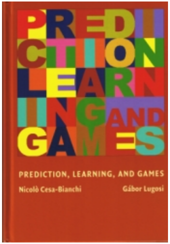
# Prédiction avec experts : Statistiques déterministes appliquées à la prédiction de la qualité de l'air

Gilles Stoltz

CNRS — École normale supérieure — INRIA, équipe CLASSIC

& HEC Paris

Au début de cet exposé, je voudrais saluer




Prediction, Learning, and Games

Nicolò Cesa-Bianchi et Gábor Lugosi

Et indiquer que tout ce qui suit est relaté dans un article de survol paru à la fin 2010 dans le Journal de la Société Française de Statistique

SFdS SOCIÉTÉ FRANÇAISE DE STATISTIQUE

Journal de la Société Française de Statistique
**Vol. 151** No. 2 (2010)

**Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique**

**Title:** Sequential aggregation of predictors: General methodology and application to air-quality forecasting and to the prediction of electricity consumption

Gilles Stoltz *

**Résumé :** Cet article fait suite à la conférence que j'ai eu l'honneur de donner lors de la réception du prix Marie-Jeanne Laurent-Duhamel, dans le cadre des XL$^e$ Journées de Statistique à Ottawa, en 2008. Il passe en revue les résultats fondamentaux, ainsi que quelques résultats récents, en prévision séquentielle de suites arbitraires par agrégation d'experts. Il décline ensuite la méthodologie ainsi décrite sur deux jeux de données, l'un pour un problème de prévision de qualité de l'air, l'autre pour une question de prévision de consommation électrique. La plupart des résultats mentionnés dans cet article reposent sur des travaux en collaboration avec Yannig Goude (EDF R&D) et Vivien Mallet (INRIA), ainsi qu'avec les stagiaires de master que nous avons co-encadrés : Marie Devaine, Sébastien Gerchinovitz et Boris Mauricette.

**Abstract:** This paper is an extended written version of the talk I delivered at the "XL$^e$ Journées de Statistique" in Ottawa, 2004, when being awarded the Marie-Jeanne Laurent-Duhamel prize. It is devoted to surveying some fundamental as well as some more recent results in the field of sequential prediction of individual sequences with expert advice. It then performs two empirical studies following the stated general methodology: the first one to air-quality forecasting and the second one to the prediction of electricity consumption. Most results mentioned in the paper are based on joint works with Yannig Goude (EDF R&D) and Vivien Mallet (INRIA), together with some students whom we co-supervised for their M.Sc. theses: Marie Devaine, Sébastien Gerchinovitz and Boris Mauricette.

***Classification AMS 2000 :*** primaire 62-02, 62L99, 62P12, 62P30

***Mots-clés :*** Agrégation séquentielle, prévision avec experts, suites individuelles, prévision de la qualité de l'air, prévision de la consommation électrique

***Keywords:*** Sequential aggregation of predictors, prediction with expert advice, individual sequences, air-quality forecasting, prediction of electricity consumption

Ecole normale supérieure, CNRS, 45 rue d'Ulm, 75005 Paris
& HEC Paris, CNRS, 1 rue de la Libération, 78350 Jouy-en-Josas
E-mail : gilles.stoltz@ens.fr
URL : http://www.math.ens.fr/~stoltz

* L'auteur remercie l'Agence nationale de la recherche pour son soutien à travers le projet JCJC06-137444 ATLAS ("From applications to theory in learning and adaptive statistics").
† Ces recherches ont été menées dans le cadre du projet CLASSIC de l'INRIA, hébergé par l'Ecole normale supérieure et le CNRS.

*Journal de la Société Française de Statistique*, **Vol. 151** No. 2 66-106
http://www.sfds.asso.fr/journal

Un statisticien accepte la mission de prédire une suite $y_1, y_2, \ldots$ d'observations vivant dans un ensemble $\mathcal{Y}$.

Ses prédictions $\widehat{p}_1, \widehat{p}_2, \ldots$ sont formées dans un ensemble $\mathcal{X}$.

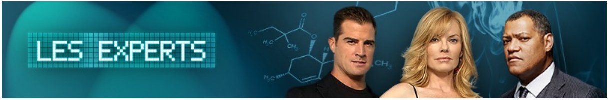
Les observations et prédictions (1) sont effectuées de manière séquentielle et (2) ne reposent sur aucun modèle stochastique.

(1) signifie qu'à chaque échéance, la prévision $\widehat{p}_t$ pour $y_t$ est déterminée

- sur le seul fondement du passé, $y_1^{t-1} = \big(y_1, \ldots, y_{t-1}\big)$,
- et avant que la vraie valeur $y_t$ ne soit révélée au grand jour.

(2) indique qu'il ne s'agira pas d'estimer des paramètres pour en déduire un bon modèle et donc de bonnes prévisions.

Pour que le problème ait un sens dans un cadre aussi général, on introduit des experts.



Pour que le problème ait un sens dans un cadre aussi général, on introduit des experts ; il leur sera fait référence par $j = 1, \ldots, N$.

A chaque échéance, l'expert $j$ procure une prédiction $f_{j,t} = f_{j,t}\bigl(y_1^{t-1}\bigr) \in \mathcal{X}$.

Le statisticien fonde maintenant ses prédictions $\widehat{p}_t$ sur les observations passées $y_1^{t-1} = (y_1, \ldots, y_{t-1})$ et sur les conseils passés et présents des experts, $f_{j,s}$ pour $s = 1, \ldots, t$.

L'objectif du statisticien est de prédire presqu'aussi bien que le meilleur expert.

Remarquez cependant que le meilleur expert ne saurait être déterminé que rétrospectivement tandis que le statisticien est assujetti à une contrainte de prédiction séquentielle.

On se limite à former des combinaisons linéaires ou convexes des conseils des experts (convexes, pour commencer).

On suppose donc que $\mathcal{X}$ est convexe.

A chaque échéance, le statisticien forme une combinaison convexe des conseils des experts,

$$\widehat{p}_t = \sum_{j=1}^{N} p_{j,t}\, f_{j,t}$$

où $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{N,t})$ est un élément du simplexe (de probabilité) d'ordre $N$.

On renforce l'objectif : le statisticien veut en fait prédire presqu'aussi bien que la meilleure combinaison convexe constante des experts.

Pour quantifier mathématiquement la notion de meilleur expert, on introduit une fonction de perte $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

On définit les pertes cumulées du statisticien et des combinaisons convexes constantes $\boldsymbol{q} = (q_1, \ldots, q_N)$ des experts par

$$\widehat{L}_n = \sum_{t=1}^{n} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\, y_t\right) \quad \text{et} \quad L_n(\boldsymbol{q}) = \sum_{t=1}^{n} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\, y_t\right)$$

Le regret face à $\boldsymbol{q}$ est la différence entre ces deux quantités,

$$R_n(\boldsymbol{q}) = \widehat{L}_n - L_n(\boldsymbol{q})$$

On veut construire des stratégies de prédiction telles que le regret moyen converge vers 0 pour toute suite d'observations $y_1, y_2, \ldots$, uniformément en les $\boldsymbol{q}$, soit

$$\limsup \ \frac{1}{n} \max_{\boldsymbol{q}} R_n(\boldsymbol{q}) \ \leqslant \ 0$$

Ce cadre admet une interprétation méta-statistique :

- chaque expert peut correspondre à une méthode statistique, éventuellement réglée avec un certain jeu de paramètres ;
- on combine de manière robuste et déterministe ces prévisions fondamentales issues d'une modélisation stochastique.

Or, la perte cumulée du statisticien est égale à

$$\widehat{L}_n = \min_{\boldsymbol{q}} L_n(\boldsymbol{q}) + \max_{\boldsymbol{q}} R_n(\boldsymbol{q})$$

On peut donc interpréter

- le terme de performance de la meilleure combinaison convexe des experts comme une erreur d'approximation,
- le terme de regret comme mesurant une difficulté d'estimation séquentielle.

On résume le cadre mathématique introduit par un jeu répété.

*Paramètres* : un ensemble convexe $\mathcal{X}$ de prédictions, un ensemble $\mathcal{Y}$ d'observations, une fonction de perte $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

*A chaque* échéance $t = 1, 2, \ldots,$

- les experts procurent leurs conseils $f_{j,t}$
- le statisticien choisit une combinaison convexe $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{N,t})$ et prédit $\widehat{p}_t = \sum_{j=1}^{N} p_{j,t} \, f_{j,t}$
- l'environnement choisit simultanément l'observation $y_t$
- $y_t$ et $\widehat{p}_t$ (et donc les pertes respectives) sont révélées.

On veut minimiser le regret face à tous les $\boldsymbol{q}$,

$$R_n(\boldsymbol{q}) = \widehat{L}_n - L_n(\boldsymbol{q}) = \sum_{t=1}^{n} \ell \left( \sum_{j=1}^{N} p_{j,t} \, f_{j,t}, \; y_t \right) - \sum_{t=1}^{n} \ell \left( \sum_{j=1}^{N} q_j \, f_{j,t}, \; y_t \right)$$

Voici le fruit d'une collaboration avec Vivien Mallet (INRIA), publiée par *Journal of Geophysical Research*.

On veut prédire, jour après jour, les hauteurs des pics d'ozone du lendemain (ou les concentrations horaires, heure après heure).

On dispose d'un réseau de stations météorologiques à travers l'Europe pour les relever (tout se passe pendant l'été 2001).

On construit tout d'abord 48 prédicteurs fondamentaux en choisissant pour chacun d'entre eux un modèle, défini par une formulation physico-chimique (parmi plusieurs possibles), un schéma numérique (parmi plusieurs possibles) de résolution approchée des EDPs en jeu, et un jeu de données d'entrée.

Voici le fruit d'une collaboration avec Vivien Mallet (INRIA), publiée par *Journal of Geophysical Research*.

On veut prédire, jour après jour, les hauteurs des pics d'ozone du lendemain (ou les concentrations horaires, heure après heure).

On dispose d'un réseau de stations météorologiques à travers l'Europe pour les relever (tout se passe pendant l'été 2001).

On construit tout d'abord 48 prédicteurs fondamentaux en choisissant pour chacun d'entre eux un modèle.

Au lieu de devoir se fier à un prédicteur plutôt qu'un autre en le sélectionnant, on recourt à une procédure plus gloutonne qui les considère tous et les agrège séquentiellement.

On dispose d'un réseau $\mathcal{S}$ de stations à travers l'Europe et chaque modèle $j = 1, \ldots, 48$ procure une prédiction $f_{j,t}^s$ pour le pic à la station $s$ et au jour $t$, qui est ensuite comparée au pic réalisé $y_t^s$.

Le statisticien détermine chaque jour une unique combinaison convexe $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{N,t})$ à utiliser en toutes les stations pour agréger les prédictions (et obtenir ainsi un champ de prévisions).

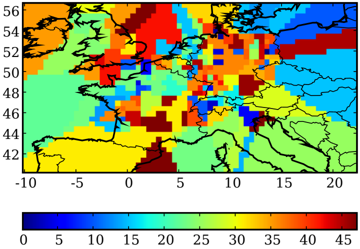
Les écarts sont mesurés en perte quadratique moyenne, ce qui revient à considérer la fonction de perte

$$\ell\big(\boldsymbol{p}_t, (y_t^s)_{s \in \mathcal{S}_t}\big) = \sum_{s \in \mathcal{S}_t} \left( \sum_{j=1}^{48} p_{j,t}\, f_{j,t}^s - y_t^s \right)^2$$

où $\mathcal{S}_t$ est le sous-ensemble des stations actives au jour $t$.

La définition s'étend au cas des combinaisons linéaires $\boldsymbol{u}_t$ (qui permettent par exemple de réduire le biais des modèles).

Les figures ci-dessous montrent que tous les experts sont utiles et apportent de l'information.



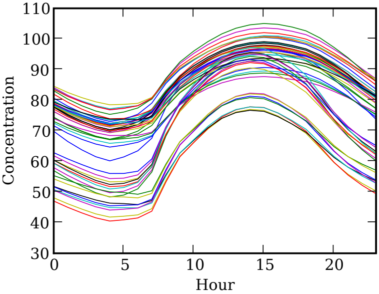




Figure: A gauche : Coloration de l'Europe en fonction de l'indice du meilleur expert local. A droite : Profils moyens de prédiction sur une journée (moyennes spatiales et temporelles, en $\mu g/m^3$).

Les erreurs cumulées de la méthode d'agrégation et de la combinaison linéaire constante induite par $\boldsymbol{u}$ valent respectivement

$$\widehat{L}_n = \sum_{t=1}^{n} \sum_{s \in \mathcal{S}_t} \left( \sum_{j=1}^{48} u_{j,t} \, f_{j,t}^s - y_t^s \right)^2$$

$$\text{et} \qquad L_n(\boldsymbol{u}) = \sum_{t=1}^{n} \sum_{s \in \mathcal{S}_t} \left( \sum_{j=1}^{48} u_j \, f_{j,t}^s - y_t^s \right)^2$$

où $\mathcal{S}_t$ est le sous-ensemble des stations actives au jour $t$.

Les erreurs quadratiques moyennes associées sont données par

$$\widehat{r}_n = \sqrt{\frac{\widehat{L}_n}{\sum_{t=1}^{n} |\mathcal{S}_t|}} \qquad \text{et} \qquad r_n(\boldsymbol{u}) = \sqrt{\frac{L_n(\boldsymbol{u})}{\sum_{t=1}^{n} |\mathcal{S}_t|}}$$

L'espoir est qu'un bon ensemble d'experts et la considération d'une procédure avec un faible regret entraînent à leur tour une faible erreur quadratique moyenne.

En effet,

$$\widehat{L}_n \leqslant \inf_{\boldsymbol{u} \in U} L_n(\boldsymbol{u}) + o(n)$$

se ré-écrit comme

$$\left(\widehat{r}_n\right)^2 \leqslant \inf_{\boldsymbol{u} \in U} \left(r_n(\boldsymbol{u})\right)^2 + o(1)$$

($U$ est par exemple le simplexe des probabilités ou une boule $\ell^1$).

| Moyenne | M. fondamental | M. convexe | M. linéaire | Prescient |
|---|---|---|---|---|
| 24.41 | 22.43 | 21.45 | 19.24 | 11.99 |

Ci-dessus, les erreurs quadratiques moyennes (en $\mu g/m^3$)

- de la moyenne des prédictions des 48 modèles, i.e., $r_n\big((1/48,\ldots,1/48)\big)$,
- du meilleur modèle fondamental parmi $j = 1,\ldots,48$,
- de la meilleure combinaison convexe $\boldsymbol{q}$ des 48 modèles, i.e., $\min_{\boldsymbol{q}} r_n(\boldsymbol{q})$,
- de la meilleure combinaison linéaire $\boldsymbol{u}$ (parmi tous les vecteurs de $\mathbb{R}^{48}$) des 48 modèles, i.e., $\min_{\boldsymbol{u}} r_n(\boldsymbol{u})$,
- du prédicteur prescient qui aurait connaissance des $y_t^s$ avant de former sa prédiction et ne serait contraint que par l'obligation de choisir une combinaison linéaire des prédictions des modèles.

Nous avons mis en œuvre environ 20 méthodes d'agrégation différentes et nous concentrons ici sur deux familles qui ont obtenu de bons résultats, EG et la régression ridge (et leurs variantes).

EG est l'abréviation d'exponentielle des gradients. Cette méthode forme des combinaisons convexes dont les composantes sont données par une pondération exponentielle des sommes des composantes des gradients des pertes passées.

Son regret moyen par rapport à l'ensemble des combinaisons convexes constantes est plus petit que $1/\sqrt{n}$.

La régression ridge est une méthode d'estimation classique en perte quadratique et qui utilise la meilleure combinaison linéaire pénalisée sur les données passées (pénalisation en terme de norme $\ell^2$).

Son regret moyen par rapport à toute combinaison linéaire constante est plus petite qu'une quantité de l'ordre de $(\ln n)/n$.

Les versions fenêtrées n'utilisent qu'un nombre fixe des plus récentes pertes passées, pour ensuite pondérer exponentiellement leurs gradients (EG) ou calculer sur elles seulement une meilleure combinaison linéaire pénalisée (régression ridge).

L'escompte multiplie chaque perte passée par un facteur d'autant plus petit que ce passé est lointain.

| EG | EG fenêtré | EG esc. | Ridge | Ridge fenêtrée | Ridge esc. |
|---|---|---|---|---|---|
| 21.47 | 21.37 | 21.31 | 20.77 | 20.03 | 19.45 |

La meilleure combinaison convexe constante est battue et la version escomptée de la régression ridge a des performances très proches de celles de la meilleure combinaison linéaire constante.

| Moyenne | M. fondamental | M. convexe | M. linéaire | Prescient |
|---|---|---|---|---|
| 24.41 | 22.43 | 21.45 | 19.24 | 11.99 |

Les méthodes d'agrégation séquentielle ne se concentrent pas sur un seul expert.

Les poids attribués aux modèles peuvent changer rapidement et de manière significative au cours du temps.

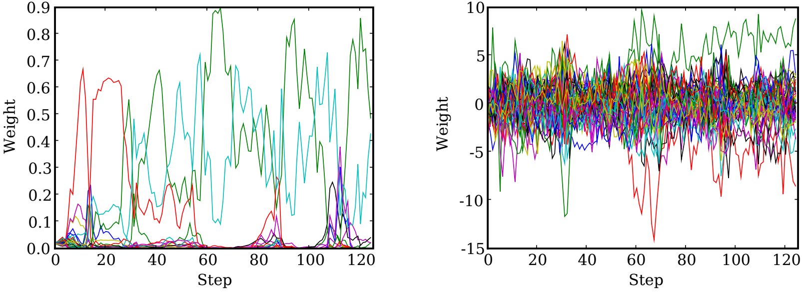



Figure: Poids produits au cours du temps par (à gauche) EG et la version escomptée de la régression ridge (à droite).

D'une part, on peut s'intéresser à la sélection séquentielle de portefeuilles d'investissement (par exemple, Cover'91).

Le tout est de déterminer de bons experts. Les valeurs boursières individuelles prises comme experts ne donnent pas de bons résultats.

De plus, ici l'on est amené à considérer des log-rendements dans la définition du regret.

Au final, les performances pratiques sont mauvaises.

D'autre part, la thèse de Yannig Goude (EDF & Université Paris-Sud), soutenue en janvier 2008, a porté sur la prévision de consommation électrique.

Les experts sont le système de prédiction Eventail, calibré avec différents jeux de paramètres.

Ici, à cause des variations fortes du profil de charge au cours de la semaine et de l'année, l'approche prise est de se comparer non pas seulement à la meilleure combinaison convexe constante, mais à la meilleure suite de combinaisons convexes avec $k$ changements au plus dans les $n$ pas de prédiction.

Nous avons continué ces travaux en 2009 avec Marie Devaine (voir l'article de survol dans le Journal de la SFdS).

Soit un ensemble convexe de prédictions $\mathcal{X}$, l'espace des observations $\mathcal{Y}$ et une fonction de perte convexe $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

L'algorithme EG choisit successivement des combinaisons convexes $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots$ des prédictions des experts et ses pertes sont données par

$$\widetilde{\ell}_t(\boldsymbol{p}_t) = \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\, y_t\right)$$

Chaque $\boldsymbol{p}_t$ ne dépend que des (gradients des) pertes passées,

$$\boldsymbol{p}_1 = (1/N, \ldots, 1/N)$$

et la $j$-ième composante de $\boldsymbol{p}_t$ est définie, pour $t \geqslant 2$, par une pondération exponentielle,

$$p_{j,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(\nabla \widetilde{\ell}_s(\boldsymbol{p}_s)\right)_j\right)}{\sum_{i=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \left(\nabla \widetilde{\ell}_s(\boldsymbol{p}_s)\right)_i\right)}$$

## Theorem

*Le regret de EG face à toute combinaison convexe constante $\boldsymbol{q}$ est uniformément borné (en $\boldsymbol{q}$ et en les suites $y_1, y_2, \ldots$) selon*

$$\sum_{t=1}^{n} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\, y_t\right) - \sum_{t=1}^{n} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\, y_t\right) \leqslant \frac{\ln N}{\eta} + \frac{\eta n}{2} B^2$$

*où $B$ est une borne sur les gradients, $\left\|\nabla\widetilde{\ell}_t\right\|_\infty \leqslant B$ pour tout $t$.*

Deux éléments de démonstration : par convexité,

$$\ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\, y_t\right) - \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\, y_t\right) \leqslant \nabla\widetilde{\ell}_t(\boldsymbol{p}_t) \cdot (\boldsymbol{p}_t - \boldsymbol{q})$$

et l'analyse de ce majorant (linéaire en $\boldsymbol{q}$) repose sur le lemme de Hoeffding.

Il faut calibrer $\eta$ : le regret est plus petit que

$$\frac{\ln N}{\eta} + \frac{\eta n}{2} B^2 = \square \, B \sqrt{n \ln N}$$

avec le choix (optimal pour la théorie) $\eta = \square \, (1/B) \sqrt{(\ln N)/n}$.

Mais $n$ et/ou $B$ peuvent être inconnus ; une version adaptative de EG est donnée par les mêmes formules où l'on remplace simplement le paramètre fixe $\eta$ par une suite adaptative $(\eta_t)$,

$$\eta_{t+1} = \frac{1}{\max_{s \leqslant t} \left\| \nabla \widetilde{\ell}_s \right\|_\infty} \sqrt{\frac{\ln N}{t}}$$

La borne sur son regret est similaire.

Bibliographie : EG a été introduit par Vovk '90, Littlestone et Warmuth '94, Cesa-Bianchi '99 et des versions adaptatives ont été étudiées par Auer, Cesa-Bianchi et Gentile '02, Cesa-Bianchi, Mansour et Stoltz '07.

La version fenêtrée repose sur une largeur de fenêtre $T$ et produit les combinaisons convexes données par

$$p_{j,t} = \frac{\exp\left(-\eta \sum_{s=\max\{1,\,t-T\}}^{t-1} \left(\nabla\widetilde{\ell}_s(\boldsymbol{p}_s)\right)_j\right)}{\sum_{i=1}^{N} \exp\left(-\eta \sum_{s=\max\{1,\,t-T\}}^{t-1} \left(\nabla\widetilde{\ell}_s(\boldsymbol{p}_s)\right)_i\right)}$$

La version escomptée utilise une suite décroissante $(\beta_s)$ pour former

$$p_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} (1+\beta_{t-s}) \left(\nabla\widetilde{\ell}_s(\boldsymbol{p}_s)\right)_j\right)}{\sum_{i=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} (1+\beta_{t-s}) \left(\nabla\widetilde{\ell}_s(\boldsymbol{p}_s)\right)_i\right)}$$

On peut exhiber une borne théorique sur le regret de la version escomptée, dépendant de $(\beta_s)$.

Pour la prédiction de la qualité de l'air, nous avons utilisé des escomptes assez forts, $\beta_s = 100/s^2$.

En s'inspirant de Cover '91 et Blum et Kalai '97, on peut définir, pour les pertes exp-concaves, un mélange uniforme adaptatif sur le simplexe.

S'il existe $\eta > 0$ tel que pour tous $y_t$ et $f_{j,t}$,

$$\boldsymbol{p} \mapsto \exp\bigl(-\eta\, \ell_t(\boldsymbol{p})\bigr)$$

est concave, alors agréger selon

$$\boldsymbol{p}_t = \frac{\int \boldsymbol{p}\, \exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(\boldsymbol{p})\right)\, \mathrm{d}\mu(\boldsymbol{p})}{\int \exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(\boldsymbol{p})\right)\, \mathrm{d}\mu(\boldsymbol{p})}$$

(où $\mu$ est la mesure uniforme sur le simplexe) assure que le regret est plus petit que $\square\, (N \ln n)/\eta$, ce qui est d'un ordre de grandeur bien plus petit que $\sqrt{n}$.

Cette hypothèse est vérifiée dans le cas de l'investissement boursier et lorsque $\ell$ est la perte quadratique.

La régression ridge a été introduite dans les années 70 par Hoerl et Kennard et intensivement étudiée depuis dans un cadre stochastique.

Vovk '01 et Azoury et Warmuth '01 en proposent une analyse pour des suites individuelles.

Formellement, en perte quadratique, la régression ridge choisit des combinaisons linéaires $\boldsymbol{u}_t$ des prédictions des experts données, à l'échéance $t \geqslant 2$, par un critère de moindres carrés pénalisés,

$$\boldsymbol{u}_t \in \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^N} \left[ \lambda \, \|\boldsymbol{u}\|_2^2 + \sum_{s=1}^{t-1} \left( y_s - \sum_{j=1}^{N} u_j \, f_{j,s} \right)^2 \right]$$

Elle peut être mise en œuvre efficacement de manière séquentielle et assure que son regret est $O(\ln n)$.

Une propriété tout à fait sympathique de la régression ridge est qu'elle semble débiaiser automatiquement les experts.

On peut en effet la faire tourner sur un seul expert (proposant les prédictions $f^s_{j,t}$) et faire ainsi presqu'aussi bien que le meilleur des experts, indexés chacun par $\gamma$, proposant les prédictions $\gamma\, f^s_{j,t}$.

S'il y a un facteur de biais multiplicatif à-peu-près contant $1/\gamma$, il est donc corrigé.

Sur les données d'ozone, cela donne les erreurs quadratiques moyennes suivantes, par exemple sur le meilleur et le moins bon modèle :

| Sans Ridge | Avec Ridge | Sans Ridge | Avec Ridge |
|---|---|---|---|
| 35.79 | 24.78 | 22.43 | 21.66 |

Dans ce qui suit, je vais passer rapidement en revue quelques extensions.

Elles portent sur

- une meilleure calibration des paramètres d'apprentissage,
- la prédiction lacunaire : la sélection séquentielle d'un sous-ensemble de modèles pour la prédiction,
- d'autres objectifs, que je n'aurai pas le temps de détailler, mais juste de mentionner.

J'ai fait allusion à la calibration automatique de l'exponentielle des gradients.

Mais cette dernière est trop précautionneuse en pratique, même si les bornes théoriques correspondantes ne sont modifiées que d'un facteur multiplicatif et pas dans leurs ordres de grandeur.

On veut ici une méthode plus efficace.

On rappelle que l'exponentielle des gradients prédit, pour $t \geqslant 2$, avec $\boldsymbol{p}_t$ défini, composante $j$ par composante $j$ selon

$$p_{j,t}(\eta) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1}\left(\nabla\widetilde{\ell}_s(\boldsymbol{p}_s)\right)_j\right)}{\sum_{i=1}^{N}\exp\left(-\eta \sum_{s=1}^{t-1}\left(\nabla\widetilde{\ell}_s(\boldsymbol{p}_s)\right)_i\right)}$$

L'idée ici est de faire varier $\eta$ en fonction de $t$ et considérer pour $\eta_t$ le meilleur paramètre $\eta$ sur les échéances $1, \ldots, t-1$,

$$\eta_t \in \operatorname*{argmin}_{\eta > 0} \sum_{s=1}^{t-1} \ell_s\big(\boldsymbol{p}_s(\eta)\big) \,.$$

On utilise alors $\boldsymbol{p}_t(\eta_t)$ pour la prédiction au jour $t$.

On peut définir de manière similaire une calibration automatique de Ridge. Sur les données d'ozone :

| | |
|---|---|
| Meilleure convexe | 21.45 |
| EG avec meilleur $\eta$ | 21.47 |
| EG avec $(\eta_t)$ | 21.80 |
| Meilleure linéaire | 19.24 |
| Ridge avec meilleur $\lambda$ | 20.77 |
| Ridge avec $(\lambda_t)$ | 20.81 |

Le “meilleur” paramètre désigne le paramètre constant $\eta$ ou $\lambda$, choisi de manière rétrospective, qui aurait donné les meilleurs résultats en termes d'erreur quadratique.

Il n'y a pas encore de borne théorique pour cette méthode de calibration, mais nous y travaillons !

Pour obtenir des combinaisons linéaires ou convexes n'utilisant qu'un nombre restreint de modèles, on peut seuiller les combinaisons proposées (pour EG) ou changer le type de pénalité (pour Ridge).

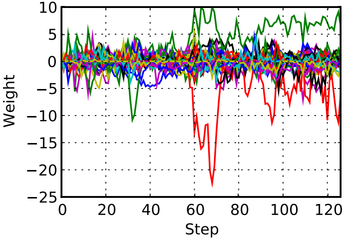
La méthode LASSO (Tibshirani, '96) choisit des combinaisons linéaires $\boldsymbol{u}_t$ des prédictions des experts données, à l'échéance $t \geqslant 2$, par un critère de moindres carrés pénalisés en norme $\ell^1$,

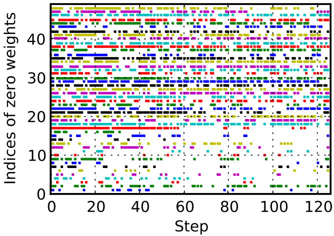
$$\boldsymbol{u}_t = \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^N} \left[ \lambda \, \|\boldsymbol{u}\|_1 + \sum_{s=1}^{t-1} \left( y_s - \sum_{j=1}^{N} u_j \, f_{j,s} \right)^2 \right]$$

Les combinaisons qui en résultent ont en général de nombreux coefficients nuls (et sont dites lacunaires).

Une version escomptée de LASSO conduit ainsi à une très forte sélection parmi les modèles (une vingtaine est éliminée sur les données d'ozone).

| Ridge esc. | LASSO esc. | M. linéaire |
|---|---|---|
| 19.45 | 19.31 | 19.24 |



D'autres résultats ont été obtenus :

- Sur un ensemble de 100 modèles et pendant un an, le gain de performance relative des méthodes présentées par rapport à la meilleure combinaison convexe ou linéaire est encore meilleur !
- La prédiction horaire, en faisant tourner 24 méthodes d'agrégation en parallèle est possible (et améliore également la performance relative).
- Pour la prédiction des événements extrêmes, comme le dépassement de seuils d'alerte par exemple, on peut voir que Ridge est plus performante que le meilleur modèle.

Malheureusement, il faut savoir s'arrêter...

Conclusion :

La prédiction de suites individuelles est un cadre méta-statistique où il s'agit d'agréger séquentiellement les prédictions de méthodes fondamentales, par exemple, celles de différentes méthodes statistiques.

J'espère avoir réussi aujourd'hui ma mission d'évangélisateur en vous montrant le potentiel et la flexibilité de ces techniques d'agrégation séquentielle...

Et si jamais vous avez d'autres idées de cadre applicatif...