# Classification de consommateurs électriques à l'aide de modèles de mélanges en régression en grande dimension.

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

[1] *Department of Mathematics and Leuven Statistics Research Centre (LStat), KU Leuven, Leuven, Belgium*
*emilie.devijver@wis.kuleuven.be*
[2] *Université d'Orsay, Laboratoire de Mathématiques, bât. 425, 91405 Orsay, France*
[3] *EDF R&D, 1 avenue du général De Gaulle, Clamart, France*
*yannig.goude@edf.fr*
[4] *Université Paris Descartes, Paris, France*
*Jean-Michel.Poggi@math.u-psud.fr*

**Résumé.** De nombreuses informations sur les consommateurs individuels sont désormais disponibles grâce, par exemple, aux nouvelles techniques de *smart grid*. L'exploitation de ces résultats passe par la modélisation à différentes échelles et l'exploitation du profil de charge. La segmentation des consommateurs basée sur la classification de charge de consommation est une approche naturelle dans cette direction. On illustre dans cet exposé l'utilisation d'une méthode basée sur les modèles de régression en grande dimension qui effectue simultanément la classification et la sélection de modèles (pour réduire la dimension) sur des données réelles de consommations électriques. On insistera sur les avantages de la méthode par rapport à ce jeu de données.

**Mots-clés.** Modèles de mélange en régression, consommation électrique

**Abstract.** Massive informations about individual (household, small and medium enterprise) consumption are now provided with new metering technologies and the smart grid. A major exploitation of those data are load profiling and modeling at different scales on the grid. Customer segmentation based on load classification is a natural approach for that and is a prolific way of research. We illustrate in this talk a methodology based on high-dimensional regression models which performs clustering and model selection (our pattern reduction step) at the same time, on real data set of Irish customers. We highlight advantages of the method according to this dataset.

**Keywords.** Model-based regression clustering, electricity consumption

# 1 Introduction

Les nouvelles technologies donnent accès à de nouvelles informations (potentiellement beaucoup) sur la consommation électrique individuelle. Les *smart meters* sont des boîtiers électroniques permettant d'enregistrer la consommation individuelle par demi-heure. ERDF prévoit d'en installer 35 millions d'ici 2020. Un enjeu important pour les statisticiens est de développer des méthodes pour tirer profit de ces données.

Plusieurs applications venant de l'analyse de données individuelles peuvent être trouvées dans la littérature. La segmentation des consommateurs basées sur la classification de la charge est un outil important. Figueiredo et al (2005) comparent des techniques existantes pour caractériser les consommateurs et montrent ainsi l'importance de tels outils pour proposer une offre adéquate aux clients. Le Zhou et al (2013) ont comparé les méthodes les plus populaires, concluant que la classification pour ces données issues de *smart grid* est un problème difficile à cause de la complexité, de la grande dimension, de la masse et de l'hétérogénéité des données. Un autre problème est dû à la structure dynamique des données et particulièrement aux variations de portfolio (perte et gain de consommateurs), la mise à jour de la classification avec ces variations étant un problème difficile.

Une approche classique consiste à construire des classes dans la population telles que chaque classe est différente des autres mais les consommateurs ont un comportement similaire à l'intérieur d'une même classe. Un point crucial consiste à trouver le bon compromis entre des classes suffisamment larges pour capter des informations intéressantes, mais pas trop grandes pour discriminer parmi différentes habitudes. L'effet de cette agrégation est étudié dans Pompey et al (2015) et Sevlian et Rajagopal (2014) par exemple.

La majorité des méthodes consiste en une étape de réduction de dimension et une étape de classification, les deux étapes étant effectuées séparément. Dans cet exposé, on propose une méthodologie basée sur de la classification par modèles de mélanges en régression où ces deux étapes sont réalisées conjointement. L'idée avait été introduite par Misiti et al (2010), où un algorithme de classification avait été introduit pour améliorer la classification de la consommation nationale française. Nous proposerons ici une procédure qui permet de reconnaître les variables sélectionnées, pour expliquer la classification.

Dans la Section 2, on présente brièvement la méthode. Dans la Section 3, on illustre quelques aspects de cette méthode sur la courbe agrégée. Enfin, dans la Section 4, on décrit nos résultats sur les données individuelles.

# 2 Méthode

On va utiliser un modèle de mélange en régression, avec $K$ densités Gaussiennes multivariées. La Figure 1 illustre ce modèle dans le cas univarié. La densité conditionnelle est

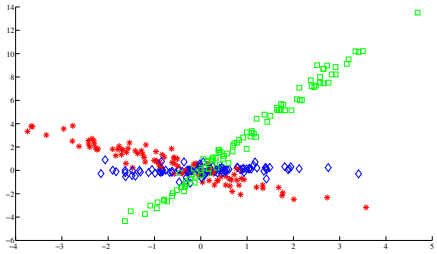

Figure 1: Données simulées avec des régresseurs et une réponse univariés. Les régresseurs $(x_i)_{1\leq i\leq n}$ sont simulés, de loi $\mathcal{N}(0,1)$ et $(y_i)_{1\leq i\leq n}$ sont simulés à partir d'un mélange de $K=3$ Gaussiennes, avec une variance égale à 0.25 dans chaque classe et $\beta=[-1,0.1,3]$.

alors

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

où les $\pi_k$ sont les proportions associés à chaque classe et où le modèle de la $k$ème classe est déterminé par $Y=\beta_k X+\varepsilon_k$ où $\varepsilon_k$ est un bruit gaussien centré de matrice de covariance $\Sigma_k$.

Un tel modèle peut être interprété et utilisé de deux façons différentes. Une première approche consiste à se focaliser sur la classification, étant données les estimations des paramètres, à l'aide du principe du maximum a posteriori. La deuxième approche se concentre plutôt sur le modèle, son interprétation permettant de comprendre la relation entre les régresseurs et la variable réponse à l'aide des paramètres $\beta_k$ et $\Sigma_k$ dans la classe $k$.

Le choix du nombre de classes $K$ et de la sélection de variables dans la matrice $\beta$ est fait par un critère de sélection de modèles. En effet, on construit une collection de modèles avec différents nombres de classes, et où la parcimonie varie, en fonction des variables sélectionnées par l'estimateur du Lasso. L'heuristique de pentes est utilisée pour sélectionner le ou les meilleurs modèles parmi cette collection.

Les données considérées, des courbes de charges électriques, sont des données fonctionnelles. On tire parti de cette structure en les projetant sur une base de Haar et en ne gardant que certains coefficients de détails et les coefficients d'approximation.

Cette méthode est détaillée dans la prépublication Devijver (2014).

# 3 Illustration sur les données agrégées

On considère ici la courbe agrégée de 487 résidentiels sur 338 jours. On classe les jours, en fonction de la relation d'un jour sur son lendemain. Un modèle est sélectionné en utilisant

l'heuristique de pentes : la Figure 2 illustre le saut de dimension pour cette collection de modèles et le choix du modèle minimisant la log-vraisemblance pénalisée.

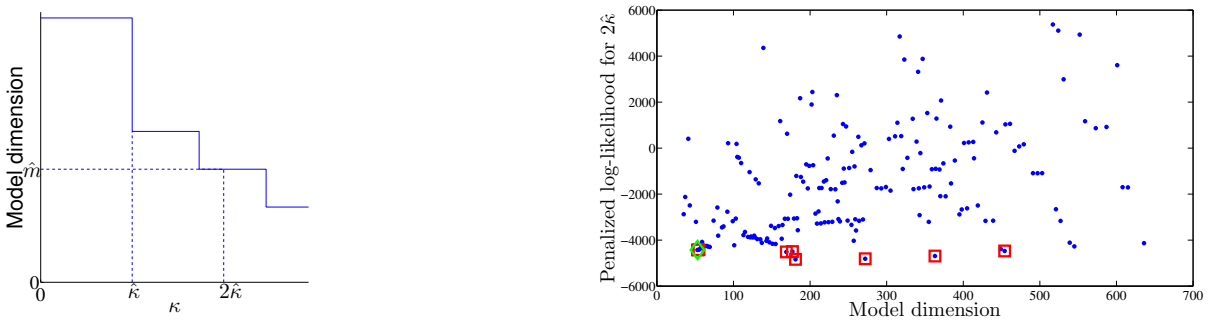


Figure 2: À gauche, on sélectionne le modèle $\hat{m}$ qui minimise le critère pénalisé, où le coefficient de proportionnalité est $2\hat{\kappa}$, avec $\hat{\kappa}$ le plus grand saut. À droite, minimisation de la log-vraisemblance pénalisée. Les modèles intéressants sont entourés de carrés rouges, le modèle sélectionné est entouré d'un losange vert.

On sélectionne un modèle à 2 classes et assez parcimonieux. On représente les estimateurs $\hat{\beta}$ et $\hat{\Sigma}$ dans les Figures 3 et 4, ou plus particulièrement $\underline{\hat{\beta}}$ et $\hat{\Sigma}$, où $\underline{\beta}_k = P_k \beta_k$ et $P_k P_k^t = \Sigma_k^{-1}$ est la racine de Cholesky de $\Sigma_k^{-1}$.

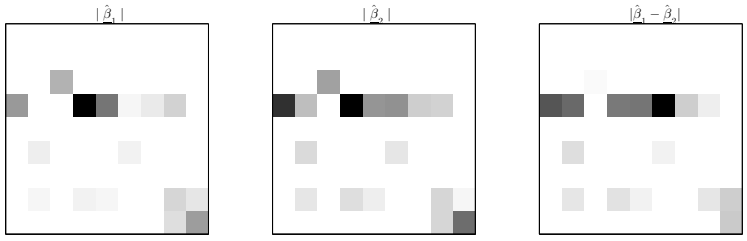


Figure 3: Pour le modèle sélectionné, on représente $\underline{\hat{\beta}}$ dans chaque classe et la différence entre les classes. Les valeurs absolues des coefficients sont représentées sur une échelle de gris, le blanc représentant le 0.

Dans la table 1, on représente les proportions de chaque couple d'appartenir à chaque classe en fonction des jours de la semaine.

# 4 Résultats sur les données individuelles

Dans cette partie, on classe les consommateurs en fonction de la transition entre le 5 et 6 janvier 2010. Avec notre procédure, nous avons construit deux modèles intéressants, notés M1 et M2 dans la suite.


Figure 4: Pour le modèle sélectionné, on représente $\hat{\Sigma}$ dans chaque classe. Les valeurs sont comprises entre 0 et $5 \times 10^{-3}$.

Table 1: On résume les proportions de chaque type de jour dans chaque classe et on interprète ces résultats.

| Interprétation | Lundi | Mardi | Mercredi | Jeudi | Vendredi | Samedi | Dimanche |
|---|---|---|---|---|---|---|---|
| Semaine | 0.88 | 0.96 | 0.94 | 0.98 | 0.96 | 0 | 0 |
| Week-end | 0.12 | 0.04 | 0.06 | 0.02 | 0.04 | 1 | 1 |

Dans la Figure 5, on représente la consommation moyenne des centres de classes sur toute l'année. On voit clairement que les deux classifications séparent les consommateurs qui ont différents niveaux moyens de consommations et différents ratio entre l'hiver et l'été, sûrement dus à un chauffage électrique. Rappelons que le modèle est fait sur les données centrées et que le niveau moyen n'entre pas en compte dans la modélisation.

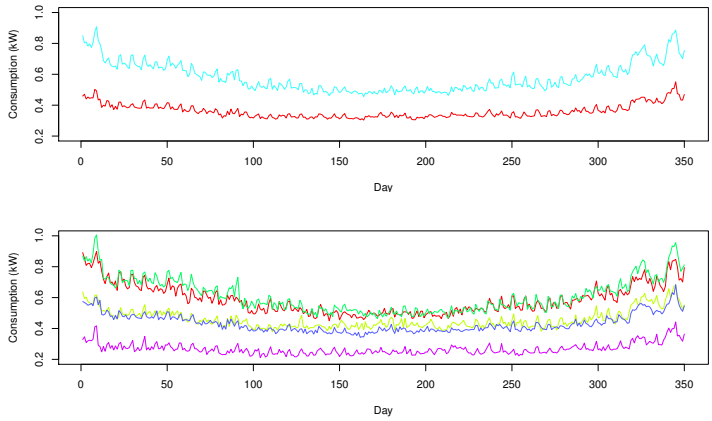


Figure 5: Consommation moyenne quotidienne pour les centres des classes sur l'année, pour le modèle M1 (en haut) et le modèle M2 (en bas).

Une autre observation importante concerne les profils hebdomadaires et quotidiens des centres de classes. Dans la Fig. 6, on représente une semaine moyenne pour les centres de classes pour les deux classifications. Ici, on remarque des différences entre les 5 classes dans le modèle M2, sûrement dues à différents tarifs.

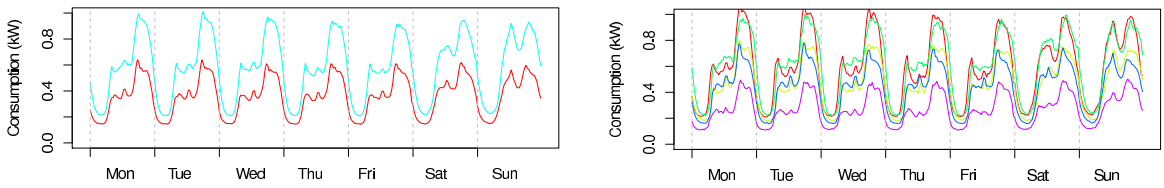


Figure 6: Semaine moyenne pour les centres des classes, pour le modèle M1 (à gauche) et le modèle M2 (à droite).

# 5 Conclusion

Nous proposons une méthode pour classifier les consommateurs électriques sur des données individuelles. À partir d'une analyse métier, on obtient une classification cohérente avec divers critères classiques. Cette analyse fait l'objet d'une prépublication (Devijver, Goude, et Poggi (2015)).

La prochaine étape consiste à utiliser cette classification pour améliorer la prédiction.

# Bibliographie

[1] E. Devijver, *Model-based clustering for high-dimensional data. Application to functional data*, 2014, arXiv:1409.1333.

[2] E. Devijver, Y. Goude et J.-M. Poggi, *Clustering electricity consumers using high-dimensional regression mixture models*, 2015, arXiv:1507.00167.

[3] V. Figueiredo, F. Rodrigues, Z. Vale, et J. Gouveia, *An electric energy consumer characterization framework based on data mining techniques*, IEEE Transactions on Power Systems, vol. 20, no. 2, pp. 596–602, 2005.

[4] K. le Zhou, S. lin Yang, et C. Shen, *A review of electric load classification in smart grid environment*, Renewable and Sustainable Energy Reviews, vol. 24, pp. 103 – 110, 2013.

[5] M. Misiti, Y. Misiti, G. Oppenheim, et J.-M. Poggi, *Optimized clusters for disaggregated electricity load forecasting*, REVSTAT, vol. 8, no. 2, pp. 105–124, 2010.

[6] P. Pompey, A. Bondu, Y. Goude, et M. Sinn, *Massive-scale simulation of electrical load in smart grids using generalized additive models*, Modeling and Stochastic Learning for Forecasting in High Dimensions, ser. Lecture Notes in Statistics, A. Antoniadis, J.-M. Poggi, and X. Brossat, Eds., Springer International Publishing, 2015, vol. 217, pp. 193–212.

[7] R. Sevlian et R. Rajagopal, *A model for the effect of aggregation on short term load forecasting*, in PES General Meeting — Conference Exposition, 2014 IEEE, 2014, pp. 1–5.