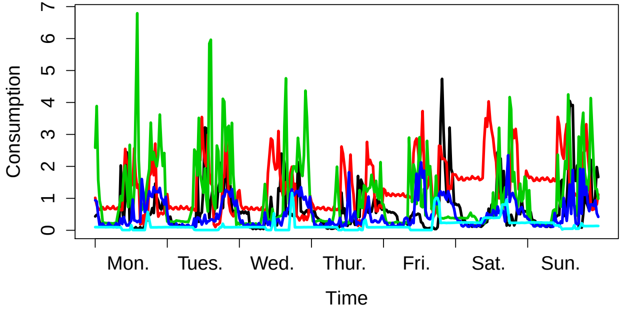
Interpretation	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Sun.
week	0.88	0.96	0.94	0.98	0.96	0	0
weekend	0.12	0.04	0.06	0.02	0.04	1	1

Table: We summarize the proportion of day type in each cluster, and interpret it.

Individual consumption

Data

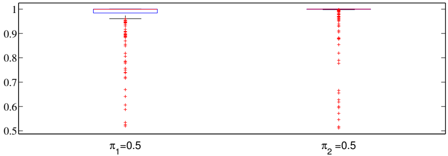
- $n = 487$ consumers
- X: consumption of Tuesday January 5th 2010, projected onto Haar basis
- Y: consumption of Wednesday January 6th 2010, projected onto Haar basis

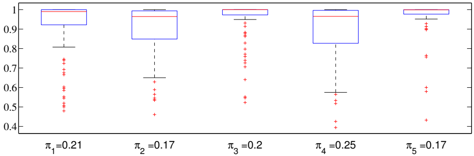


Our method: Model 1 (2 clusters) and Model 2 (5 clusters)

Individual consumption

A posteriori probabilities for each observation





Individual consumption

Interpretation of the clusters along a year

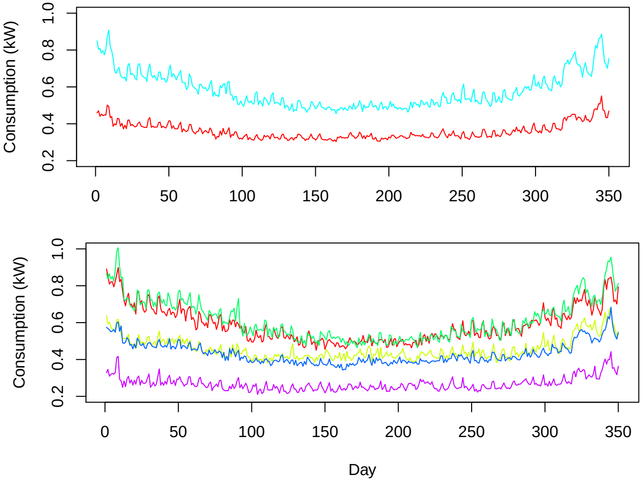


Figure: Daily mean consumptions of the cluster centers along the year for 2 (top) and 5 clusters (bottom).

Individual consumption

Estimation of the parameters

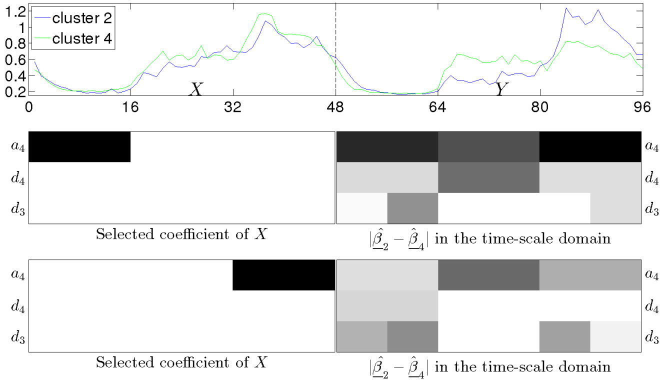


Figure: Clustering representation for the two medium consumer clusters.

Individual consumption

Interpretation of the clusters according to the temperature

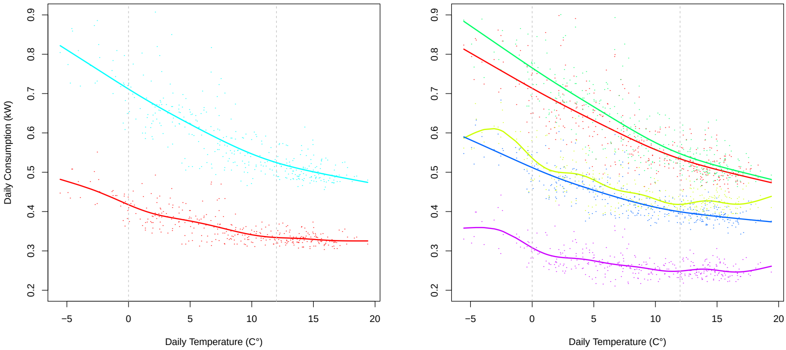


Figure: Daily mean consumptions of the cluster centers in function of the daily mean temperature for 2 (on the left) and 5 clusters (on the right).

Individual consumption

Interpretation of the clusters according to the tariffs

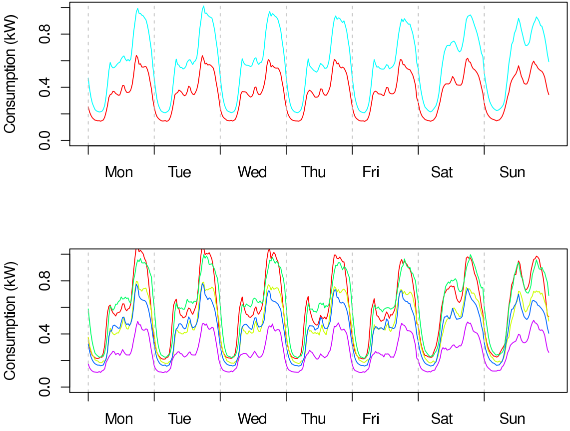


Figure: Average (over time) week of consumption for the centers of each cluster.

Conclusion

- Unsupervised clustering method for regression data in high-dimension
- *Theoretical result proving the model selection step*[4]
- Real data analysis

[4] Devijver, E., *Finite mixture regression: a sparse variable selection by model selection for clustering*, EJS, 2015

Conclusion

- Unsupervised clustering method for regression data in high-dimension
- *Theoretical result proving the model selection step*[4]
- Real data analysis

Thank you for your attention!

- E. Devijver, Y. Goude et J.-M. Poggi, *Clustering electricity consumers using high- dimensional regression mixture models*, 2015, submitted, arXiv:1507.00167
- Matlab code: http://git.auder.net/?p=select.git

[4] Devijver, E., *Finite mixture regression: a sparse variable selection by model selection for clustering*, EJS, 2015