Marie Devaine

Prévision de consommation électrique par mélange de prédicteurs intermittents

sous la direction de Yannig Goude (EDF R&D)
et Gilles Stoltz (CNRS - ENS Paris - HEC Paris).

Table des matières

Introduction

L'objet de ce mémoire est de présenter une application de la théorie des suites individuelles et d'agrégation de prédicteurs à des données de consommation d'électricité d'EDF. La prévision de consommation électrique à court terme constitue un enjeu central pour une entreprise comme EDF. Des modèles performants ont été développés dans le but d'en être de bons prédicteurs, mais l'estimation de leurs paramètres reste un problème délicat. De plus, le développement de nouveaux modes de consommation et l'ouverture récente du marché nécessite l'utilisation de prédicteurs plus adaptatifs, notamment grâce à des modèles semi-paramétriques voire non paramétriques. Dans ces conditions, l'agrégation de prédicteurs par apprentissage séquentiel permet de tirer profit de la diversité des modèles et de gagner en robustesse. On trouve dans la littérature de nombreux algorithmes de mélange de prédicteurs, la question est de savoir lesquels seront le plus adaptés à ce type de données.

D'autre part, il peut être intéressant de considérer des prédicteurs spécialisés temporellement (week-end, été..). Le cadre des experts intermittents (*sleeping experts*) se révèle alors particulièrement adapté. Dans ce cadre cependant, il existe peu de résultats théoriques sur les performances des algorithmes. Un des enjeux est donc d'adapter les algorithmes qui semblent intéressants au cadre des experts intermittents.

1 Présentation des données

Particularités des données

J'ai été confrontée à deux jeux de données, un jeu de données de consommation slovaque et un jeu de données de consommation française.

- Le jeu de données slovaques comprend la consommation électrique slovaque sur 3 ans (du 01/01/05 au 31/12/07) à chaque heure en Megawatt et les prédictions de 35 experts (prédicteurs obtenus à partir d'un certain modèle et d'un certain paramétrage, par exemple) sur cette période.
  La particularité de ces données est le caractère intermittent des experts (*sleeping experts*). En effet, certaines prévisions de certains experts sont manquantes : tous les experts ne s'expriment pas à tous les tours. On a bien cependant qu'à tous les tours au moins un des experts s'exprime.

On ne dispose que de peu d'informations sur la manière dont sont construits ces experts et sur les raisons pour lesquelles ils sont inactifs à certains tours (pour certains cela peut être dû à une spécialisation). Comme le jeu de données est très grand, on l'a divisé en 24 sous-jeux de données à heure fixe, et on applique nos différents algorithmes de mélange sur les sous-jeux.

– Le jeu de données françaises comprend la consommation France au pas demi-horaire sur un an (du 01/09/07 au 31/08/08) en Gigawatt et la prédiction sur cette période de 24 experts, décrits dans le paragraphe suivant.
  Dans ce jeu de données aussi, même s'il est moins prononcé que dans le jeu de données slovaques, on retrouve le caractère intermittent (*sleeping*) des experts. De plus comme ces données prennent en compte une période plus courte que le précédent nous n'avons pas voulu le séparer en 48 sous-jeux de données.
  Pour rester cohérent avec la contrainte opérationnelle, les poids du mélange d'experts sont calculés à heure fixe (12h) et restent fixes pour toute la journée (mises à part des renormalisations sur les experts actifs dues à des experts qui se désactiveraient ou s'activeraient au cours de la journée). La prédiction est donc formée à horizon variable : la prédiction de 12h pour 12h30 est faite à l'horizon d'une demi-heure, tandis que la prédiction faite à 12h pour 12h le lendemain est faite à l'horizon d'un jour.

La Figure 1 met en relation le pourcentage d'activité des experts et leurs performances.

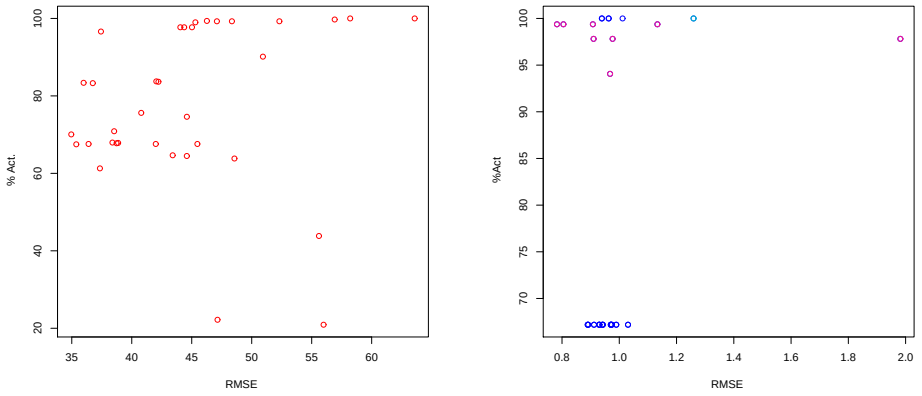

FIGURE 1 – Pourcentage d'activité vs RMSE pour les données slovaques (gauche) et françaises (droite)

Les experts français

Les experts français utilisés sont de trois types : paramétriques (dérivés du modèle de prévision de consommation Eventail), semi-paramétriques (dérivés du modèle GAM) et non paramétrique (un seul expert suivant un modèle de similarité fonctionnelle). Pour les deux premiers types, les prédicteurs de base (Eventail en paramétrique et GAM en semi-paramétrique) ont été déclinés selon plusieurs valeurs des paramètres. En effet, les paramètres issus d'une estimation basée sur les quatre années précédentes ne sont pas nécessairement ceux qui donnent les meilleures performances a posteriori. Pour Eventail

on a fait varier le gradient d'hiver et les poids de recalage court terme, pour GAM on a fait varier les tendances, et on a testé différents modes de recalage. Les prédicteurs issus de GAM et d'Eventail, donnent une prédiction à horizon constant d'une journée. Le prédicteur non paramétrique calcule ses prédictions à 23h30 pour toute la journée du lendemain.

2 Théorie des suites individuelles dans le cadre *sleeping*

2.1 Cadre théorique

Suites individuelles et notations

Dans le cadre des suites individuelles, un statisticien doit prédire une suite $y_1, y_2, \ldots$ d'observations dans $[0, B]$. Il forme ses prédictions $\widehat{y}_1, \widehat{y}_2, \ldots$ de manière séquentielle c'est-à-dire qu'à chaque échéance t, y_t est prédite sur le seul fondement du passé et en ne recourrant à aucun modèle stochastique. La prédiction $\widehat{y}_t$ est formée avant que la vraie valeur y_t ne soit révélée, et les deux sont ensuite comparées. Pour que cette tâche ait un sens dans un cadre aussi général on introduit des experts : des prédicteurs de référence. A chaque tour le statisticien formera donc sa prédiction en s'appuyant sur les prédictions des experts. Plus précisément : pour chaque tour t dans $\{1, \ldots, N\}$ on dispose de N experts. Le sous-ensemble des experts actifs, c'est-à-dire donnant effectivement une prédiction, au temps t est donné par $E_t \subset \{1, \ldots, N\}$. Ainsi, la prédiction de l'expert i au temps t existe si et seulement si $i \in E_t$, elle est alors notée $f_{i,t}$.

La consommation d'électricité (en Megawatt ou Gigawatt) au temps t est notée y_t. Par commodité, on pose $f_{i,t} = 0$ si $i \notin E_t$. On peut alors définir le vecteur de prédiction des experts au temps t par $\mathbf{f}_t = (f_{1,t}, \ldots, f_{N,t})$.

Les méthodes d'agrégation proposées par la suite consistent en la donnée d'une séquence de vecteurs de pondération $\mathbf{u}_1^n = (\mathbf{u}_1, \ldots, \mathbf{u}_n) \in \mathbb{R}^N \times \ldots \times \mathbb{R}^N$ qui permettent au statisticien de former la prédiction linéaire $\widehat{y}_t = \mathbf{u}_t \cdot \mathbf{f}_t$ au temps t. Suivant les méthodes, les vecteurs de pondérations peuvent prendre leurs valeurs dans tout $\mathbb{R}^N$ ou bien on les contraint à rester dans le simplexe correspondant aux indices d'experts actifs que l'on note

$$\mathcal{X}_{E_t} = \left\{ \mathbf{q} \in \mathbb{R}_+^N,\ q_i = 0 \text{ pour } i \notin E_t,\ \text{et} \sum_{i \in E_t} q_i = 1 \right\},$$

(on notera simplement $\mathcal{X}$ pour $\mathcal{X}_{\{1,\ldots N\}}$). Dans le cas contraint au simplexe, la séquence de prédiction est notée $\mathbf{p}_1^n = (\mathbf{p}_1, \ldots, \mathbf{p}_n) \in \mathcal{X}_{E_1} \times \ldots \times \mathcal{X}_{E_n}$.

Performance

On se munit de la perte quadratique $\ell : (x, y) \mapsto \ell(x, y) = (x - y)^2$. Dans la suite on aura besoin d'un majorant de cette perte, on introduit à cet effet B, un réel positif, tel que les prédictions et les consommations d'électricité sont dans $[0, B]$. Un majorant de la perte est alors B^2. On note $\ell_t(\mathbf{v}) = \left(\sum_{i \in E_t} v_i f_{i,t} - y_t \right)^2$ la perte instantanée au temps t associée au vecteur de pondération $\mathbf{v}$ ($\mathbf{v} \in \mathbb{R}^N$ ou $\mathbf{v} \in \mathcal{X}_{E_t}$). Dans le cas particulier où $\mathbf{v}$ est la masse de Dirach en $i \in E_t$, δ_i, on note $\ell_{i,t} = \ell_t(\delta_i) = (f_{i,t} - y_t)^2$.

A ces quantités correspondent les pertes cumulées $L_n(\mathbf{v}) = \sum_{t=1}^n \ell_t(\mathbf{v})$ et $L_{i,n} = \sum_{t=1}^n \ell_{i,t}$.

A une méthode d'agrégation $\mathbf{u}_1^n = (\mathbf{u}_1, \ldots, \mathbf{u}_n) \in \mathbb{R}^N \times \ldots \times \mathbb{R}^N$ (indifféremment contrainte au simplexe ou non), on associe la perte instantanée $\widehat{\ell}_t = \ell_t(\mathbf{u}_t)$ et la perte cumulée $\widehat{L}_n = \sum_{t=1}^n \widehat{\ell}_t$.

Un critère de performance naturel pour une séquence de vecteurs de pondération $\mathbf{v}_1^n = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \mathbb{R}^N \times \ldots \times \mathbb{R}^N$ découlant de ce choix de fonction de perte est l'erreur quadratique moyenne ou RMSE.

Définition. L'erreur quadratique moyenne d'une séquence de vecteurs de pondération $\mathbf{v}_1^n$ est définie par

$$\mathrm{RMSE}(\mathbf{v}_1^n) = \sqrt{\frac{1}{n}\sum_{t=1}^n \left(\sum_{i \in E_t} v_{i,t} f_{i,t} - y_t\right)^2}.$$

Remarque. Grâce à notre convention on peut réécrire plus simplement :

$$\mathrm{RMSE}(\mathbf{v}_1^n) = \sqrt{\frac{1}{n}\sum_{t=1}^n (\mathbf{v}_t \cdot \mathbf{f}_t - y_t)^2}.$$

2.2 Comparaison experts/mélange

Comparaison avec un expert fixe

Comme on se base sur les experts pour former les prédictions, il est naturel de comparer les performances obtenues par un algorithme de mélange à celles obtenues par les experts de base.

Dans le cadre classique des suites individuelles (i.e. quand tous les experts sont actifs à tous les tours), il suffit de comparer les pertes cumulées du mélange et des experts, et d'introduire la notion de regret individuel pour un expert donné, différence entre la perte cumulée du mélange et celle de cet expert. Dans le cadre intermittent, on ne peut cumuler les pertes que pour les tours où l'expert de référence est actif. Le regret individuel associé à l'expert d'indice i est donc :

$$R'_{i,n} = \sum_{t=1}^n \mathbb{I}_{\{i \in E_t\}}(\widehat{\ell}_t - \ell_{i,t}) \ .$$

Comparaison avec une combinaison convexe d'experts

Un but plus ambitieux est de se comparer avec une combinaison convexe fixe d'experts. Dans le cadre intermittent, il faut opérer à des renormalisations lorsque certains experts sont inactifs. Pour toute distribution de probabilité $\mathbf{q} \in \mathcal{X} = \mathcal{X}_{\{1,\ldots,N\}}$, on définit donc le regret convexe renormalisé par :

$$R'_n(\mathbf{q}) = \sum_{t=1}^n \mathbf{q}(E_t)\left(\widehat{\ell}_t - \ell_t\left(\mathbf{q}^{E_t}\right)\right) \ ,$$

où $\mathbf{q}(E_t)$ est le facteur de renormalisation $\sum_{i \in E_t} q_i$,

et $\mathbf{q}^{E_t}$ la restriction renormalisée de $\mathbf{q}$ à $\mathcal{X}_{E_t}$ définie par

$$q_i^{E_t} = q_i \mathbb{I}_{\{i \in E_t\}} / \mathbf{q}(E_t)$$

pour $i \in \{1, \dots, N\}$ avec la convention que si $\mathbf{q}(E_t) = 0$ alors $\mathbf{q}^{E_t} = (0, \dots, 0)$.

On remarque que le regret individuel se déduit de ce regret convexe renormalisé et que dans le cadre classique on retrouve bien une simple différence de pertes cumulées.

Comparaison avec une comparaison linéaire d'experts

Une extension naturelle plus ambitieuse dans le cadre classique est de ne plus se comparer avec une combinaison convexe mais linéaire (en prenant la différence entre perte cumulée du mélange et celle de la combinaison linéaire fixe). Malheureusement nous n'avons pas trouvé d'adaptation satisfaisante au cadre intermittent. Il y a pour cela plusieurs possibilités mais pas vraiment de bonne solution :

- On peut considérer le regret linéaire classique : $R_n(\mathbf{v}) = \sum_{t=1}^{n} \left(\widehat{\ell}_t - \ell_t(\mathbf{v}) \right)$, ce qui a un sens, assez artificiel, grâce à notre convention. Cette définition ne donne pas un but très ambitieux à cause du biais apporté par le cadre intermittent, et n'utilise pas l'information apportée par les sous-ensembles d'indices inactifs à chaque tour. De plus on n'en déduit pas par exemple le regret renormalisé
- On peut essayer de construire un regret partitionné : on introduit une partition des tours selon les ensembles d'experts en activité. Si K désigne le nombre d'ensembles différents parmi les E_t pour t parcourant $1, \dots, n$, on note $U_1, \dots, U_K$ les K classes de E_t différents. Pour tout t on note k_t l'indice tel qu'on ait $E_t = U_{k_t}$. On peut ainsi se ramener au cadre classique en sommant les regrets classiques de chaque sous-ensemble : on se compare alors à la meilleure séquence fixe de K combinaisons linéaires. Plus précisément pour $\mathbf{v}_1^K = (\mathbf{v}_1, \dots, \mathbf{v}_K) \in \left(\mathbb{R}^N\right)^K$, on définit le regret linéaire partitionné
$$R_n(\mathbf{v}_1^K) = \sum_{t=1}^{n} \ell_t(\mathbf{v}_{k_t}) \ .$$
Le but qu'on se fixe est alors bien plus ambitieux. Cette idée peut-être déclinée pour une définition de regret convexe partitionné ou encore de regret individuel partitionné. De telles définitions ne sont cependant pas très pertinentes car les adaptations des algorithmes classiques qu'elles impliquent amènent à une perte d'information.
- On peut reprendre l'idée du regret "renormalisé" et l'étendre aux combinaisons convexes (on étend également les notations).
$$R'_n(\mathbf{v}) = \sum_{t=1}^{n} \mathbf{v}(E_t) \left(\widehat{\ell}_t - \ell_t\left(\mathbf{v}^{E_t}\right) \right) \ .$$
Cependant on ne se compare plus vraiment à la meilleure combinaison linéaire, et le but n'est donc pas beaucoup plus ambitieux que dans le cas du regret convexe renormalisé.

Comparaison avec une séquence d'experts

Au lieu de se comparer à des combinaisons d'experts fixes dans le temps (à des renormalisation dûes à des experts inactifs près) on se compare à des séquences d'experts, en s'autorisant à changer d'expert de référence un nombre fixé de fois. Cela s'adapte bien à un contexte où le meilleur expert change en fonction du temps. On considère donc des

séquences de prédictions d'experts $(f_{i_1,1}, f_{i_2,2}, \dots, f_{i_n,n})$ de taille donnée où la taille d'une telle séquence est définie par :

$$\mathrm{size}(f_{i_1,1}, f_{i_2,2}, \dots, f_{i_n,n}) = |\{t \in \{1, \dots, n-1\}, i_t \neq i_{t+1}\}| \ .$$

Pour simplifier, on se contente de désigner par $(i_1, \dots, i_n)$, la séquence des prédictions d'experts $(f_{i_1,1}, f_{i_2,2}, \dots, f_{i_n,n})$. Dans le cadre intermittent, il faut de plus imposer que la séquence soit admissible ce qui correspond à la condition naturelle :

$$\forall 1 \leqslant k \leqslant n, \ i_k \in E_k \ ,$$

c'est à dire que tous les experts composant la séquence sont actifs au tour considéré. Le regret *tracking* d'une séquence admissible d'experts $(i_1, \dots, i_n) \in \{1, \dots, N\}^n$ est alors défini par :

$$R_n\Big((i_1, \dots, i_n)\Big) = \widehat{L}_n - \sum_{t=1}^{n} \ell_{i_t,t} \ .$$

3 Les algorithmes

Comme on l'a dit précedemment, le statisticien fait sa prédiction en pondérant à chaque tour les prédictions des expert. En pratique on va donc développer des algorithmes de mélange. Dans le cadre classique il existe beaucoup de tels algorithmes pour lesquels on a en plus des resultats théoriques sur des contrôles de regret (voir par exemple [CBL06] pour une très bonne référence dans le cadre classique). Dans la suite on présente les généralisations au cadre intermittent des plus pertinents de ses algorithmes.

3.1 Poids exponentiels

On cherche dans un premier temps à construire un algorithme qui permettrait d'avoir un contrôe sur le regret individuel. Dans le cadre classique il s'agit de l'algorithme de mélange par poids exponentiels (voir par exemple [CBL06, Section 2.1]) On peut essayer, grâce à cette définition, d'étendre l'algorithme de poids exponentiels (EWA) au cadre intermittent, Blum et Mansour donnent une version intermittente de cet algorithme (EWA-β) dans [BM05]. On donne ici une version un peu plus intuitive, avec le même contrôle théorique sur le regret et équivalente numériquement pour les valeurs de paramètres utilisés.

Cette version fait intervenir le regret $R'_{i,t}$ en exposant. L'encadré ci-dessous donne l'expression des poids du mélange à chaque tour.

EWA

paramètre : η

$$p_{i,t} = \frac{\mathbb{I}_{\{i \in E_t\}} e^{\eta R'_{i,t-1}}}{\sum_{j \in E_t} e^{\eta R'_{j,t-1}}}$$

Pour cet algorithme on a le résultat théorique suivant (on peut le déduire de [BM05] comme c'est fait dans le rapport technique [DGS09])

Théorème 1. *Pour les poids exponentiels (EWA), avec* η *de l'ordre de* $\frac{1}{B}\sqrt{\frac{\log N}{n}}$*, on a*

$$\max_{i\in\{1,\dots,N\}} R'_{i,n} \leqslant \mathcal{O}\left(B\sqrt{n\log N}\right) \ .$$

Démonstration. Le lemme suivant est utilisé dans [BM05] pour prouver la borne théorique de EWA-beta. Il vaut pour tout β-regret construit à partir d'une fonction de perte convexe en son premier argument, où on appelle β-regret la quantité $R'_{\beta,i,t-1} = \sum_{t\in E_i} \beta\ell_t - \ell_{i,t}$, pour tout $0 < \beta < 1$.

Lemme 1. *En posant* $W'_t = \sum_{i=1}^N \beta^{-R'_{\beta,i,t-1}}$ *pour tout* $t \geqslant 0$*, on a que* $W'_t \leqslant W'_{t+1}$ *quelque soit* $t \geqslant 0$*.*

On écrit $e^{\eta R'_{i,n}} = \beta^{-R'_{\beta,i,n}} \times e^{\eta(1-\beta)\sum_{t=1}^n \widehat{\ell}_t \mathbb{I}_{\{i\in E_t\}}}$.

D'après le Lemme 1, on a en particulier que $\beta^{-R'_{\beta,i,n}} \leqslant N$ pour tout $i \in \{1\dots,n\}$, ce qui nous donne :

$$e^{\eta R'_{i,n}} \leqslant N\, e^{\eta(1-\beta)\sum_{t=1}^n \widehat{\ell}_t \mathbb{I}_{\{i\in E_t\}}} \ .$$

En utilisant de plus que $e^{-\eta} \geqslant 1-\eta$, on obtient pour $\eta > 0$,

$$\eta R'_{i,n} \leqslant \log N + \eta(1-e^{-\eta})\sum_{t=1}^n \widehat{\ell}_t \mathbb{I}_{\{i\in E_t\}} \leqslant \log N + \eta^2 \sum_{t=1}^n \widehat{\ell}_t \mathbb{I}_{\{i\in E_t\}} \ .$$

En réarrangeant les termes et en bornant brutalement mais de manière uniforme la somme, on a

$$R'_{i,n} \leqslant \frac{\log N}{\eta} + \eta B^2 n \ .$$

La borne voulue est alors obtenue pour $\eta = \sqrt{\log N/(B^2 n)}$,

$$R'_{i,n} \leqslant 2B\sqrt{n\log N} \ .$$

□

Exponentielle des gradients

Comme la perte quadratique est convexe et différentiable, on a grâce à l'inégalité des pentes :

$$\ell(\widehat{y}_t, y_t) - \ell(f_{i,t}, y_t) = \ell_t(\mathbf{p}_t) - \ell_t(\delta_i) \leqslant \nabla\widehat{\ell}_t \cdot (\mathbf{p}_t - \delta_i) = \widetilde{\ell}(\widehat{y}_t, y_t) - \widetilde{\ell}(f_{i,t}, y_t) \ ,$$

où on a noté $\nabla\widehat{\ell}_t = \nabla\ell_t(\mathbf{p}_t)$ le gradient de la fonction $\mathbf{v} \mapsto \ell_t(\mathbf{v})$ en $\mathbf{p}_t$ et où on a introduit les pertes linéarisées définies pour tout $\mathbf{q} \in \mathcal{X}$ par

$$\widetilde{\ell}(\mathbf{q}\cdot\mathbf{f}_t, y_t) = \nabla\widehat{\ell}_t \cdot \mathbf{q} \ .$$

L'encadré ci-dessous définit les poids pour $t \geqslant 2$ de la méthode d'agrégation de l'exponentielle du gradient (EG), qui généralise celle du cas classique (on prend des poids

initiaux uniformes sur l'ensemble des experts actifs).

EG
paramètre : η

$$p_{i,t} = \frac{\mathbb{I}_{\{i\in E_t\}} \exp\left(\eta \sum_{s=1}^{t-1} \mathbb{I}_{\{i\in E_s\}} \nabla\widehat{\ell}_s \cdot (p_t - \delta_i)\right)}{\sum_{j\in E_t} \exp\left(\eta \sum_{s=1}^{t-1} \mathbb{I}_{\{j\in E_s\}} \nabla\widehat{\ell}_s \cdot (p_t - \delta_j)\right)}$$

On peut majorer le regret individuel par le regret des pertes linéarisées,

$$R'_{i,n} = \sum \mathbb{I}_{\{i\in E_t\}} \left(\widehat{\ell}_t - \ell_{i,t}\right) \leqslant \sum_{t=1}^{n} \mathbb{I}_{\{i\in E_t\}} \left(\widetilde{\ell}(\widehat{y}_t, y_t) - \widetilde{\ell}(f_{i,t}, y_t)\right)$$

et appliquer le Théorème précédant (les pertes linéarisées sont en particulier convexes) en remplaçant B par $2B^2$ qui borne les pertes linéarisées.

On obtient le Corollaire 1.

Corollaire 1. *Pour les poids donnés par EG, avec* η *de l'ordre de* $\frac{1}{2B^2}\sqrt{\frac{\log N}{n}}$*, on a*

$$\max_{i\in\{1,\dots,N\}} R'_{i,n} \leqslant \mathcal{O}\left(2B^2\sqrt{n \log N}\right) .$$

Remarque. Dans le cadre classique on a une majoration plus forte pour cet algorithme : on arrive en effet à avoir un contrôle sur un regret plus large, le regret convexe. L'algorithme suivant, un peu différent (un peu plus compliqué aussi), permet d'avoir un contrôle du regret convexe.

3.2 Régression Ridge

Dans le cadre classique des suites individuelles ridge regression est un algoritme simple et efficace pour contrôler le regret linéaire. L'idée est de définir le vecteur de pondération de la manière suivante,

$$u_t = \operatorname*{argmin}_{v\in\mathbb{R}^N} \left\{ \lambda\|v\|^2 + \sum_{s=1}^{t-1} (v \cdot f_s - y_s)^2 \right\} .$$

Cependant, on a vu que la généralisation du regret linéaire au cadre intermittent est problématique et pour cette raison, nous n'avons pas non plus trouvé d'adaptation très satisfaisante de cet algorithme.

- La première tentative est de garder la même définition du vecteur de pondération que dans le cas classique. Cela a bien un sens pour peu qu'on ait pris la convention suivante : si $i \notin E_t$ alors $f_{i,t} = 0$. Cependant, dans ce cas on visera à faire presqu'aussi bien que la meilleure combinaison linéaire fixe, les résultats numériques sont donc à la hauteur de l'objectif : décevants.
- Une deuxième idée (qui peut être appliquée aux autres algorithmes mais qui n'est pas très judicieuse) peut-être de se ramener au cadre classique en partitionnant. Cela revient en pratique à lancer K algorithmes en parallèle sur chacun des ensembles de

tours tels que $k_t = k$. Ce faisant, on se garantit l'existence de bornes théoriques déduites du cadre classique sur chaque sous-ensemble, en les sommant on obtient donc un majorant (plus grossier que dans le cas classique) du regret partitionné. On perd toutefois beaucoup d'information en lançant ces algorithmes en parallèle et les résultats numériques sont mauvais.

– En s'inspirant de la notion de regret renormalisé, on peut définir une troisième variante de la régression ridge dont les poids au temps t sont :

$$\mathbf{u}_t = \underset{\mathbf{v}\in\mathbb{R}^N}{\operatorname{argmin}} \left\{ \lambda\|\mathbf{v}\|^2 + \sum_{s=1}^{t-1} \mathbf{v}(E_s)\left(\mathbf{v}^{E_s}\cdot \mathbf{f}_s - y_s\right)^2 \right\} .$$

Les performances de cet algorithme, bien que meilleures, n'ont pas non plus été très concluantes. Cela est peut-être dû à des limites calculatoires. En effet l'implémentation de cet algorithme ne peut se faire récursivement contrairement aux autres méthodes, il faut donc effectuer réellement la minimisation à chaque tour.

3.3 Fixed-Share

Fixed-Share est une méthode qui permet de contrôler le regret *tracking* pour une taille de séquence fixe, voici comment l'extension (FSEWA) peut s'implémenter.

FSEWA

Paramètres : $\eta > 0$ et $0 \leqslant \alpha \leqslant 1$

Initialisation : $\mathbf{w}_0 = (1/N, \dots, 1/N)$

A chaque tour $t = 1, 2, \dots, n$,

(1) on forme la prédiction à partir des poids :

$$\mathbf{p}_t = \frac{\mathbf{w}_{t-1}}{\sum_{j=1}^N w_{j,t-1}} ;$$

(2) Mise à jour de la perte : on observe y_t et on actualise pour $i = 1, \dots, N$,

$$v_{i,t} = \begin{cases} w_{i,t-1}\, e^{\eta\left(\widehat{\ell}_t - \ell_{i,t}\right)} & \text{si } i \in E_t, \\ \text{non défini} & \text{si } i \notin E_t ; \end{cases}$$

(3) Répartition des poids : on pose $w_{i,t} = 0$ si $i \notin E_{t+1}$, et

$$w_{i,t} = \frac{1}{|E_{t+1}|} \sum_{j\in E_t \setminus E_{t+1}} v_{j,t} + \frac{\alpha}{|E_{t+1}|} \sum_{j\in E_t \cap E_{t+1}} v_{j,t} + (1-\alpha)\, \mathbb{I}_{\{i\in E_t\cap E_{t+1}\}} v_{i,t}$$

si $i \in E_{t+1}$ (avec la convention qu'une somme vide est nulle).

Remarque. Il existe aussi une version exponentielle des gradients de cette méthode (FSEG) : il suffit de remplacer les pertes par les pertes linéarisées dans l'étape de mise à jour de la perte. Comme dans la Section 3.1 les résultats théoriques valent pour les deux versions, en remplaçant B par $2B^2$ pour la version gradient.

On a le résultat théorique suivant pour $m \in \mathbb{N}$ (la preuve se déduit de [CBL06], voir [DGS09]).

Théorème 2. *En utilisant les poids donnés par FSEWA, pour* $\alpha = m/(n-1)$ *et* $\eta = \frac{1}{B^2}\sqrt{\frac{8}{n}\left((m+1)\log N + (n-1)\,H\left(\frac{m}{n-1}\right)\right)}$*, on a*

$$\max R_n((i_1,\dots,i_n)) \leqslant B^2\sqrt{\frac{n}{2}\left((m+1)\log N + (n-1)\,H\left(\frac{m}{n-1}\right)\right)}\ ,$$

où H *désigne l'entropie binaire :* $H(x) = -x\log x - (1-x)\log(1-x)$ *pour* $x \in [0,1]$ *et où le maximum est pris sur l'ensemble des séquences admissibles* $(i_1,\dots,i_n)$ *de taille* m*.*

La valeur du paramètre α théorique donnée par ce théorème peut s'interpréter comme un taux de ruptures, m étant le nombre de ruptures que l'on autorise au sein de notre séquence d'experts. Ce paramètre permet en pratique d'affiner la sensibilité aux ruptures (qui correspondent à un changement de meilleur expert). Cette adaptativité fait de cette méthode une des plus efficace dans le cadre intermittent.

4 Calibration des paramètres via une grille

Les différentes méthodes d'agrégation présentées dans les sections précédentes font intervenir un paramètre ou un couple de paramètres (qui peut être considéré comme le nouveau paramètre). Le choix de ce paramètre est crucial dans leur application à un jeu de données.

En pratique, on a utilisé une calibration automatique du paramètre à l'aide d'une grille de valeurs en prenant à chaque instant la valeur du paramètre qui a donné les meilleures performances sur le passé parmi cette grille.

Plus précisément, si on considère une méthode dépendant d'un paramètre λ (pour la méthode fixed-share le paramètre est en fait un couple de paramètres) et dont le vecteur de poids au tour t est noté $\mathbf{u}_t = \mathbf{u}_t^{(\lambda)}$, notre but est de choisir le paramètre $\lambda \in \Lambda$ de manière automatique. L'espace des paramètres est, par exemple, $\Lambda =]0,+\infty[$ pour la régression ridge, la méthode des poids exponentiels, et celle de l'exponentielle des gradients.

On suppose de plus que $\mathbf{u}_1^{(\lambda)} = \mathbf{u}_1^\star$ ne dépend pas de λ, ce qui est bien le cas pour toutes les méthodes que l'on a étudiées jusqu'à présent.
La méthode de calibration proposée se base sur la minimisation de la perte empirique. On choisit $\mathbf{u}_1 = \mathbf{u}_1^\star$ et, pour $t \geqslant 2$,

$$\mathbf{u}_t = \mathbf{u}_t^{(\widehat{\lambda}_t)} \qquad \text{où} \qquad \widehat{\lambda}_t \in \operatorname*{argmin}_{\lambda\in\Lambda} \sum_{s=1}^{t-1} \left(\mathbf{u}_s^{(\lambda)}\cdot \mathbf{f}_s - y_s\right)^2 .$$

Résultats pratiques

J'ai testé les différentes méthodes exposées ainsi que quelques unes de leurs variantes sur les deux jeux de données. On retrouve plusieurs conclusions dans les deux cas : on

gagne à utiliser les pertes linéarisées plutôt que les vraies pertes dans les algorithmes de base (EG/EWA) et on obtient d'assez bons résultats avec la calibration en ligne même avec des grilles assez grossières, FSEG faisant exception sur ce point dans le cas des données slovaques. Certaines caractéristiques du cadre intermittent semblent également apparaître : fixed-share est une méthode plus adaptée quand les experts sont fortement intermittents. Les tableaux suivants présentent les résultats obtenus, la Table 1 contient quelques valeurs de référence dont

- n le nombre total de tour, pour les données slovaques comme on a séparé les données en vingt-quatre on donne donne le nombre de tours pour un jeu de donnée à heure fixe,
- K le nombre d'ensemble d'indices d'experts actifs différents,
- l'erreur moyenne (rapportée à sa présence) du meilleur expert (ME)
- l'erreur moyenne de la meilleure combinaison linéaire d'expert (MCL)
- l'erreur moyenne de la meilleure combinaison convexe renormalisée à chaque tour (MCC)
- l'erreur moyenne de la meilleure séquence d'experts de taille inférieure ou égale à 50 ($\mathrm{MSE}_{\leqslant 50}$).

La Table 2 donne les performances obtenues par les algorithmes pour les meilleures valeurs des paramètres hors ligne (HL) et avec la méthode de calibration (cf Section 4) pour une grille logarithmique de 21 points dans le cas des algorithmes EWA et EG pour les données slovaques et françaises, et pour les deux dérivés de fixed-share des grilles de 20 points ont été utilisées. L'algorithme Ridge utilisé est le premier présenté dans la Section 3.2.

	n	K	ME	MCL	MCC	$\mathrm{MSE}_{\leqslant 50}$
Données slovaques	1095 (x24)	74	30.4	40.7	29.2	23.1
Données françaises	15 360	7	0.782	0.620	0.696	0.534

TABLE 1 – Références

	EWA		EG		Ridge	FSEWA		FSEG	
	HL	EL	HL	EL	HL	HL	EL	HL	EL
Don. slov.	30.5	30.6	28.2	28.2	41.8	27.0	27.4	27.2	28.3
Don. fran.	0.683	0.692	0.651	0.661	0.833	0.675	0.697	0.655	0.662

TABLE 2 – Performances

La Figure 2 représente l'évolution de la perte cumulée dans le cas des données françaises, des experts (en couleur) et des algorithmes EG (trait noir) et FSEWA (trait pointillé) pour le meilleur choix de paramètres, quand les experts sont inactifs leur perte instantanée est représentée comme étant nulle et leur perte cumulée comme étant constante.

Les résultats sont assez satisfaisants dans le sens où plusieurs algorithmes permettent de faire beaucoup mieux que le meilleur expert. Cependant on a noté une grande instabilité numérique dans le fonctionnement des algorithmes. Par exemple, mettre le regret ou la

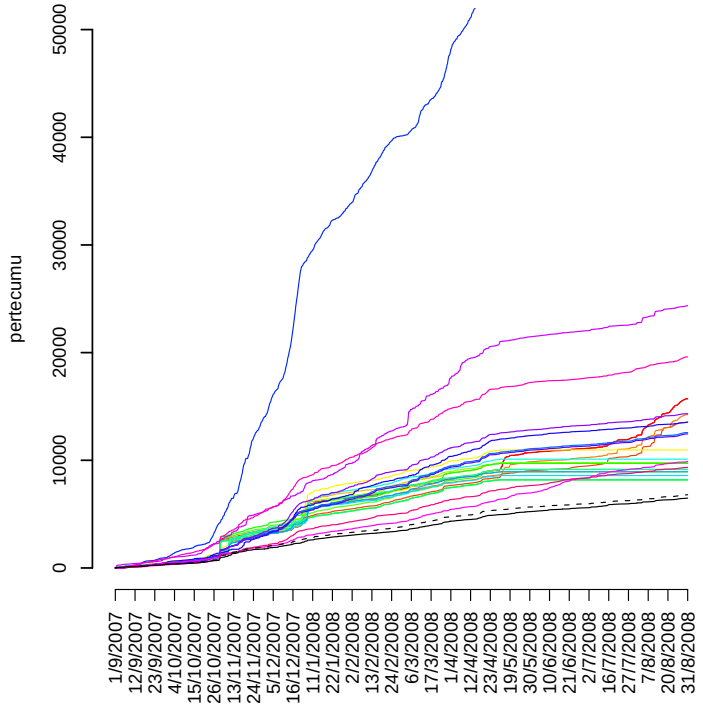

FIGURE 2 – Perte cumulée des experts et des méthodes EG et FSEWA

perte en exposant dans fixed-share devrait théoriquement donner exactement les mêmes vecteurs de pondération, or ce n'est pas le cas numériquement.

Conclusion

Comme les résultats pratiques le montrent, il peut s'avérer très avantageux d'appliquer la théorie des suites individuelles à des données de consommation d'électricité. De plus, le cadre intermittent offre une souplesse supplémentaire : il permet d'utiliser de manière optimale des experts "spécialisés" (par exemple, certains modèles avec certaines valeurs de paramètres peuvent-être plus adaptés pour former des prédctions en hiver..). Compte tenu de cette possibilité il pourrait être avantageux de développer plus spécifiquement des modèles dans une optique de spécialisation.

En ce qui concerne la calibration des paramètres, il serait intéressant d'avoir une calibration vraiment automatique : choisir la grille de manière prédeterminée en se basant sur les valeurs théoriques optimales des paramètres. Une piste de recherche serait à ce moment de chercher des bornes théoriques sur les "méta-algorithmes" obtenus. On peut aussi se poser la question d'intervalles de confiance existant pour un algorithme de mélange en fonction des intervalles de confiance des experts de base.

Références

[BM05] Blum, Avrim et Yishay Mansour : *From External to Internal Regret*. Dans *Proceedings of COLT*, pages 621–636, 2005.

[CBL06] Cesa-Bianchi, Nicolò et Gàbor Lugosi : *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[DGS09] Devaine, Marie, Yannig Goude et Gilles Stoltz : *Aggregation of sleeping predictors to forecast electricity consumption*, 2009.

[GMS08] Gerchinovitz, Sébastien, Vivien Mallet et Gilles Stoltz : *A Further Look at Sequential Aggregation Rules for Ozone Ensemble Forecasting*, 2008. Available at http://www.dma.ens.fr/~stoltz/GeMaSt-report.pdf.

[HW95] Herbster, Mark et Manfred Warmuth : *Tracking the Best Expert*. Dans *Machine Learning*, pages 286–294. Morgan Kaufmann, 1995.