Contributions à l'agrégation séquentielle robuste d'experts

Travaux sur l'erreur d'approximation et la prévision en loi. Applications à la prévision pour les marchés de l'énergie

Thèse Cifre







Laboratoires de ratachement







Pierre Gaillard

Directeur de thèse : Gilles Stoltz

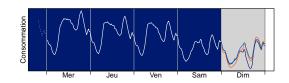
Responsable industriel: Yannig Goude

6 juillet 2015

Contexte industriel

000

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



C'est un enjeu majeur pour EDF car l'électricité se stocke mal.

L'équilibre entre l'offre et la demande doit être maintenu à tout moment pour éviter

- les risques physiques : coupure d'électricité, reconfiguration de réseau, blackout
- les risques financiers.

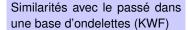
Production



Demande

De nombreux experts construisent des modèles

Modèles additifs généralisés (GAM)



Régression linéaire sur des courbes (CLR)



Introduction 00













Parallèlement, le paysage électrique français évolue



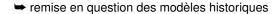












Quel modèle utiliser? On se propose de les mélanger séquentiellement.

Introduction

Agrégation séquentielle robuste d'experts

Cadre théorique
Application à la consommation électrique
Diminuer l'erreur d'approximation
Une méthode universelle!

La prévision d'incertitudes

Incertitudes et suites arbitraires
Application à la prévision de prix d'électricité

Opera

[Cesa-Bianchi et Lugosi, 2006]

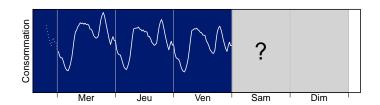


On dispose d'une série temporelle y_1, y_2, \ldots à prévoir instant par instant

À chaque instant t

- K experts nous proposent des prévisions x_{k,t}
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts et on prévoit

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

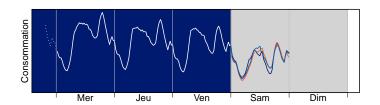


On dispose d'une série temporelle y_1, y_2, \ldots à prévoir instant par instant

À chaque instant t

- K experts nous proposent des prévisions x_{k,t}
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts et on prévoit

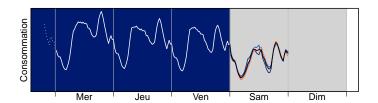
$$\widehat{y}_t = \sum_{k=1}^K \widehat{p}_{k,t} x_{k,t}$$



On dispose d'une série temporelle y_1, y_2, \ldots à prévoir instant par instant

À chaque instant t

- K experts nous proposent des prévisions x_{k,t}
- on forme une prévision \hat{y}_t de la prochaine observation à partir
 - des observations passées y₁,..., y_{t-1}
 - des prévisions actuelles et passées des experts $(x_{k,s})$ pour $s \leq t$
- on observe y_t ; on subit la perte $\hat{\ell}_t = \ell(\hat{y}_t, y_t)$; les experts subissent $\ell_{k,t} = \ell(x_{k,t}, y_t)$

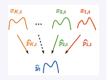


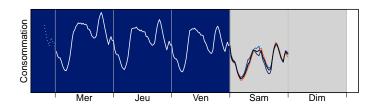
On dispose d'une série temporelle y_1, y_2, \ldots à prévoir instant par instant

À chaque instant t

- K experts nous proposent des prévisions x_{k,t}
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts et on prévoit

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



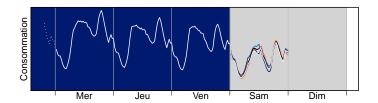


On dispose d'une série temporelle y_1, y_2, \ldots à prévoir instant par instant

À chaque instant t

- K experts nous proposent des prévisions X_{k,t}
- on attribue un poids $\hat{p}_{k,t}$ à chacun des experts et on prévoit

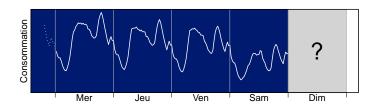
$$\widehat{y}_t = \sum_{k=1}^K \widehat{p}_{k,t} x_{k,t}$$



À chaque instant t

- K experts nous proposent des prévisions x_{k,t}
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts et on prévoit

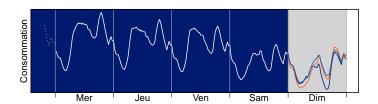
$$\widehat{y}_t = \sum_{k=1}^K \widehat{p}_{k,t} x_{k,t}$$



À chaque instant t

- K experts nous proposent des prévisions x_{k,t}
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts et on prévoit

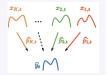
$$\widehat{y}_t = \sum_{k=1}^K \widehat{p}_{k,t} x_{k,t}$$

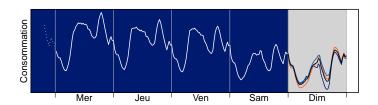


À chaque instant t

- K experts nous proposent des prévisions x_{k,t}
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts et on prévoit

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

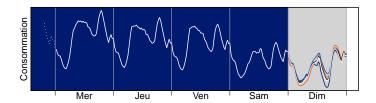




À chaque instant t

- K experts nous proposent des prévisions x_{k,t}
- on attribue un poids $\hat{p}_{k,t}$ à chacun des experts et on prévoit

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



On dispose d'une série temporelle y_1, y_2, \ldots à prévoir instant par instant

À chaque instant t

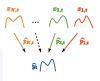
- K experts nous proposent des prévisions x_{k,t}
- on attribue un poids $\hat{p}_{k,t}$ à chacun des experts et on prévoit

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

• on observe y_t ; on subit la perte $\hat{\ell}_t = \ell(\hat{y}_t, y_t)$; les experts subissent $\ell_{k,t} = \ell(x_{k,t}, y_t)$

But : minimiser notre perte moyenne $\hat{L}_T = \frac{1}{T} \sum_{t=1}^{T} \hat{\ell}_t$

Perte	carrée	absolue	relative	quantile
$\ell(x,y) =$	$(x - y)^2$	x - y	$\frac{ x-y }{y}$	$(y-x)(\alpha-1_{\{y< x\}})$



Un cadre déterministe : aucune hypothèse stochastique sur la suite d'observations y_t .

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

Solution : on évalue donc notre performance relativement à celles des experts

$$\frac{1}{T}\sum_{t=1}^{T}\hat{\ell}_{t} = \min_{\substack{k=1,\ldots,K}} \frac{1}{T}\sum_{t=1}^{T}\ell_{k,t} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\hat{\ell}_{t} - \min_{\substack{k=1,\ldots,K}} \frac{1}{T}\sum_{t=1}^{T}\ell_{k,t}}_{\text{eff}}$$

$$\frac{\det}{\det} \hat{L}_{T}$$

$$\frac{\det}{\det} L_{T}^{\star}$$

$$\det L_{T}^{\star}$$

$$\det Reg_{T}$$

$$\text{regret moyen}$$

$$\text{performance}$$

$$\text{(erreur d'approximation)}$$

$$\text{(erreur d'estimation séquentielle)}$$

Objectif

Prévoir presque aussi bien que le meilleur expert sur le long terme :

$$\limsup_{T \to \infty} \left(\sup_{(y_t)_t, (x_{k-t})_{k-t}} \operatorname{Reg}_T \right) \leqslant 0$$

Meilleur mélange constant

[Cesa-Bianchi et Lugosi, 2006, Chapitre 2.5]

Une astuce permet de se comparer à une performance de référence plus ambitieuse

$$\min_{\mathbf{q} \in \Delta_K} \frac{1}{T} \sum_{t=1}^T \ell\left(\sum_{k=1}^K q_k x_{k,t}, y_t\right)$$

où
$$\Delta_K = \big\{ \mathbf{q} \in \mathbb{R}_+^K : \sum_{k=1}^K q_k = 1 \big\}.$$

Si un expert est mauvais mais forme des prévisions qui compensent la nôtre, on doit augmenter son poids.

→ le « gradient trick » formalise cette intuition en modifiant les pertes de l'algorithme

Exemple pour la perte carrée : $(x_{k,t} - y_t)^2 \rightarrow (\hat{y}_t - y_t)(x_{k,t} - y_t)$





Notre prévision

Expert k

Observation

Garantie de performance dans le pire des cas

[Cesa-Bianchi and Lugosi, 2006]

De nombreuses stratégies permettent d'obtenir la borne uniforme sur le regret moyen

$$\sup_{(x_k,t),(y_l)} \operatorname{Reg}_{T} \leqslant \Box \sqrt{(\log K)/T} \to 0 \quad \text{quand } T \to \infty$$

La vitesse uniforme $\mathcal{O}(1/\sqrt{T})$ est optimale.

Avantages:

- cadre séquentiel : permet de traiter potentiellement de gros volumes de données
- robustesse : les garanties sont vérifiées pour toutes suites d'observations et toutes suites de prévisions d'experts

Inconvénient : le pire des cas est rarement représentatif de la réalité.

vitesse trop lente! Peut-on parfois l'améliorer?

Le mélange par poids exponentiels (EWA)

[Vovk '90, Littlestone and Warmuth '94]

Paramètre : $\eta > 0$

Initialisation:
$$\hat{\mathbf{p}}_0 = (1/K, \dots, 1/K)$$

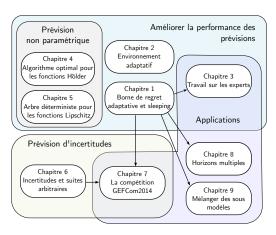
À chaque instant t, on attribue à l'expert k le poids

$$\hat{p}_{k,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{k,t}\right)}{\sum_{j=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,t}\right)}$$

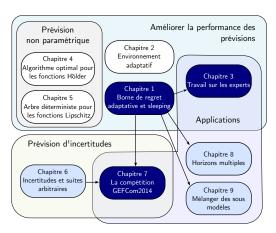
Performance : pour η bien choisi ($\approx \sqrt{(\log K)/T}$) est convexe et bornée, on a

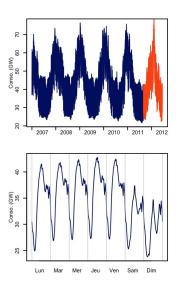
$$\mathsf{Reg}_T \stackrel{\mathrm{def}}{=} \frac{1}{T} \sum_{t=1}^T \hat{\ell}_t - \min_k \frac{1}{T} \sum_{t=1}^T \ell_{k,t} \leqslant \Box \sqrt{\frac{\log K}{T}} \to 0$$

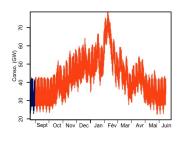
Dans cette thèse, nous avons traité divers problèmes de prévision de suites arbitraires

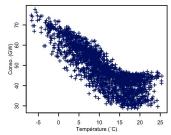


Dans cette thèse, nous avons traité divers problèmes de prévision de suites arbitraires









Jeu de données

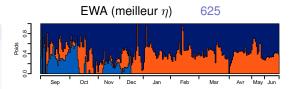
Consommation électrique + variables exogènes (calendaires, météorologiques)

Nombre de jours D	1 696 (de 2007 à 2012)				
	1 452 (entraînement) + 254 (test)				
Intervalles de temps	30 minutes				
Nombre d'experts K	3 (GAM, CLR, KWF)				
Unité	MW				
Médiane des y _t	43 496				
Borne B sur les y_t	78 922				

Performances

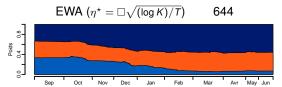
Critère

RMSE (root mean square error). Plus c'est faible, mieux c'est!



Oracles

Moyenne uniforme 725
Meilleur expert 744
Meilleur mélange constant 629



 \rightarrow II faut bien calibrer η . En pratique, on est rarement dans le pire des cas. Il faut tirer profit de la simplicité des données.

Notre solution: l'algorithme ML-Poly

[G., van Erven et Stoltz '14]

À chaque instant t,

on définit un paramètre d'apprentissage pour chaque expert k

$$\eta_{k,t} = \left(1 + \sum_{s=1}^{t-1} (\hat{\ell}_t - \ell_{k,t})^2\right)^{-1}$$

on attribue à l'expert k le poids

$$\hat{p}_{k,t} = \frac{\eta_{k,t} \left(\sum_{s=1}^{t-1} \left(\hat{\ell}_s - \ell_{k,s} \right) \right)_+}{\sum_{j=1}^{K} \eta_{j,t} \left(\sum_{s=1}^{t-1} \left(\hat{\ell}_s - \ell_{j,s} \right) \right)_+},$$

où
$$(\cdot)_+\stackrel{\mathrm{def}}{=} \max\{\cdot,0\}.$$

Garantie théorique

Si $\ell_{k,t} \in [0,1]$, notre erreur moyenne est alors majorée par

$$\underbrace{\frac{1}{T} \sum_{t=1}^{T} \hat{\ell}_{t}}_{\text{Notre perte}} \leqslant \min_{k=1,...,K} \left\{ \underbrace{\frac{1}{T} \sum_{t=1}^{T} \ell_{k,t}}_{\text{Perte de l'expert } k} + \underbrace{\frac{\square}{T} \sqrt{K \sum_{t=1}^{T} (\hat{\ell}_{t} - \ell_{k,t})^{2} (\log \log T)}}_{\text{Regret}} \right\}$$

Avantages: si un expert k^* est significativement meilleur que tous les autres, la convergence est en $\mathcal{O}((\log \log T)/T)$!

En effet, le poids de l'algorithme se concentre sur cet expert et au bout d'un certain temps

$$\hat{\ell}_t = \ell_{k,t}$$

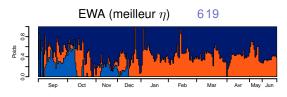
le regret n'augmente plus!

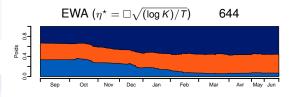
Critère

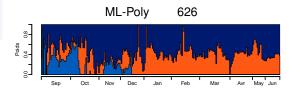
RMSE (root mean square error). Plus c'est faible, mieux c'est!

Oracles

Moyenne uniforme 725 Meilleur expert 744 Meilleur mélange constant 629







Travail sur le jeu d'experts

Notre performance est un compromis entre deux erreurs :

$$\frac{1}{T}\sum_{t=1}^{T}\hat{\ell}_{t} = \min_{\mathbf{q}\in\Delta_{K}}\frac{1}{T}\sum_{t=1}^{T}\ell(\mathbf{q}\cdot\mathbf{x}_{t},y_{t}) + \operatorname{Reg}_{T}$$
notre
performance de référence
cerreur d'approximation)
regret moyen
cerreur d'estimation séquentielle)

Diminuer l'erreur d'approximation en enrichissant le jeu d'experts :

- modification aléatoire d'un expert initial par méthode de bagging
- ajout de nouveaux experts cherchant à corriger les erreurs des experts déjà présents dans le mélange
- spécialisation d'experts à des événements météorologiques particuliers ;
- •

3 experts → 133 experts

Spécialisation d'experts

Idée

Concentrer l'apprentissage des experts sur certains scénarios

Scénarios météorologiques

- Fortes / faibles températures
- Fortes / faibles variations de températures

Autres scénarios

- Fortes / faibles consommations
- Hiver / été

De tels experts spécialisés ne prévoient que les jours correspondant à leurs scénarios.

Mélanger des experts spécialisés

Chaque jour, certains experts sont actifs et proposent des prévisions (selon leur spécialisation). Les autres experts n'en fournissent pas.

Si un expert k est inactif, on n'a pas accès à sa prévision.

Solution: supposer que les experts inactifs prévoient la même prévision que le mélange et résoudre le point fixe

$$\hat{y}_t = \sum_{j \text{ actif}} \hat{p}_{j,t} x_{j,t} + \sum_{k \text{ inactif}} \hat{p}_{k,t} \hat{y}_t$$

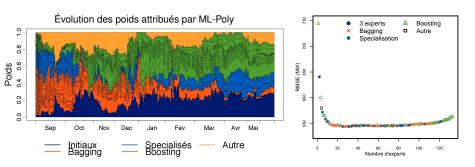
Le regret de ML-Poly n'augmente pas face aux experts inactifs :

$$\frac{1}{T} \sum_{t=1}^{T} \hat{\ell}_{t} \leq \min_{\kappa=1,...,\kappa} \left\{ \underbrace{\frac{1}{T} \sum_{t=1}^{T} \ell_{\kappa,t}}_{\text{Perte de l'expert } \kappa} + \underbrace{\frac{\square}{T} \sqrt{K \sum_{t=1}^{T} (\hat{\ell}_{t} - \ell_{\kappa,t})^{2}} (\log \log T)} \right\}$$

Perspective : optimalité de ce type de borne

Performances

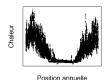
	Meilleur expert	Meilleur mélange	EWA (η^{\star})	EWA (meilleur η)	ML-Poly
3 experts	744	629	644	625	626
133 experts	744	521	737	591	565



Perspectives : création automatique de nombreux experts. Algorithmes parcimonieux.

Universalité – D'autres jeux de données réels

Chaleur (centrale de cogénération)



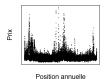


Température



Prix de l'électricité

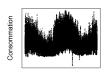
Ε̈́





Consommation

Consommation américaine

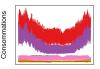




Position annuelle

Température

Groupes de consommateurs





Position annuelle Température

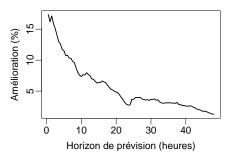
Défi : les données individuelles

Thersame Thorizon de prevision

Sur un jeu de données similaire, on a

- construit 5 experts pouvant prévoir à différents horizons de temps
- adapté les algorithmes de mélange à la prévision multi-horizon

Gain de performance (du RMSE) de ML-Poly par rapport au meilleur expert



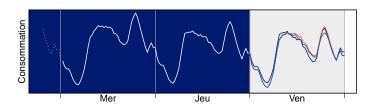
Sommaire

La prévision d'incertitudes

Incertitudes et suites arbitraires Application à la prévision de prix d'électricité

La prévision d'incertitudes

Peut-on aller plus loin et prévoir des lois à partir des prévisions des experts?



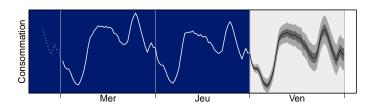
Motivations:

- planification de stratégies en fonction des scénarios possibles
- gestion des risques
- énergies renouvelables

Note interne edf: Méthodes de prévision en loi: état de l'art. Dordonnat, Collet, Oppenheim et Dacunha-Castelle 2010

La prévision d'incertitudes

Peut-on aller plus loin et prévoir des lois à partir des prévisions des experts?

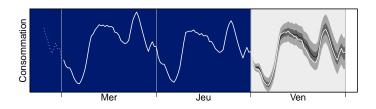


Motivations:

- planification de stratégies en fonction des scénarios possibles
- gestion des risques
- énergies renouvelables

Note interne edf: Méthodes de prévision en loi: état de l'art. Dordonnat, Collet, Oppenheim et Dacunha-Castelle 2010

Peut-on aller plus loin et prévoir des lois à partir des prévisions des experts?



Motivations:

- planification de stratégies en fonction des scénarios possibles
- gestion des risques
- énergies renouvelables

Note interne edf: Méthodes de prévision en loi: état de l'art. Dordonnat, Collet, Oppenheim et Dacunha-Castelle 2010

Des incertitudes dans un monde déterministe?

Intuition : si les prévisions des experts sont

- significativement différentes → risque accru
- semblables → risque plus faible

Pas facile de trouver un critère de performance convainquant!

S'inspirer des méthodes de la littérature stochastique :

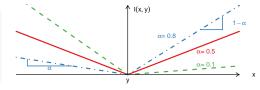
- minimiser la log-vraissemblance → minimiser la perte logarithmique : déjà très étudié
- compromis niveau-largeur pour des intervalles de confiance : impossible ou trivial!
- minimiser la perte quantile

La perte quantile

[Koenker et Bassett, 1978]

Si $\alpha \in (0,1)$ est le quantile à prévoir, celle-ci est définie par

$$\ell_{\alpha} \colon (x,y) \mapsto (y{-}x) \big(\alpha {-} \mathbb{1}_{\{y < x\}}\big)$$



Lemma

Si Y est une variable aléatoire réelle. Alors le quantile q_{α} vérifie

$$q_{\alpha} \in \operatorname*{arg\,min}_{x \in \mathbb{R}} \mathbb{E} \big[\ell_{\alpha}(Y, x) \big]$$

En particulier, le quantile empirique \hat{q}_{α} d'ordre α de $\{y_1, \dots, y_T\}$ vérifie

$$\hat{q}_{\alpha} \in \operatorname*{arg\,min}_{x \in \mathbb{R}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \ell_{\alpha}(y_{t}, x) \right\}$$

Mélange + perte quantile ℓ_{α} + experts = fonctions constantes :

$$\frac{1}{T}\sum_{t=1}^{T}\ell_{\alpha}(y_{t},\hat{y}_{\alpha,t}) \leqslant \frac{1}{T}\sum_{t=1}^{T}\ell_{\alpha}(y_{t},\hat{q}_{\alpha}) + \Box\sqrt{1/T}$$

Perte du quantile empirique

La prévision d'incertitudes

où $\hat{y}_{\alpha,t}$ est la prévision du mélange au temps t.

Si les y_t sont i.i.d de loi "régulière" : $\frac{1}{T}\sum_{t=1}^T \hat{y}_{\alpha,t} \approx \hat{q}_\alpha \to q_\alpha$ quand $T \to \infty$.

Problème : les données ne sont pas i.i.d! On s'aide donc des experts, en supposant les quantiles des y_t bien représentés par des combinaisons linéaires des experts \mathbf{x}_t .

Le mélange vérifie alors :

$$\frac{1}{T}\sum_{t=1}^{T}\ell_{\alpha}(y_{t},\hat{y}_{\alpha,t}) \leqslant \min_{\mathbf{p}\in\Delta_{K}} \frac{1}{T}\sum_{t=1}^{T}\ell_{\alpha}(y_{t},\mathbf{p}\cdot\mathbf{x}_{t}) + \Box\sqrt{\frac{\log K}{T}}$$

Application: la compétition GEFCom2014 (prix)

[Hong, Pinson, Fan et Zareipour, '15]

[G., Goude et Nedellec, '15]

But : prévision probabiliste du prix de l'électricité.

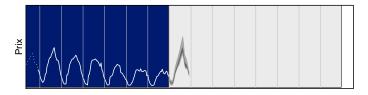
- 15 semaines de compétition (Août → Dec. 2014) : 3 essais, 12 officielles
- Chaque semaine
 - les données passées sont révélées
 - on prévoit les quantiles 0.01, ..., 0.99 des prix de la prochaine journée
 - on est évalué par la perte quantile moyenne



Application: la compétition GEFCom2014 (prix)

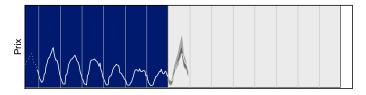
But : prévision probabiliste du prix de l'électricité.

- 15 semaines de compétition (Août → Dec. 2014) : 3 essais, 12 officielles
- Chaque semaine
 - les données passées sont révélées
 - on prévoit les quantiles 0.01,..., 0.99 des prix de la prochaine journée
 - on est évalué par la perte quantile moyenne



But : prévision probabiliste du prix de l'électricité.

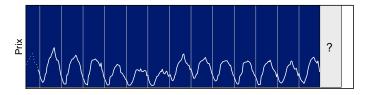
- 15 semaines de compétition (Août → Dec. 2014) : 3 essais, 12 officielles
- Chaque semaine
 - les données passées sont révélées
 - on prévoit les quantiles 0.01,..., 0.99 des prix de la prochaine journée
 - on est évalué par la perte quantile moyenne



Application: la compétition GEFCom2014 (prix)

But : prévision probabiliste du prix de l'électricité.

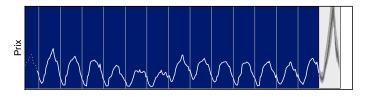
- 15 semaines de compétition (Août → Dec. 2014) : 3 essais, 12 officielles
- Chaque semaine
 - les données passées sont révélées
 - on prévoit les quantiles 0.01, ..., 0.99 des prix de la prochaine journée
 - on est évalué par la perte quantile moyenne



Application: la compétition GEFCom2014 (prix)

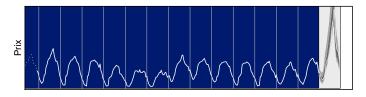
But : prévision probabiliste du prix de l'électricité.

- 15 semaines de compétition (Août → Dec. 2014) : 3 essais, 12 officielles
- Chaque semaine
 - les données passées sont révélées
 - on prévoit les quantiles 0.01, ..., 0.99 des prix de la prochaine journée
 - on est évalué par la perte quantile moyenne

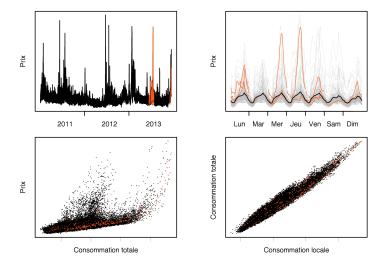


But : prévision probabiliste du prix de l'électricité.

- 15 semaines de compétition (Août → Dec. 2014) : 3 essais, 12 officielles
- Chaque semaine
 - les données passées sont révélées
 - on prévoit les quantiles 0.01, ..., 0.99 des prix de la prochaine journée
 - on est évalué par la perte quantile moyenne



Les données en images



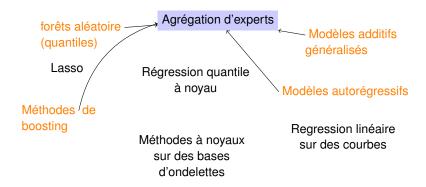
Les méthodes de prévision utilisées

Nouveau dans le domaine, nous avons essayé beaucoup de méthodes différentes !

Agrégation d'experts forêts aléatoire Modèles additifs (quantiles) généralisés Lasso Régression quantile à noyau Modèles autorégressifs Méthodes de boosting Regression linéaire Méthodes à noyaux sur des courbes sur des bases d'ondelettes

Toutes ces méthodes avec différentes variables exogènes, calibrations,...

Nouveau dans le domaine, nous avons essayé beaucoup de méthodes différentes!

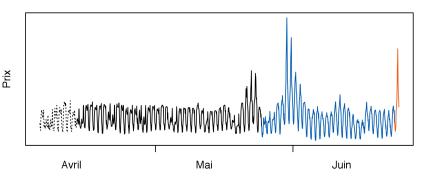


Toutes ces méthodes avec différentes variables exogènes, calibrations,...

Méthode par agrégation d'experts

Protocole:

- on entraîne des experts en s'inspirant des méthodes de la litérature : ensemble d'entraînement
- on mélange les experts avec la perte quantile : ensemble de mélange
- on prévoit les quantiles à l'aide des poids du mélange : jour d'évaluation



12 individual forecasters (or experts)

[Weron, 2014]

Notations: prix P_t , consommations TL_t et ZL_t , jour D_t , prix maximal P_t^{\max} , température lissée $TL_t^{(0.8)}$

Modèles autoregressifs issus de la littérature :

• AR:
$$\log(P_t) = \alpha_1 \log(P_{t-24}) + \alpha_2 \log(P_{t-48}) + \alpha_3 \log(P_{t-168}) + \alpha_4 \log(P_{t-24}) + h(D_t) + \varepsilon_t$$
,

$$\begin{aligned} \bullet \ \ \mathsf{ARX} : \log(P_t) &= \alpha_1 \log(P_{t-24}) + \alpha_2 \log(P_{t-48}) + \alpha_3 \log(P_{t-168}) \\ &+ \alpha_4 \log(P_{t-24}^{\min}) + \alpha_5 \log(\mathsf{TL}_t) + \alpha_6 \log(\mathsf{ZL}_t) + h(\mathsf{D}_t) + \varepsilon_t \,. \end{aligned}$$

- TAR et TARX : AR et ARX avec changements de régimes
- PAR et PARX : AR et ARX avec prétraitement des données

Nouveaux modèles :

• IM1:

$$\log(P_{t}) = \alpha_{1} \log(P_{t-24}) + \alpha_{2} \log(P_{t-48}) + \alpha_{3} \log(P_{t-168}) + \alpha_{4} \log(P_{t}^{\max}) + \alpha_{5} Z L_{t}^{(0.95)} + \alpha_{6} Z L_{t}^{(0.95)} + \alpha_{7} Z L_{t}^{(0.8)} + \alpha_{8} Z L_{t}^{(0.8)} + h(D_{t}) + \varepsilon_{t}$$

- LM2: $\log(P_t) = \alpha_1 \log(P_t^{\text{last}}) + \alpha_2 \log(P_{t-24}) + \alpha_3 \log(P_t^{\text{max}}) + \alpha_4 Z L_t + \alpha_5 Z L_t + h(D_t) + \varepsilon_t.$
- GAM : un modèle additif généralisé
- RF1 and RF2 : deux forêts aléatoires et GBM : une méthode de boosting

12 individual forecasters (or experts)

veron, 2014]

Notations: prix P_t , consommations TL_t et ZL_t , jour D_t , prix maximal $P_t^{\mathbf{m.x}}$, température lissée $TL_t^{(0.8)}$

Modèles autoregressifs issus de la littérature :

• AR:
$$\log(P_t) = \alpha_1 \log(P_{t-24}) + \alpha_2 \log(P_{t-48}) + \gamma_5 \log(P_{t-168}) + \alpha_4 \log(P_{t-24}^{\min}) + h(D_t) + \varepsilon_t$$

• ARX:
$$\log(P_t) = \alpha_1 \log(P_{t-24}) + \alpha_2 \log(P_{t-34}) + \alpha_3 \log(P_{t-168}) + \alpha_4 \log(P_{t-24}^{\min}) + \alpha_5 \log(T_{L_t}) + \alpha_6 \log(Z_{L_t}) + h(D_t) + \varepsilon_t$$
.

- TAR et TARX : AR et ARX avec changements de régimes
- PAR et PARX : AR et A Tix a rec prétraitement des données

Nouveaux modèles :

• LM1:

$$\begin{split} \log(P_{t}) &= \alpha_{1} \log(P_{t-24}) + \alpha_{2} \log(P_{t-48}) + \alpha_{3} \log(P_{t-168}) + \alpha_{4} \log(\frac{P_{t}^{\max}}{t}) + \alpha_{5} Z L_{t}^{(0.95)} \\ &+ \alpha_{6} T L_{t}^{(0.95)} + \alpha_{7} Z L_{t}^{(0.8)} + \alpha_{8} T L_{t}^{(0.8)} + h(D_{t}) + \varepsilon_{t} \end{split}$$

• UM2 :

$$\log(P_t) = \alpha_1 \log(P_t^{\mathrm{last}}) + \alpha_2 \log(P_{t-24}) + \alpha_3 \log(P_t^{\mathrm{max}}) + \alpha_4 Z L_t + \alpha_5 Z L_t + h(D_t) + \varepsilon_t.$$

- GAM : un modèle additif généralisé
- RF1 and RF2 : deux forêts aléatoires et GBM : une méthode de boosting

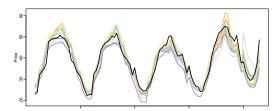
Les prévisions des experts

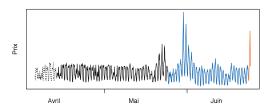
Les 12 experts ne prévoient que des prévisions ponctuelles du prix attendu. MAPE (en %) des experts évalués du 16 juin au 17 décembre 2013.

AR	ARX	PAR	PARX	TAR	TARX
11.30	10.50	11.30	10.40	11.30	10.50
1 1 1 1	LMO	C 4 8 4	DE4	DEO	ODM
LIVII	LIVIZ	GAIVI	RF1	RF2	GBIVI

Moyenne	Meilleur	
uniforme	mélange	
9.17	8.85	

La prévision d'incertitudes 0000000000



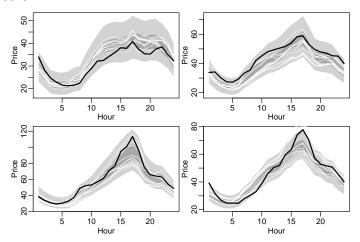


- « Gradient trick »: combinaison convexes
- Astuce de Kivinen et Warmuth, '97 : combinaisons linéaires Idée : ajouter des versions négatives des experts
- Échantillonage des données pour aggrandir l'ensemble de mélange
- On moyenne les poids à la fin

$$\mathbf{p} = \frac{1}{T} \sum_{t=1}^{T} \hat{p}_{k,t}$$

Résultats (Semaines 1 à 4)

Plutôt bons!



Perspectives : création automatique des experts. Construction séquentielle des experts pour entraîner le mélange sur tout le jeu de données.

Sommaire

Opera

Opera

Package R implémentant les différents algorithmes de mélange.

```
# EWA avec eta = 1 avec la perte carrée sans le ''gradient trick''
mod = mixture(y=Y, experts=X, aggregationRule=list(name="EWA", eta=1,
loss.type='squareloss', loss.gradient=FALSE), awake=awake)

# EWA avec calibration automatique de eta sur une grille
mod = mixture(y=Y, experts=X, aggregationRule="EWA", awake=awake)

# ML-Poly
mod = mixture(y=Y, experts=X, aggregationRule="MLpol", awake=awake)

# Fixed—share: adaptativité aux ruptures
mod = mixture(y=Y, experts=X, aggregationRule="FS", awake=awake)
```

Perspectives : à publier sur le CRAN!

Merci!

Introduction

Agrégation séquentielle robuste d'experts

Cadre théorique Application à la consommation électrique Diminuer l'erreur d'approximation Une méthode universelle!

La prévision d'incertitudes

Incertitudes et suites arbitraires Application à la prévision de prix d'électricité

Opera