Robust online aggregation of ensemble (demand) forecasts

Gilles Stoltz

CNRS — HEC Paris

The framework of this talk

Sequential and worst-case deterministic prediction of time series based on ensemble forecasts

A time series $y_1, y_2, \ldots \in \mathbb{R}^d$ is to be predicted

Ensemble forecasts are available, e.g., given by some stochastic or machine-learning models (for us: black boxes)

At each instance t, forecasting black-box $j \in \{1, \ldots, N\}$ outputs

$$f_{j,t} \equiv f_{j,t}\bigl(y_1^{t-1}\bigr)$$

Observations and predictions are made in a sequential fashion:

The prediction $\widehat{y}_t$ of y_t is determined based

- on the past observations $y_1^{t-1} = (y_1, \ldots, y_{t-1})$,
- and the current and past ensemble forecasts $f_{j,s}$, where $s \in \{1, \ldots, t\}$ and $j \in \{1, \ldots, N\}$

Typical solution: convex (or linear) combinations of the ensemble forecasts, with adaptive weights $\mathbf{p}_t = \bigl(p_{1,t}, \, \ldots, \, p_{N,t}\bigr)$

Aggregated forecasts: $$\widehat{y}_t = \sum_{j=1}^{N} p_{j,t} \, f_{j,t}$$

The observations y_t will not be considered stochastic anymore at this stage; thus the performance criterion will be a relative one

Given a convex loss function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, e.g., the square loss $\ell(x,y) = \|x-y\|^2$:

The cumulative losses of the statistician and of the constant convex combinations $\mathbf{q} = (q_1, \ldots, q_N)$ of the forecasts equal

$$\widehat{L}_T = \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t} \, f_{j,t}, \, y_t\right) \quad \text{and} \quad L_T(\mathbf{q}) = \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j \, f_{j,t}, \, y_t\right)$$

The regret R_T is defined as the difference

$$\widehat{L}_T - \min_{\mathbf{q}} L_T(\mathbf{q}) = \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\ y_t\right) - \min_{\mathbf{q}} \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\ y_t\right)$$

We are interested in aggregation rules with (uniformly) vanishing per-round regret,

$$\limsup_{T\to\infty} \quad \frac{1}{T} \sup \left\{ \widehat{L}_T - \min_{\mathbf{q}} L_T(\mathbf{q}) \right\} \leqslant 0$$

The supremum is over all possible sequences of observations and of ensemble forecasts (not just over most of these sequences!)

Remarks:

- Hence the name “prediction of individual sequences” (or robust aggregation of ensemble forecasts)
- The best convex combination $\mathbf{q}^\star$ is known in hindsight whereas the statistician has to predict in a sequential fashion

This framework leads to a meta-statistical interpretation:

- ensemble forecasts are given by some statistical forecasting methods, each possibly tuned with a different given set of parameters
- these ensemble forecasts relying on some stochastic model are then combined in a robust and deterministic manner

The cumulative loss of the statistician can be decomposed as

$$\widehat{L}_T = \min_{\mathbf{q}} L_T(\mathbf{q}) + R_T$$

In words:

cumulative loss $=$ approximation error $+$ sequential estimation error

Disclaimer

We could also consider batch learning methods to aggregate forecasts, like

- BMA (Bayesian model averaging),
- CART (classification and regression trees),
- random forests, etc.,

or even selection methods, and apply them online, by running a batch analysis at each step

→ We instead resort to “real” online techniques that, in addition, come up with theoretical guarantees even in non-stochastic scenarios

First study

Forecasting of the electricity load

Data source: EDF R&D

Authors: Pierre Gaillard and Yannig Goude

Reference: Proceedings of WIPFOR '2013

Some characteristics of one among the studied data sets:

- January 1, 2008 – August 31, 2011 as a training data set
- September 1, 2011 – June 15, 2012 (excluding some special days) as testing set

- Electricity demand for EDF clients, at a half-hour step

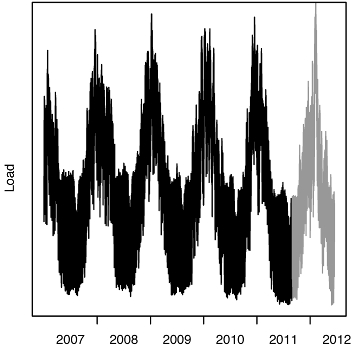
- Typical values: median $= 43\,496$ MW
  maximum $= 78\,922$ MW

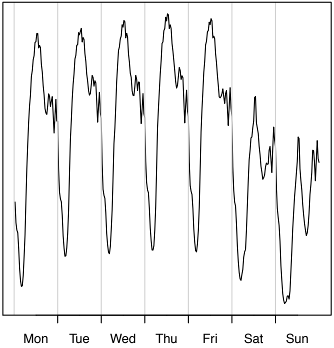
- Three forecasters: GAM, CLR, KWF

→ Instead of trusting only one model/base forecaster ("selection"), we proceed in a more greedy way and consider ensemble forecasts, which we combine sequentially ("aggregation")

This leads to more accurate and more stable (meta-)predictions

Data looks like...





Convex loss functions considered:

- square loss $\ell(x, y) = (x - y)^2$ → RMSE
- absolute percentage of error $\ell(x, y) = |x - y|/|y|$ → MAPE

Operational constraint:

One-day ahead prediction at a half-hour step, i.e., 48 aggregated forecasts

Ensemble forecasters:

- GAM / generalized additive models
  (see Wood 2006; Wood, Goude, Shaw 2014)
- CLR / curve linear regression
  (see Cho, Goude, Brossat, Yao 2013, 2014)
- KWF / functional wavelet-kernel approach
  (see Antoniadis, Paparoditis, Sapatinas 2006; Antoniadis, Brossat, Cugliari, Poggi 2012, 2013)

RMSE and MAPE on the testing set (with no warm-up period):

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(y_t - \widehat{y}_t\right)^2} \qquad \text{and} \qquad \frac{1}{T}\sum_{t=1}^{T}\left|y_t - \widehat{y}_t\right|$$

How good are our building blocks? See the “oracles” below

	Uniform mean	Best forecaster	Best convex **p**	Best linear **u**
RMSE (MW)	725	744	629	629
MAPE (%)	1.18	1.29	1.06	1.06

In this article the focus is to create more base forecasting methods and to improve the oracles (and in turn the performance of the aggregation methods) accordingly

A strategy to pick convex weights

Let’s do some maths!

Reminder of the aim and setting:

Given a loss function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

Choose sequentially the convex weights $p_{j,t}$

To uniformly bound the regret with respect to all sequences of observations y_t and ensemble forecasts $f_{j,t}$:

$$\sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\ y_t\right) - \min_{\mathbf{q}} \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\ y_t\right)$$

When ℓ is convex and differentiable in its first argument:

For all $x, y \in \mathbb{R}^d$,

$$\forall x' \in \mathbb{R}^d, \qquad \ell(x,y) - \ell(x',y) \leqslant \nabla \ell(x,y) \cdot (x - x')$$

Assumption OK for RMSE, MAE, MAPE, etc.

To uniformly bound the regret with respect to all convex weight vectors $\mathbf{q}$, we write

$$\max_{\mathbf{q}} \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\, y_t\right) - \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\, y_t\right)$$

$$\leqslant \quad \max_{\mathbf{q}} \sum_{t=1}^{T} \nabla\ell\left(\sum_{k=1}^{N} p_{k,t} f_{k,t},\, y_t\right) \cdot \left(\sum_{j=1}^{N} p_{j,t} f_{j,t} - \sum_{j=1}^{N} q_j f_{j,t}\right)$$

$$= \quad \max_{\mathbf{q}} \sum_{t=1}^{T} \left(\sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \sum_{j=1}^{N} q_j \widetilde{\ell}_{j,t}\right)$$

$$= \quad \sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \widetilde{\ell}_{i,t}$$

where we denoted

$$\widetilde{\ell}_{j,t} = \nabla\ell\left(\sum_{k=1}^{N} p_{k,t} f_{k,t},\, y_t\right) \cdot f_{j,t}$$

Considering the (signed) pseudo-losses $\widetilde{\ell}_{j,t} = \nabla \ell \left(\sum_{k=1}^{N} p_{k,t} f_{k,t},\; y_t \right) \cdot f_{j,t}$

the regret is smaller than $\displaystyle \sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \widetilde{\ell}_{i,t}$

Exponentially weighted averages [also called AFTER]:

$p_{j,1} = 1/N$ then

$$p_{j,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{k,s}\right)}$$

ensure that if all $\widetilde{\ell}_{j,t} \in [m, M]$, then

$$\sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \widetilde{\ell}_{i,t} \leqslant \frac{\ln N}{\eta} + \eta \frac{(M-m)^2}{8} T$$

References: Vovk '90; Littlestone and Warmuth '94

Proof by mere calculus

Hoeffding's lemma: for all convex weights $(p_1, \ldots, p_N)$ and all numbers $u_1, \ldots, u_N$ with range $[b, B]$,

$$\ln \sum_{j=1}^{N} p_j e^{u_j} \leqslant \frac{(B-b)^2}{8} + \sum_{j=1}^{N} p_j u_j$$

For all $t = 1, 2, \ldots,$

$$\begin{aligned} -\eta \sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} &= -\eta \sum_{j=1}^{N} \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{k,s}\right)} \widetilde{\ell}_{j,t} \\ &\geqslant \ln \frac{\sum_{j=1}^{N} \exp\left(-\eta \sum_{s=1}^{t} \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{k,s}\right)} - \frac{\eta^2}{8}(M-m)^2 \end{aligned}$$

A telescoping sum appears and leads to

$$\sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} \leqslant \underbrace{-\frac{1}{\eta} \ln \frac{\sum_{j=1}^{N} \exp\left(-\eta \sum_{s=1}^{T} \widetilde{\ell}_{j,s}\right)}{N}}_{\leqslant \min_{i=1,\ldots,N} \sum_{t=1}^{T} \widetilde{\ell}_{i,t} + \frac{\ln N}{\eta}} + \eta \frac{(M-m)^2}{8} T$$

Obtained regret bound optimized over η:

$$R_T \leqslant \min_{\eta>0} \left\{ \frac{\ln N}{\eta} + \eta \frac{(M-m)^2}{8} T \right\} = (M-m)\sqrt{\frac{T}{2}\ln N}$$

for the (theoretical) optimal choice

$$\eta^\star = \frac{1}{M-m}\sqrt{\frac{8\ln N}{T}}$$

Issue: the parameters T and $[m, M]$ not always known beforehand

Solutions:

- "doubling trick"
- adaptive learning rates η_t, picked according to some theoretical formulas

However, these theoretically satisfactory solutions would not work well in practice!

This is what we do instead. (It is very different from techniques like cross-validation: we exploit the sequential fashion.)

The exponentially weighted average strategy $\mathcal{E}_\eta$ with fixed learning rate η picks

$$p_{j,t}(\eta) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{k,s}\right)}$$

We denote its cumulative loss $\widehat{L}_t(\eta) = \displaystyle\sum_{s=1}^{t} \ell\left(\sum_{j=1}^{N} p_{j,s}(\eta) f_{j,s},\, y_s\right)$

Based on the family of the $\mathcal{E}_\eta$, we build a data-driven meta-strategy which at each instance $t \geqslant 2$ resorts to

$$\mathbf{p}_{t+1}(\eta_t) \qquad \text{where} \qquad \eta_t \in \mathop{\arg\min}_{\eta > 0} \widehat{L}_t(\eta)$$

Reference: An idea of Vivien Mallet

Other natural variants: Focus on the most recent losses

Moving sums (with window of size H)

$$p_{j,t} = \frac{\exp\left(-\eta \sum_{s=\max\{1,t-H\}}^{t-1} \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=\max\{1,t-H\}}^{t-1} \widetilde{\ell}_{k,s}\right)}$$

Regret is $\geqslant \Box\, T$ in the worst case

Discounted losses (with discounts given by a sequence $\beta_t \searrow 0$)

$$p_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} (1+\beta_{t-s})\widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} (1+\beta_{t-s})\widetilde{\ell}_{k,s}\right)}$$

Sublinear regret bounds hold for suitable sequences (β_t) and (η_t):

$$t\eta_t \longrightarrow 0 \qquad \text{and} \qquad \eta_t \sum_{s \leqslant t} \beta_s \longrightarrow 0$$

(We often take $\beta_s = \Box/s^2$ in the experimental studies.)

A strategy to pick linear weights

You all know it in a stochastic setting!

Linear combinations: Ridge regression (and the LASSO?)

Ridge regression — introduced in the 70s by Hoerl and Kennard:

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

It also exhibits a sublinear regret against individual sequences: for all $y_t \in [-B, B]$ and $f_{j,t} \in [-B, B]$, for all $\mathbf{u} \in \mathbb{R}^N$

$$\sum_{t=1}^{T} \left(y_t - \sum_{j=1}^{N} v_{j,t} \, f_{j,t} \right)^2 - \sum_{t=1}^{T} \left(y_t - \sum_{j=1}^{N} u_j \, f_{j,t} \right)^2 \leqslant \lambda \|\mathbf{u}\|_2^2 + 2NB^2 \left(1 + \frac{NTB^2}{\lambda} \right) \ln \left(1 + \frac{TB^2}{N\lambda} \right)$$

References: Vovk '01; Azoury and Warmuth '01; Gerchinovitz '11

The bound can be $O\big(\sqrt{T} \ln T\big)$ with λ of the order of $1/\sqrt{T}$

Same comments as before when T is unknown

We do not know any such regret bounds for the LASSO (yet?)

These methods can compensate for biases in either direction (the weights do not need to sum up to 1)

Can/should even be used as a pre-treatment on each single forecaster:

- turn it into a forecaster with predictions $\gamma_t \, f_{j,t}$
- performing on average almost as well as the best forecaster of the form $\gamma \, f_{j,t}$ for some constant $\gamma \in \mathbb{R}$

This would improve greatly the predictions if there existed, for instance, an almost constant multiplicative bias of $1/\gamma$

First study, continued

Prediction of electricity load

Benchmark and oracles

(RMSE of the ensemble forecasts and of fixed combinations thereof)

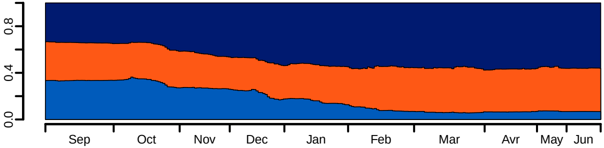
Uniform mean	Best forecaster	Best convex $\mathbf{p}$	Best linear $\mathbf{u}$
725	744	629	629

vs.

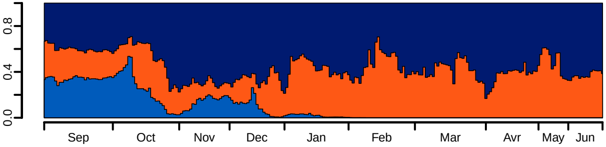
Aggregated forecasts with convex weights
(No discount considered)

Exp. weights (best η for theory)	644
Exp. weights (best η on data)	619
Exp. weights (η_t tuned on data)	625
ML-Poly (tuned according to theory)	626

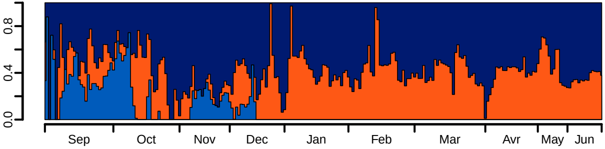
No focus on a single member! (See also the numerical performance.)



← Exp. weights (theory)



← Exp. weights (best η)



← ML-Poly (theory)

Weights change quickly and significantly over time and do not converge

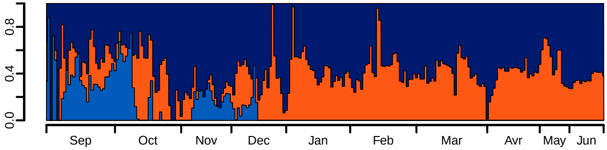
(illustrates that the performance of forecasters varies over time)

Are all forecasters useful? ... Definitely yes!

3 forecasters → only best 2

ML-Poly	626	→	646
Exp. weights	625	→	644

Forecasters not considered anymore can come back to life if needed



← ML-Poly

Benchmark and oracles

(RMSE of the ensemble forecasts and of fixed combinations thereof)

Uniform mean	Best forecaster	Best convex $\mathbf{p}$	Best linear $\mathbf{u}$
725	744	629	629

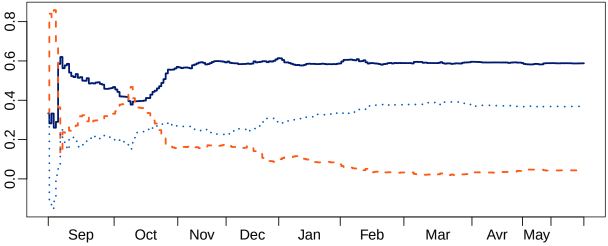
vs.

Aggregated forecasts with linear weights

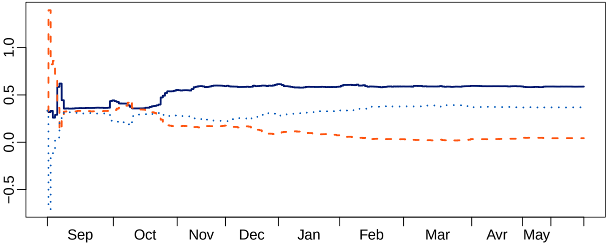
(No discount considered)

Ridge (best λ on data)	636
Ridge (λ_t tuned on data)	638
Ridge pre-treatment	744 → 745
on the forecasters	887 → 873
(tuned on data)	1 287 → 1 293

Weight vectors chosen by ridge regression



← Best λ



← λ_t on data

This was only a small glimpse into the work performed by Pierre Gaillard, Yannig Goude, and others, at EDF R&D

Other data sets studied include the forecasting of

- Slovakian demand for clients of an EDF subbranch
- GEFCom '2014 electricity price
- GEFCom '2014 electricity load
- Heat load of an Ukrainian co-generation plant
- Electricity demand of sub-groups of EDF clients

$\longrightarrow$ Universality of the aggregation methods!

Reference: Pierre Gaillard's Ph.D. dissertation, July 2015

Methodological summary

Methodological summary

1. Build the N base forecasters, possibly on a training data set, and pick another data set for the evaluation, with T instances
2. Compute some benchmarks and some reference oracles
3. Evaluate our strategies when run with fixed parameters (i.e., with the best parameters in hindsight)
4. The performance of interest is actually the one of the data-driven meta-strategies

We typically expect $T \geqslant 5N$ or even $T \geqslant 10N$

Hope arises when the oracles are 10% or 20% better than the methods used so far (e.g., the best ensemble forecast when the latter is known in advance)

This usually requires the ensemble forecasters to be as different as possible!

Second study

Forecasting of exchange rates

Data source: IMF / Fed

Authors: Christophe Amat, Tomasz Michalski, Gilles Stoltz

Reference: SSRN–2448655, April 2015

Predict 1-month ahead monthly averages r_{t+1} of exchange rates

Based on 4×2 macro-economic indicators describing the state of each country in month t:

- inflation rates (Infl)
- short-term interest rates (IR)
- changes in monetary mass (Mon)
- changes in industrial production (Prod)
- changes in interest rates (IR.Diff)

Difficult to improve on the no-change (NC) prediction, i.e., on forecasting r_{t+1} by r_t

Reference: Meese and Rogoff '83

Some (limited) results as well for end-of-month values

Convex or linear combinations of the 5 × 2 macro-economic indicators for countries A and B:

$$\ln \widehat{r}_{t+1} - \ln r_t = \sum_{j=1}^{5} \bigl(u^A_{j,t+1} x^A_{j,t} - u^B_{j,t+1} x^B_{j,t} \bigr)$$

Evaluation through RMSE with a short training period of $t_0 = 30$:

$$\sqrt{\frac{1}{T - t_0 + 1} \sum_{t=t_0}^{T} \bigl(\ln \widehat{r}_t - \ln r_t \bigr)^2}$$

Data-driven meta-strategies based on discounted versions of:

- Exponential weights (no gradient) ← interpretable weights
- Ridge regression ← pushes in favor of no-change

Some orders of magnitude for the prediction problems at hand are indicated below.

Time intervals	Every month
Period	March 1973 – December 2014
Time instances T	about 500
Size N of ensemble	$5\ (= 1 + 4)$
USD / GBP	
Median of the Δ_t	1.48×10^{-2}
Maximum of the $\lvert\Delta_t\rvert$	11.08×10^{-2}
JPY / USD	
Median of the Δ_t	1.57×10^{-2}
Maximum of the $\lvert\Delta_t\rvert$	10.52×10^{-2}

Results for USD / GBP

Pairs of indicators	RMSE	Oracle	RMSE
NC	2.4410×10^{-2}	Best member	2.4400×10^{-2}
Infl	2.4561×10^{-2}	Best convex	2.4315×10^{-2}
IR	2.4620×10^{-2}	Best linear	2.3453×10^{-2}
Mon	2.5037×10^{-2}		
Prod	2.4390×10^{-2}		
IR.Diff	2.4400×10^{-2}		

vs.

Rolling OLS	2.5960×10^{-2}	(worse!)	
Exp. weights	2.3777×10^{-2}	(-2.51%)	→ P-value: 3.3%
Ridge	2.3512×10^{-2}	(-3.68%)	→ P-value: 2.7%

Results for JPY / USD

Pairs of indicators	RMSE	Oracle	RMSE
NC	2.7042×10^{-2}	Best member	2.7003×10^{-2}
Infl	2.7003×10^{-2}	Best convex	2.6751×10^{-2}
IR	2.7203×10^{-2}	Best linear	2.6411×10^{-2}
Mon	2.7551×10^{-2}		
Prod	2.7406×10^{-2}		
IR.Diff	2.7038×10^{-2}		

vs.

Rolling OLS	2.8189×10^{-2}	(worse!)	
Exp. weights	2.6125×10^{-2}	(-3.39%)	→ P-value: 0.5%
Ridge	2.6031×10^{-2}	(-3.74%)	→ P-value: 0.2%

Quantile prediction

Uncertainty measures in this deterministic setting

Pinball loss: $\ell_\alpha : (x,y) \mapsto (y-x)\big(\alpha - \mathbb{I}_{\{y<x\}}\big)$

Quantile of order α of the law of Y as a minimizer:

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

$\longrightarrow$ Substitute $\ell(x,y) = (x-y)^2$ or $\ell(x,y) = |x-y|/|y|$ with ℓ_α to predict an α–quantile $\widehat{y}_t^\alpha$ for y_t

I.e., control a per-round regret of the form

$$\frac{1}{T}\sum_{t=1}^{T} \ell_\alpha\big(\widehat{y}_t^\alpha,\, y_t\big) - \frac{1}{T}\min_x \sum_{t=1}^{T} \ell_\alpha\big(x,\, y_t\big)$$

The $\widehat{y}_t^\alpha$ are based on forecasts $f_{j,t}$ of central tendencies or of α–quantiles

Strategy: exponential weights

(+ trick from Kivinen and Warmuth '97 to compete with linear combinations)

Does it work well?

Ask Pierre Gaillard, Yannig Goude, Raphaël Nedellec:

Winners of the two GEFCom'2014 competitions (demand + price)

Other empirical studies

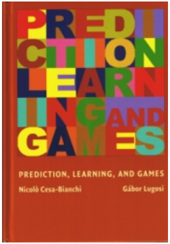
- Forecasting of air quality (INRIA and INERIS)
- Forecasting of the production data of oil reservoirs (IFP–EN)

→ Universality, versatility and efficiency!

But time is over...

Reference for theory

The so-called “red bible!”


PREDICTION LEARNING AND GAMES
PREDICTION, LEARNING, AND GAMES
Nicolò Cesa-Bianchi
Gábor Lugosi

Prediction, Learning, and Games

Nicolò Cesa-Bianchi and Gábor Lugosi