



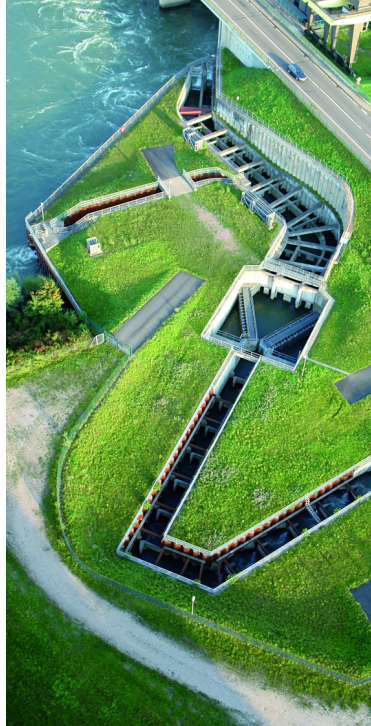
Agrégation séquentielle d'experts avec application à la prévision de consommation électrique.

Pierre Gaillard (EDF R&D, HEC Paris)

avec Yannig Goude (EDF R&D)

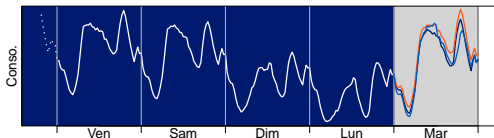
Gilles Stoltz (CNRS, HEC Paris)

Avril 2014 – Colloque Jeunes Probabilistes et Statisticiens



Objectif

Prévision à **court terme** (1 jour à l'avance) de la consommation électrique française



Parallèlement,

- EDF R&D développe de **nombreux modèles** de prévision
- Le paysage électrique français **évolue**



➔ remise en question des modèles historiques

Quel modèle utiliser?

On se propose de les **mélanger** séquentiellement.

Cadre – Prédiction séquentielle à l'aide d'experts

On dispose d'une suite bornée d'observations y_1, y_2, \dots à prévoir **instant par instant**

À chaque instant t

- un nombre fini K d'experts nous proposent des prévisions $x_{k,t}$
- on forme une prévision \hat{y}_t de la prochaine observation à partir
 - des **observations passées** y_1, \dots, y_{t-1}
 - des **prévisions actuelles et passées** des experts $(x_{k,s})$
- on observe y_t et on subit la perte $\hat{\ell}_t = \ell(\hat{y}_t, y_t)$

Notre but est de minimiser notre perte cumulée

$$\sum_{t=1}^T \hat{\ell}_t$$

Cadre – Prédiction séquentielle à l'aide d'experts

On dispose d'une suite bornée d'observations y_1, y_2, \dots à prévoir **instant par instant**

À chaque instant t

- un nombre fini K d'experts nous proposent des prévisions $x_{k,t}$
- on attribue un poids $\hat{p}_{k,t}$ à chacun des experts et on prévoit

$$\hat{y}_t = \sum_{k=1}^K \hat{p}_{k,t} x_{k,t}$$

- on observe y_t et on subit la perte $\hat{\ell}_t = \ell(\hat{y}_t, y_t)$

Notre but est de minimiser notre perte cumulée

$$\sum_{t=1}^T \hat{\ell}_t$$

Cadre – Prévision séquentielle à l'aide d'experts

On dispose d'une suite bornée d'observations y_1, y_2, \dots à prévoir **instant par instant**

À chaque instant t

- un nombre fini K d'experts nous proposent des prévisions $x_{k,t}$
- on attribue un poids $\hat{p}_{k,t}$ à chacun des experts et on prévoit

$$\hat{y}_t = \sum_{k=1}^K \hat{p}_{k,t} x_{k,t}$$

- on observe y_t et on subie la perte $\hat{\ell}_t = \ell(\hat{y}_t, y_t)$

Notre but est de minimiser notre perte cumulée $\sum_{t=1}^T \hat{\ell}_t$.

Exemple de fonction de perte $\ell(x, y)$

	Perte carrée	Perte absolue	Perte relative
$\ell(x, y) =$	$(x - y)^2$	$ x - y $	$\frac{ x - y }{y}$

Cadre – Prévision séquentielle à l'aide d'experts

On dispose d'une suite bornée d'observations y_1, y_2, \dots à prévoir **instant par instant**

À chaque instant t

- un nombre fini K d'experts nous proposent des prévisions $x_{k,t}$
- on attribue un poids $\hat{p}_{k,t}$ à chacun des experts et on prévoit

$$\hat{y}_t = \sum_{k=1}^K \hat{p}_{k,t} x_{k,t}$$

- on observe y_t et on subit la perte $\hat{\ell}_t = \ell(\hat{y}_t, y_t)$

Notre but est de minimiser notre perte cumulée $\sum_{t=1}^T \hat{\ell}_t$.

Remarques

- On ne fait **aucune hypothèse stochastique** sur la suite d'observations y_t .
- On observe les données et on prévoit **séquentiellement**.

Cadre – Prévision séquentielle à l'aide d'experts

On dispose d'une suite bornée d'observations y_1, y_2, \dots à prévoir **instant par instant**

À chaque instant t

- un nombre fini K d'experts nous proposent des prévisions $x_{k,t}$
- on attribue un poids $\hat{p}_{k,t}$ à chacun des experts et on prévoit

$$\hat{y}_t = \sum_{k=1}^K \hat{p}_{k,t} x_{k,t}$$

- on observe y_t et on subit la perte $\hat{\ell}_t = \ell(\hat{y}_t, y_t)$

Si tous les experts sont mauvais, on ne peut pas espérer avoir une faible erreur.

On évalue donc notre performance **relativement** à celles des experts

$$\underbrace{\sum_{t=1}^T \hat{\ell}_t}_{\text{Notre perte}} = \underbrace{\min_{k=1, \dots, K} \sum_{t=1}^T \ell(x_{k,t}, y_t)}_{\text{Perte du meilleur expert}} + \underbrace{R_T^{\text{expert}}}_{\text{Regret}}$$

Notre but est d'avoir un regret moyen R_T^{expert} / T qui tend vers 0 **quoi qu'il arrive**.

Cadre – Prévision séquentielle à l'aide d'experts

On dispose d'une suite bornée d'observations y_1, y_2, \dots à prévoir **instant par instant**

À chaque instant t

- un nombre fini K d'experts nous proposent des prévisions $x_{k,t}$
- on attribue un poids $\hat{p}_{k,t}$ à chacun des experts et on prévoit

$$\hat{y}_t = \sum_{k=1}^K \hat{p}_{k,t} x_{k,t}$$

- on observe y_t et on subit la perte $\hat{\ell}_t = \ell(\hat{y}_t, y_t)$

Si tous les experts sont mauvais, on ne peut pas espérer avoir une faible erreur.

On évalue donc notre performance **relativement** à celles des experts

$$\underbrace{\sum_{t=1}^T \hat{\ell}_t}_{\text{Notre perte}} = \underbrace{\min_{\mathbf{q} \in \Delta_K} \sum_{t=1}^T \ell(\mathbf{q} \cdot \mathbf{x}_t, y_t)}_{\text{Perte du meilleur mélange constant}} + \underbrace{R_T^{\text{convex}}}_{\text{Regret}}$$

Notre but est d'avoir un regret moyen R_T^{convex}/T qui tend vers 0 **quoi qu'il arrive**.

Meilleur expert \rightarrow meilleur mélange constant

Si la fonction de perte $\ell(\cdot, y)$ est convexe pour tout y , alors on peut se ramener au cadre linéaire suivant

À chaque instant t

- on attribue un poids $\hat{p}_{k,t}$ à chacun des experts
- la nature choisie un vecteur de pertes $\ell_t = (\ell_{1,t}, \dots, \ell_{k,t}) \in [0, 1]^K$
- chaque expert subit la perte $\ell_{k,t}$
- on subit la perte linéaire $\hat{\ell}_t = \hat{\mathbf{p}}_t \cdot \ell_t$ $\left(= \sum_{k=1}^K \hat{p}_{k,t} \ell_{k,t} \right)$

Notre but étant de minimiser notre perte cumulée

$$\underbrace{\sum_{t=1}^T \hat{\ell}_t}_{\text{Notre perte}} = \underbrace{\min_{k=1, \dots, K} \sum_{t=1}^T \ell_{k,t}}_{\text{Perte du meilleur expert}} + \underbrace{R_T}_{\text{Regret}}$$

On peut montrer que $R_T^{\text{expert}} \leq R_T^{\text{convex}} \leq R_T$ pour ℓ_t bien choisi.

Meilleur expert \rightarrow meilleur mélange constant

Si la fonction de perte $\ell(\cdot, y)$ est convexe pour tout y , alors on peut se ramener au cadre linéaire suivant

À chaque instant t

- on attribue un poids $\hat{p}_{k,t}$ à chacun des experts
- la nature choisie un vecteur de pertes $\ell_t = (\ell_{1,t}, \dots, \ell_{k,t}) \in [0, 1]^K$
- chaque expert subit la perte $\ell_{k,t}$
- on subit la perte linéaire $\hat{\ell}_t = \hat{\mathbf{p}}_t \cdot \ell_t$ $\left(= \sum_{k=1}^K \hat{p}_{k,t} \ell_{k,t} \right)$

De nombreuses stratégies permettent de contrôler notre perte cumulée

$$\underbrace{\sum_{t=1}^T \hat{\ell}_t}_{\text{Notre perte}} \leq \underbrace{\min_{k=1, \dots, K} \sum_{t=1}^T \ell_{k,t}}_{\text{Perte du meilleur expert}} + \underbrace{\square \sqrt{T \log K}}_{\text{Regret}}$$

On peut montrer que $R_T^{\text{expert}} \leq R_T^{\text{convex}} \leq R_T$ pour ℓ_t bien choisi.

Stratégie ML-prod

Paramètre: un vecteur $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$ de paramètres d'apprentissages

Initialisation: un vecteur $\widehat{\boldsymbol{w}}_0 = (\widehat{w}_{1,0}, \dots, \widehat{w}_{K,0})$ de poids initiaux

À chaque instant t

- on attribue à l'expert k le poids $\widehat{p}_{k,t} = \eta_k \widehat{w}_{k,t-1} / \boldsymbol{\eta} \cdot \widehat{\boldsymbol{w}}_{t-1}$
- on observe ℓ_t et on subit la perte $\widehat{\ell}_t = \widehat{\boldsymbol{p}}_t \cdot \boldsymbol{\ell}_t$
- on met à jour les poids

$$\widehat{w}_{k,t} = \widehat{w}_{k,t-1} \left(1 + \eta_k (\widehat{\ell}_t - \ell_{k,t}) \right)$$

Si $\eta_k \leq 1/2$ et $\ell_{k,t} \in [0, 1]$, notre erreur cumulée est alors majorée par

$$\underbrace{\sum_{t=1}^T \widehat{\ell}_t}_{\text{Notre perte}} \leq \min_{k=1, \dots, K} \left\{ \underbrace{\sum_{t=1}^T \ell_{k,t}}_{\text{Perte de l'expert } k} + \underbrace{\frac{1}{\eta_k} \log \frac{1}{w_{k,0}} + \eta_k \sum_{t=1}^T (\widehat{\ell}_t - \ell_{k,t})^2}_{\text{Regret}} \right\}$$

Stratégie ML-prod

Paramètre: un vecteur $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$ de paramètres d'apprentissages

Initialisation: un vecteur $\widehat{\mathbf{w}}_0 = (\widehat{\mathbf{w}}_{1,0}, \dots, \widehat{\mathbf{w}}_{K,0})$ de poids initiaux

À chaque instant t

- on attribue à l'expert k le poids $\widehat{\mathbf{p}}_{k,t} = \eta_k \widehat{\mathbf{w}}_{k,t-1} / \boldsymbol{\eta} \cdot \widehat{\mathbf{w}}_{t-1}$
- on observe ℓ_t et on subit la perte $\widehat{\ell}_t = \widehat{\mathbf{p}}_t \cdot \ell_t$
- on met à jour les poids

$$\widehat{\mathbf{w}}_{k,t} = \widehat{\mathbf{w}}_{k,t-1} \left(1 + \eta_k (\widehat{\ell}_t - \ell_{k,t}) \right)$$

Si on peut optimiser $\eta_k = \sqrt{\log(1/w_{k,0}) / \sum_t (\widehat{\ell}_t - \ell_{k,t})^2}$

$$\underbrace{\sum_{t=1}^T \widehat{\ell}_t}_{\text{Notre perte}} \leq \min_{k=1, \dots, K} \left\{ \underbrace{\sum_{t=1}^T \ell_{k,t}}_{\text{Perte de l'expert } k} + \underbrace{2 \sqrt{\log \frac{1}{w_{k,0}} \sum_{t=1}^T (\widehat{\ell}_t - \ell_{k,t})^2}}_{\text{Regret}} \right\}$$

Borne de regret de ML-prod – Pire des cas

Pour η_k bien choisis, la perte cumulée est contrôlée par

$$\sum_{t=1}^T \hat{\ell}_t \leq \min_{k=1, \dots, K} \left\{ \sum_{t=1}^T \ell_{k,t} + 2\sqrt{\log \frac{1}{w_{k,0}} \sum_{t=1}^T (\hat{\ell}_t - \ell_{k,t})^2} \right\}$$

En particulier, si on attribue des poids initiaux uniformes $w_{k,0} = 1/K$

$$\underbrace{\sum_{t=1}^T \hat{\ell}_t}_{\text{Notre perte}} \leq \underbrace{\min_{k=1, \dots, K} \sum_{t=1}^T \ell_{k,t}}_{\text{Perte du meilleur expert}} + \underbrace{2\sqrt{T \log K}}_{\text{Regret}}$$

Notre perte moyenne se rapproche de celle du meilleur expert à une vitesse $O(\sqrt{\log K/T})$.

Borne de regret de ML-prod – Faibles pertes

Si une stratégie vérifie la borne de regret

$$\sum_{t=1}^T \hat{\ell}_t \leq \min_{k=1, \dots, K} \left\{ \sum_{t=1}^T \ell_{k,t} + 2\sqrt{\log K \sum_{t=1}^T (\hat{\ell}_t - \ell_{k,t})^2} \right\}$$

Alors elle vérifie aussi

$$\sum_{t=1}^T \hat{\ell}_t \leq \min_{k=1, \dots, K} \left\{ \sum_{t=1}^T \ell_{k,t} + 2\sqrt{\log K \sum_{t=1}^T \ell_{k,t} + 16 \log K} \right\}$$

Le regret par rapport à l'expert k est d'autant plus faible que l'expert est bon !

Borne de regret de ML-prod – Pertes i.i.d.

Pour l'instant toutes nos bornes étaient déterministes (dans le pire des cas).
Que se passe-t-il si on suppose nos pertes i.i.d?

Hypothèse

Les vecteurs de pertes $\ell_t \in [0, 1]^K$ sont des variables aléatoires i.i.d telles qu'il existe un meilleur expert k^* et $\alpha > 0$ tel que

$$\forall t \geq 1 \quad \min_{k \neq k^*} \mathbb{E}[\ell_{k,t} - \ell_{k^*,t}] \geq \alpha$$

Si une stratégie vérifie

$$\sum_{t=1}^T \hat{\ell}_t \leq \sum_{t=1}^T \ell_{k^*,t} + 2\sqrt{\log K \sum_{t=1}^T (\hat{\ell}_t - \ell_{k^*,t})^2}$$

alors

- $\mathbb{E} \left[\sum_{t=1}^T \hat{\ell}_t \right] \leq \mathbb{E} \left[\sum_{t=1}^T \ell_{k^*,t} \right] + \frac{4 \log K}{\alpha}$

- Pour tout $\delta \in (0, 1)$, avec probabilité au moins $1 - \delta$

$$\sum_{t=1}^T \hat{\ell}_t \leq \sum_{t=1}^T \ell_{k^*,t} + \frac{4 \log K}{\alpha} + \frac{12}{\alpha} \sqrt{\log \frac{1}{\delta} \log K}$$

Experts spécialisés

À chaque instant t

- chaque expert donne une **mesure de confiance** en sa prévision $I_{k,t} \in [0, 1]$
- on attribue un poids $\hat{p}_{k,t}$ à chacun des experts
- la nature choisit un vecteur de pertes $\ell_t = (\ell_{1,t}, \dots, \ell_{K,t}) \in [0, 1]^K$
- chaque expert subit la perte $\ell_{k,t}$
- on subit la perte linéaire $\hat{\ell}_t = \hat{\mathbf{p}}_t \cdot \ell_t \quad \left(= \sum_{k=1}^K \hat{p}_{k,t} \ell_{k,t} \right)$

On souhaite minimiser notre **confiance regret** par rapport à chaque expert

$$R_{k,T} = \sum_{t=1}^T I_{k,t} (\hat{\ell}_t - \ell_{k,t})$$

Le cas particulier $I_{k,t} = 0$ exprime l'inactivité de l'expert k à l'instant t . Les jours où l'expert pense être meilleur sont plus pondérés.

Experts spécialisés

À chaque instant t

- chaque expert donne une **mesure de confiance** en sa prévision $I_{k,t} \in [0, 1]$
- on attribue un poids $\hat{p}_{k,t}$ à chacun des experts
- la nature choisit un vecteur de pertes $\ell_t = (\ell_{1,t}, \dots, \ell_{k,t}) \in [0, 1]^K$
- chaque expert subit la perte $\ell_{k,t}$
- on subit la perte linéaire $\hat{\ell}_t = \hat{\mathbf{p}}_t \cdot \ell_t \quad \left(= \sum_{k=1}^K \hat{p}_{k,t} \ell_{k,t} \right)$

Si on applique l'algorithme ML-prod avec les pseudo pertes

$$\tilde{\ell}_{k,t} = I_{k,t} \ell_{k,t} + (1 - I_{k,t}) \hat{\ell}_t,$$

alors

$$R_{k,T} = \sum_{t=1}^T I_{k,t} (\hat{\ell}_t - \ell_{k,t}) \leq 2 \sqrt{\log K \sum_{t=1}^T I_{k,t}^2}$$

Application – Le jeu de données

Il inclut 1696 jours du 1er janvier 2007 au 15 juin 2012 et contient

- la **consommation électrique** des clients EDF
- **De l'information exogène**
 - météo: température, nébulosité, vent
 - temporelle: date, tarif spécial
 - perte de clients

On retire les jours fériés ± 2 .

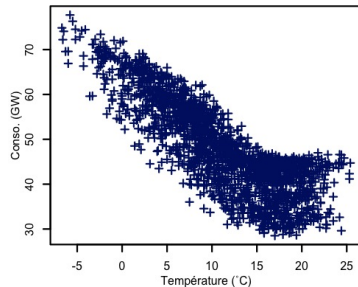
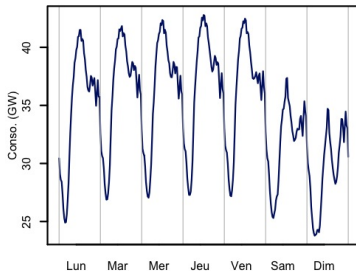
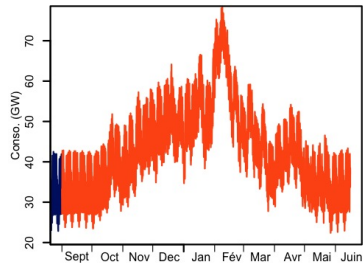
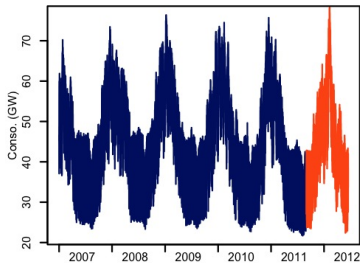
On divise le jeu de données en deux

- Jan. 2007 – Août. 2011: **ensemble d'entraînement** → construire les experts
- Sept. 2011 – Juin. 2012: **ensemble de test** → tester les experts et le mélange

On considère 3 experts construit à partir de trois modèles statistiques différents

- Régression régularisée sur une base de splines
- Régression linéaire sur un espace fonctionnel
- Clustering de données fonctionnelles à partir d'une base d'ondelettes

Les données en images...



Performances des experts et du mélange

On évalue les performances à l'aide du RMSE (root mean square error) sur l'année test. Plus c'est faible, mieux c'est !

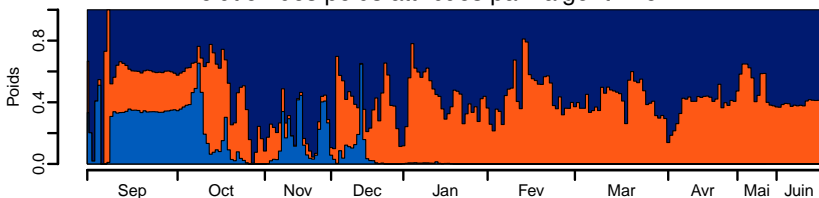
Method	RMSE (MW)
--------	-----------

Meilleur expert	744
-----------------	-----

Meilleur mélange constant	629
---------------------------	-----

ML-prod	626
---------	-----

Évolution des poids attribués par l'algorithme



Merci !