Sélection automatique de composantes dans les modèles additifs par des méthodes de régression pénalisée

Vincent Thouvenot [1] & Anestis Antoniadis [2]& Xavier Brossat [1,3]& Yannig Goude [1,4]& Jean-Michel Poggi [5]

[1] *EDF, 1 avenue Général de Gaulle, 92140 Clamart, France, e-mail: vincent.thouvenot@edf.fr*
[2] *Université Joseph Fourier Tour IRMA, B.P.53, 38041 Grenoble cedex 9, France, e-mail: Anestis.Antoniadis@imag.fr*
[3] *e-mail: xavier.brossat@edf.fr*
[4] *e-mail: yannig.goude@edf.fr*
[5] *Université Paris-Sud, Bâtiment 425, 91405 Orsay cedex, France, e-mail: Jean-Michel.Poggi@math.u-psud.fr*

Résumé. L'électricité ne se stockant pas aisément, EDF a besoin d'outils de prévision de consommation et de production efficaces. De plus, grâce au développement de nouvelles technologies, EDF peut étudier les mailles locales du réseau, conduisant à un nombre important de séries chronologiques à étudier, conduisant à un besoin de développer des nouvelles méthodes de sélection et d'estimation efficaces et automatiques des modèles de prévision. Nous utilisons des modèles additifs, qui sont flexibles et ne souffrent pas du fléau de la dimension. L'objectif de ce travail est de présenter deux procédures automatiques de sélection et d'estimation de composantes d'un modèle additif. Nous utilisons du Group LASSO, qui est, sous certaines conditions, consistant en sélection, et des P-Splines, qui sont consistantes en estimation. Nous présentons deux manières de combiner ces deux méthodes de régression pénalisée. Les procédures sont illustrées sur des simulations et une application pratique.

Mots-clés. Group LASSO, Modèle additif, Prévision de charge électrique, P-Splines, Régression pénalisée

Abstract. As electricity can't easily be stored, EDF, French company of production and distribution of electricity, needs efficient tools of load forecasting. Moreover, thanks to development of new technologies, EDF has data on the local network. That means there are a lot of series to study, so we need new automatic methods for selection and estimation of forecasting models. The aim of this work is to propose two automatic methods of selection and estimation in additive model, combining Group LASSO, which is selection consistent and P-Splines, which is estimation consistent. Numerical experiments on simulations and a real application are provided.

Keywords. Additive Model, Group LASSO, Load Forecasting, Penalized Regression, P-Splines

1 Introduction

Comme l'électricité ne se stocke pas, ou très difficilement, EDF doit toujours équilibrer sa production d'électricité avec la consommation. Ceci amène à un besoin de modélisation et de prévision de consommation d'électricité à différents horizons. Plusieurs évolutions majeures s'ajoutent à cette problématique d'équilibre, notamment grâce au développement de nouvelles technologies (Smart Grids), qui permettent de s'intéresser à la prévision de la maille locale. L'agrégation des prévisions locales peuvent permettre d'améliorer les prévisions globales. De plus, cela permet de répondre à des problématiques d'optimisation du réseau.

Pour répondre à la question de prévision de consommation à court et moyen terme, nous avons choisi d'utiliser des modèles additifs. Ils ont été introduits par Hastie et Tibshirani (1990):

$$E\left(Y|(X_1,\ldots,X_d)=(x_1,\ldots,x_d)\right)=\beta_0+\sum_{i=1}^{d}f_i(x_i), \tag{1}$$

où Y est la variable à expliquer (par exemple la charge d'un poste source), $(X_1,\ldots,X_d)$ les variables explicatives (par exemple la température, le moment dans l'année,...), f_i la composante de la *ieme* variable explicative, β_0 une constante. Nous avons choisi ce modèle, parce qu'il est suffisament flexible pour la problématique de prévision de consommation d'électricité, et contrairement à des modèles complétement non paramétriques, ne souffre pas du fléau de la dimension.

L'étude de la maille locale du réseau électrique français amène à un nombre important de séries chronologiques à étudier (typiquement de l'ordre de milliers à quelques dizaines de milliers). Nous avons donc besoin d'automatiser la sélection et l'estimation des composantes des modèles. Il peut y avoir quatre objectifs pour réaliser de la sélection de composantes. Nous pouvons chercher à :

- améliorer la prédiction grâce à la sélection de composantes
- exclure les composantes qui influencent peu la variable à expliquer
- conserver uniquement les composantes qui influencent le plus la variable à expliquer
- sélectionner le vrai modèle si celui-ci est présent dans les modèles possibles

Notre principal objectif est prédictif.

Nous utilisons des bases de B-Splines pour approximer localement les composantes f_i. Cette projection permet de nous ramener au cas du modèle linéaire, mais possiblement en grande dimension. En effet, le modèle 1 peut se réécrire:

$$E\left(Y|(X_1,\ldots,X_d)=(x_1,\ldots,x_d)\right)=\beta_0+\sum_{j=1}^{d}\sum_{k=1}^{m_j}\beta_{j,k}B_{j,k}^{q_j}(x_j), \tag{2}$$

avec $\mathbf{B_j}(-) = \left(B^{q_j}_{j,k}(-)|k = 1, \ldots, K_j + q_j = m_j\right)$ la base de B-Splines de degré q_j et K_j noeuds, où est projetée la variable X_j. Soit $\mathbf{B_j} = (\mathbf{B_j}(x_{1j})^T, \ldots, \mathbf{B_j}(x_{nj})^T)^T \in \mathbb{R}^{n\times m_j}$. Les paramètres à estimer du modèle sont alors β_0 et $\beta = (\beta_{1,1}, \ldots, \beta_{d,m_d})^T$.

2 Régression pénalisée

2.1 MCO

Nous nous plaçons dans le cas de l'estimation du modèle 2. L'estimateur des moindres carrés ordinaires (MCO) minimise la fonction objective suivante:

$$Q^{MCO}(\beta) = \sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{d}\sum_{k=1}^{m_j}\beta_{j,k}B^{q_j}_{j,k}(X_{i,j})\right)^2,$$

Nous appelons alors estimateurs des moindres carrés ordinaires (EMCO) les estimateurs $\hat{\beta}$ de β les estimateurs tels que $\hat{\beta} = \arg\min_\beta Q(\beta)^{MCO}$.

La méthode par MCO ne permet pas de sélectionner les composantes, et de plus, nécessite que $\sum_{j=1}^{d} m_j \leq n-1$.

2.2 Group LASSO

Le Group LASSO a été introduit par Yuan et Lin (2006). Appliquée à l'estimation du modèle 2, la fonction objective est donnée par :

$$Q^{GroupLASSO}(\beta) = Q^{MCO}(\beta) + \lambda\sum_{j=1}^{d}\sqrt{m_j}||\beta_{\mathbf{j}}||_2, \tag{3}$$

avec $\beta_{\mathbf{j}} = (\beta_{j1}, \ldots, \beta_{jm_j})^T$ et m_j l'ordre de la base de B-Splines où est projetée la jeme variable. Il est à noter que la constante n'est pas pénalisée. Le Group LASSO permet de sélectionner les composantes.

L'expression 3 ne tient pas compte des échelles. Dans sa forme actuelle, toutes les variables sont pénalisées de la même manière, sans tenir compte des échelles. Chaque groupe de variables peut être orthonormalisé pour supprimer les unités. Cependant, cela change le problème initial et nous allons par la suite utiliser une version standardisée du Group LASSO.

2.3 P-Splines

Nous utilisons des P-Splines, introduits par Eilers et Marx (1996). La fonction objective à optimiser est alors dans ce cas donnée par :

$$Q^{pspline}(\beta) = Q^{MCO}(\beta) + \sum_{j=1}^{d} \lambda_j ||D_{k,j}\beta_{\mathbf{j}}||_2^2, \tag{4}$$

où $D_{k,j}$ représentation matricielle de la différence d'ordre k.

Lorsque les noeuds sont équirépartis, cela revient à minimiser l'équation suivante:

$$Q^{pspline}(\beta) = Q^{MCO}(\beta) + \lambda \int_0^1 \left\{ \left(\sum_{k=1}^{K+q} \beta_k B_k^q(x_i) \right)^{(p)} \right\}^2 dx,$$

La section suivante décrit les manières dont nous combinons les critères présentés.

3 Procédures en deux ou trois étapes pour sélectionner, régulariser et estimer les modèles

Nous comparons deux familles de procédures : celles de type Ante et celles de type Post.

Procédure de type Ante

1. Utilisation du Group Lasso pour sélectionner les variables. Choix du terme de lissage avec un critère BIC, AIC ou GCV (ce choix n'est pas remis en cause dans la deuxième phase)

2. Etape de régularisation et d'estimation: utilisation des MCO ou des P-Splines (respectivement avec et sans régularisation) sur les composantes conservées à l'étape précédente

Procédure de type Post

1. Utilisation du Group Lasso pour sélectionner les variables.

2. Pour l'ensemble des sous-modèles sélectionnés sur la grille de paramètres de lissage par le Group LASSO, estimer les coefficients à l'aide des MCO ou des P-Spline (sans et avec régularisation)

3. Recalculer le BIC, AIC ou le GCV avec ces nouveaux estimateurs et conserver celui qui minimise le critère choisi

Pour la première famille de procédures, la phase de sélection de variables est faite au préalable, avant l'estimation du modèle final (procédures qui peuvent être caractérisées de type "Ante" pour le choix du terme de lissage du Group LASSO). Il n'y a pas de retour sur la sélection. Les procédures de la seconde famille vont sélectionner le sous-modèle qui minimise un critère type BIC, AIC ou GCV après une phase d'estimation avec ou sans régularisation, qui permet de débiaiser le group LASSO (procédures qui peuvent caractériser de type "Post" pour le choix du terme de lissage du Group LASSO). Ce faisant, la seconde famille de procédures demande l'estimation de beaucoup plus de modèles que les deux premières, le coût informatique risque donc d'être beaucoup plus élevé.

4 Expériences numériques

4.1 Simulation

Le plan de simulation que nous utilisons est inspiré de l'exemple d'Antoniadis et *al*. (2012).

Nous simulons le modèle suivant:

$$Y_i = f_1(X_{i1}) + f_2(X_{i2}) + f_3(X_{i3}) + f_4(X_{i4}) + \sum_{j=5}^{10} f_5(X_{ij}) + \epsilon_i,$$

Dans ce modèle, seules les 4 premières composantes $f_1, \dots, f_4$ sont non nulles. Nous dirons que les quatre premières variables sont influentes ou informatives. La composante f_5 est nulle, nous dirons que $X_5, \dots, X_{10}$ sont non informatives. Le design est construit de manière à ce que la corrélation entre $(X_1, \dots, X_{10})$ puisse varier. Les bruits ϵ sont non corrélés, gaussiens, centrés et de variance σ^2.

Nous illustrons les procédures en faisant varier le nombre d'observations, le bruit et la corrélation entre les covariables. Nous sommes intéressés par les capacités de sélection et de prédiction des procédures. L'objectif des simulations est de vérifier et d'illustrer les capacités suivantes des procédures :

- L'intensité du bruit influence-t-elle la qualité de sélection et d'estimation des modèles? Si oui, comment?
- Comment le nombre d'observations influence-t-il la qualité de sélection et d'estimation des modèles?
- Quel est l'effet de la corrélation entre les variables explicatives?

4.2 Données réelles

4.2.1 Données nationales

Nous nous intéressons dans un premier temps à la consommation électrique française entre 2007 et 2012. Il s'agit d'un jeu de données relativement bien connu et avec peu de variabilité. Nous avons à notre disposition des variables métérologiques (température brute nationale, températures brutes de 32 postes métérologies, forces de vent nationale ou locale, nébulosité,...), des variables calendaires (instant dans la journée, dans l'année, jours fériés,...) et des variables tarifaires. A partir des variables existantes, nous en construisons des nouvelles (variables lissées, retardées, ...) et appliquons nos procédures. Nous travaillons à la fois sur du court-terme (jour pour le lendemain) et sur du moyen terme (une année par la suivante).

4.2.2 Données locales

Nous travaillons aussi sur des données locales de consommation d'électricité, et notamment sur les postes sources (environ 2200, voir l'article de Goude et *al.*, 2014). Nous nous concentrons sur deux zones géographiques au profil différent: les zones associées à Blois (plus de 50 postes sources) et à Lyon (plus de 80 postes sources). Les courbes locales de consommation d'électricité sont beaucoup plus variables que les courbes nationales: un départ d'un gros industriel peut entraîner une rupture et un changement de forme de la courbe de la charge sur un poste source donné par exemple. Nous avons à notre disposition plusieurs postes météorologiques dans chaque zone géographique, ainsi que des données calendaires et tarifaires. A partir des températures brutes de chaque poste météorologique, nous pouvons avoir des températures lissées, retardées,... Nous pouvons aussi utiliser la température nationale, régionale, d'un poste météorologique local,... Nous souhaitons tester nos procédures en adoptant l'idée selon laquelle qu'il peut être intéressant pour avoir la meilleure prévision possible de ne conserver qu'une partie des variables de température disponibles.

Bibliographie

[1] Antoniadis, Gijbels et Verhasselt (2012),Variable Selection in additive models using P-Splines, *Technometrics*, 54 (4), 425–438.
[2] Bach (2008), Consistency of the Group Lasso and Multiple Kernel Learning, *Journal of Machine Learning Research* , 9, 1179–1225.
[3] Eilers et Marx (1996), Flexible smoothing with B-splines and penalties, *Journal of Machine Learning Research* , 11 (2) , 89–121.
[4] Goude, Nedellec et Kong (2013), Local short and middle term electricity load forecasting with semi-parametric additive models, *IEEE Transactions on Smart Grid* , 5 (1), 440–446.
[5] Hastie et Tibshirani (1990) *Generalized Additive Models*, New York: Chapman and Hall.
[6] Yuan et Lin (2006), Model Selection and Estimation in Regression with Grouped Variables, *Journal of the Royal Statistical Society, Series B*, 68, 49–67.