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).

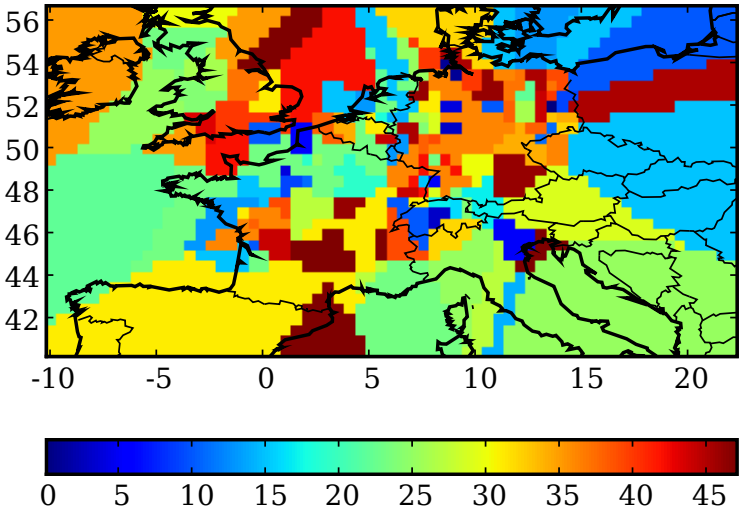
Les écarts sont mesurés en perte quadratique moyenne, ce qui revient à considérer la fonction de perte

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

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.



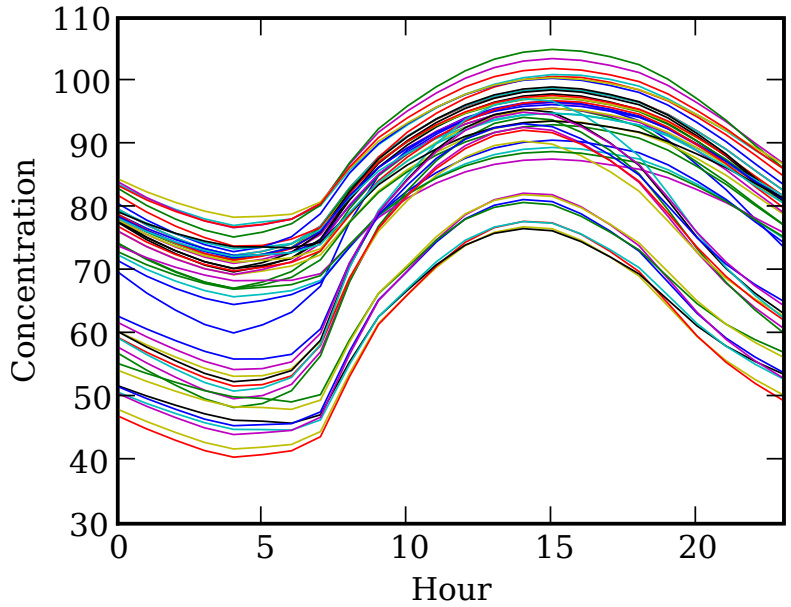


Figure: A gauche : Coloration de l'Europe en fonction de l'indice du meilleur expert local. A droite : Profils moyens de prédiction sur une journée (moyennes spatiales et temporelles, en $\mu g/m^3$).

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

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

et

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

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 ℓ^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., $r_n\big((1/48,\dots,1/48)\big)$,
- du meilleur modèle fondamental parmi $j = 1,\dots,48$,
- 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 ℓ^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.

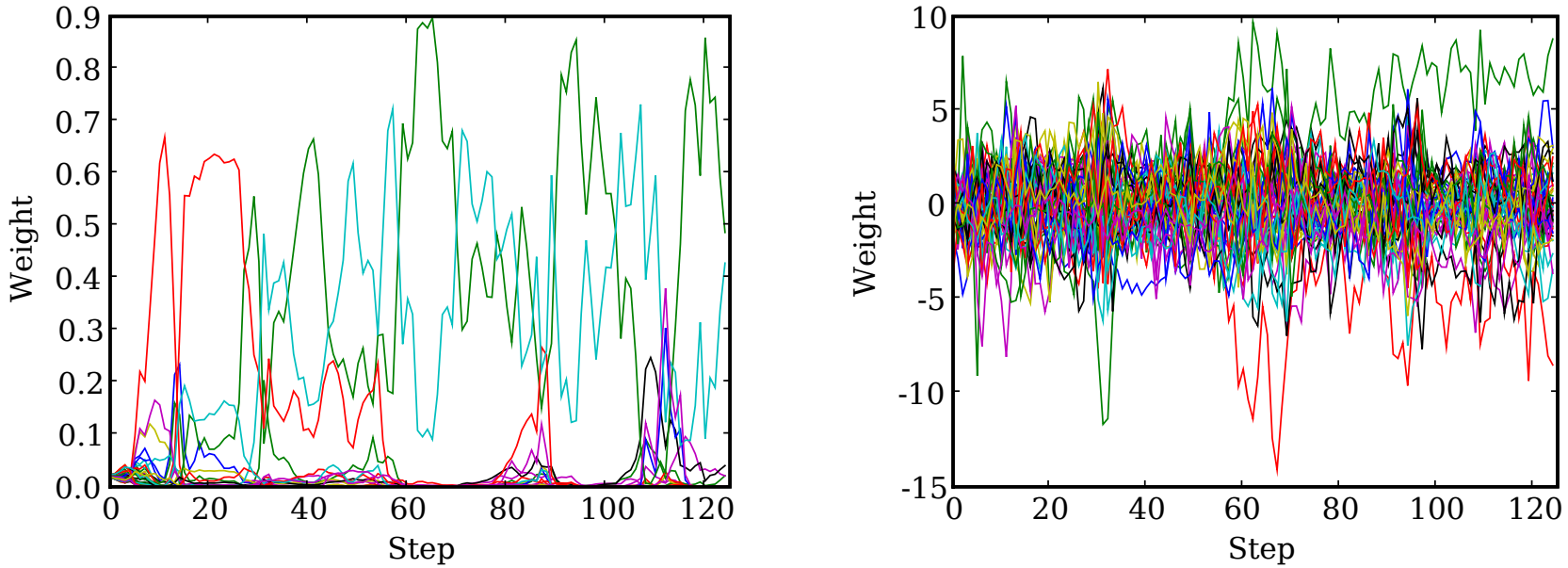


Figure: Poids produits au cours du temps par (à gauche) EG et la version escomptée de la régression ridge (à droite).

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.

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 (β_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 (β_s).

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

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 γ, 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.

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 η en fonction de t et considérer pour η_t le meilleur paramètre η 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 η	21.47
EG avec (η_t)	21.80
Meilleure linéaire	19.24
Ridge avec meilleur λ	20.77
Ridge avec (λ_t)	20.81

Le "meilleur" paramètre désigne le paramètre constant η ou λ, 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 ℓ^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