Big Data for Economics # 8

A. Charpentier (Université de Rennes 1)

https://github.com/freakonometrics/ub

UB School of Economics Summer School, 2018.

#8 New Tools for Time Series & Forecasting

Forecasts and Predictions

Mathematical statistics is based on inference and testing, using probabilistic properties.

If we can reproduce past observations, it is supposed to proved good predictions.

Why not consider a collection of scenarios likely to occur on a given time horizon (drawing fro, a (predictive) probability distribution)

The closer forecast $\widehat{y}_t$ is to observed y_t, the better the model, either according to ℓ_1-norm - with $|\widehat{y}_t - y_t|$ - or to the ℓ_2-norm - with $(\widehat{y}_t - y_t)^2$.

If this is an interesting information about central tendency, it cannot be used to anticipate extremal events.

Forecasts and Predictions

More formally, we try to compare two very different objects : a function (the predictive probability distribution) and a real value number (the observed value).

Natural idea : introduce a score, as in Good (1952, Rational Decisions) or Winkler (1969, Scoring Rules and the Evaluation of Probability Assessors), used in meteorology by Murphy & Winkler (1987, A General Framework for Forecast Verification).

Let F denote the predictive distribution, expressing the uncertainty attributed to future values, conditional on the available information.

Probabilistic Forecasts

Notion of probabilistic forecasts, Gneiting & Raftery (2007 Strictly Proper Scoring Rules, Prediction, and Estimation).

In a general setting, we want to predict value taken by random variable Y.

Let F denote a cumulative distribution function.

Let $\mathcal{A}$ denote the information available when forecast is made.

F is the ideal forecast) for Y given $\mathcal{A}$ if the law of $Y|\mathcal{A}$ has distribution F.

Suppose F continuous. Set $Z_F = F(Y)$, the probability integral transform of Y.

F is probabilistically calibrated if $Z_F \sim \mathcal{U}([0,1])$

F is marginally calibrated if $\mathbb{E}[F(y)] = \mathbb{P}[Y \leq y]$ for any $y \in \mathbb{R}$.

Probabilistic Forecasts

Observe that for a ideal forecast, $F(y) = \mathbb{P}[Y \leq y|\mathcal{A}]$, then

- $\mathbb{E}[F(y)] = \mathbb{E}[\mathbb{P}[Y \leq y|\mathcal{A}]] = \mathbb{P}[Y \leq y]$

This forecast is est marginally calibrated

- $\mathbb{P}[Z_F \leq z] = \mathbb{E}[\mathbb{P}[Z_F \leq z|\mathcal{A}]] = z$

This forecast is probabilistically calibrated

Suppose $\mu \sim \mathcal{N}(0,1)$. And that ideal forecast is $Y|\mu \sim \mathcal{N}(\mu,1)$.

E.g. if $Y_t \sim \mathcal{N}(0,1)$ and $Y_{t+1} = y_t + \varepsilon_t \sim \mathcal{N}(y_t,1)$.

One can consider $F = \mathcal{N}(0,2)$ as naïve forecast. This distribution is marginally calibrated, probabilistically calibrated and ideal.

One can consider F a mixture $\mathcal{N}(\mu,2)$ and $\mathcal{N}(\mu \pm 1,2)$ where "± 1" means $+1$ or -1 probability $1/2$, hesitating forecast. This distribution is probabilistically calibrated, but not marginally calibrated.

Probabilistic Forecasts

Indeed $\mathbb{P}[F(Y)\leq u]=u$,

$$\mathbb{P}[F(Y)\leq u]=\frac{\mathbb{P}[\Phi(Y)\leq u]+\mathbb{P}[\Phi(Y+1)\leq u]}{2}+\frac{\mathbb{P}[\Phi(Y)\leq u]+\mathbb{P}[\Phi(Y-1)\leq u]}{2}$$

One can consider $F=\mathcal{N}(-\mu,1)$. This distribution is marginally calibrated, but not probabilistically calibrated.

In practice, we have a sequence (Y_t,F_t) of pairs, $(\boldsymbol{Y},\boldsymbol{F})$.

The set of forecasts $\boldsymbol{F}$ is said to be performant if for all t, predictive distributions F_t are precise (sharpness) and well-calibrated.

Precision is related to the concentration of the predictive density around a central value (uncertainty degree).

Calibration is related to the coherence between predictive distribution F_t and observations y_t.

Probabilistic Forecasts

Calibration is poor if 80%-confidence intervals (implied from predictive distributions, i.e. $\big[F_t^{-1}(\alpha), F_t^{-1}(1-\alpha)\big]$) do not contain y_t's about 8 times out of 10.

To test marginal calibration, compare the empirical cumulative distribution function

$$\widehat{G}(y) = \lim_{n\rightarrow\infty} \frac{1}{n}\mathbf{1}_{Y_t \leq y}$$

and the average of predictive distributios

$$\overline{F}(y) = \lim_{n\rightarrow\infty} \frac{1}{n}\sum_{t=1}^{n} F_t(y)$$

To test probabilistic calibration, test if sample $\{F_t(Y_t)\}$ has a uniform distribution (PIT approach, see Dawid (1984, Present Position and Potential Developments: The Prequential Approach)).

Probabilistic Forecasts

One can also consider a score $S(F,y)$ for all distribution F and all observation y.

The score is said to be proper if

$$\forall F,G,\mathbb{E}[S(G,Y)]\leq\mathbb{E}[S(F,Y)] \text{ where } Y\sim G.$$

In practice, this expected value is approximated using $\displaystyle\frac{1}{n}\sum_{t=1}^n S(F_t,Y_t)$

One classical rule is the logarithmic score $S(F,y)=-\log[F'(y)]$ if F is (abs.) continuous.

Another classical rule is the continuous ranked probability score (CRPS, see Hersbach (2000, Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems))

$$S(F,y)=\int_{-\infty}^{+\infty}(F(x)-\mathbf{1}_{x\geq y})^2dx=\int_{-\infty}^{y}F(x)^2+\int_{y}^{+\infty}(F(x)-1)^2dx$$

Probabilistic Forecasts

with empirical version

$$\widehat{S}=\frac{1}{n}\sum_{t=1}^{n}S(F_t,y_t)=\frac{1}{n}\sum_{t=1}^{n}\int_{-\infty}^{+\infty}(F_t(x)-\mathbf{1}_{x\geq y_t})^2dx$$

studied in Murphy (1970, The ranked probability score and the probability score: a comparison).

This rule is proper since

$$\mathbb{E}[S(F,Y)]=\int_{-\infty}^{\infty}\mathbb{E}\big[F(x)-\mathbf{1}_{x\geq Y}\big]^2dx$$

$$=\int_{-\infty}^{\infty}\big[[F(x)-G(x)]^2+G(x)[1-G(x)]\big]^2dx$$

is minimal when $F=G$.

Probabilistic Forecasts

If F corresponds to the $\mathcal{N}(\mu,\sigma^2)$ distribution

$$S(F,y)=\sigma\left[\frac{y-\mu}{\sigma}\left(2\Phi\left(\frac{y-\mu}{\sigma}\right)-1\right)+2\frac{y-\mu}{\sigma}-\frac{1}{\sqrt{\pi}}\right]$$

Observe that

$$S(F,y)=\mathbb{E}\big|X-y\big|-\frac{1}{2}\mathbb{E}\big|X-X'\big| \text{ où } X,X'\sim F$$

(where X and X' are independent versions), cf Gneiting & Raftery (2007, Strictly Proper Scoring Rules, Prediction, and Estimation.

If we use for F the empirical cumulative distribution function $\widehat{F}_n(y)=\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{y_i\leq y}$ then

$$S(\widehat{F}_n,y)=\frac{2}{n}\sum_{i=1}^n(y_{i:n}-y)\left(\mathbf{1}_{y_{i:n}\leq y}-\frac{i-1/2}{n}\right)$$

Probabilistic Forecasts

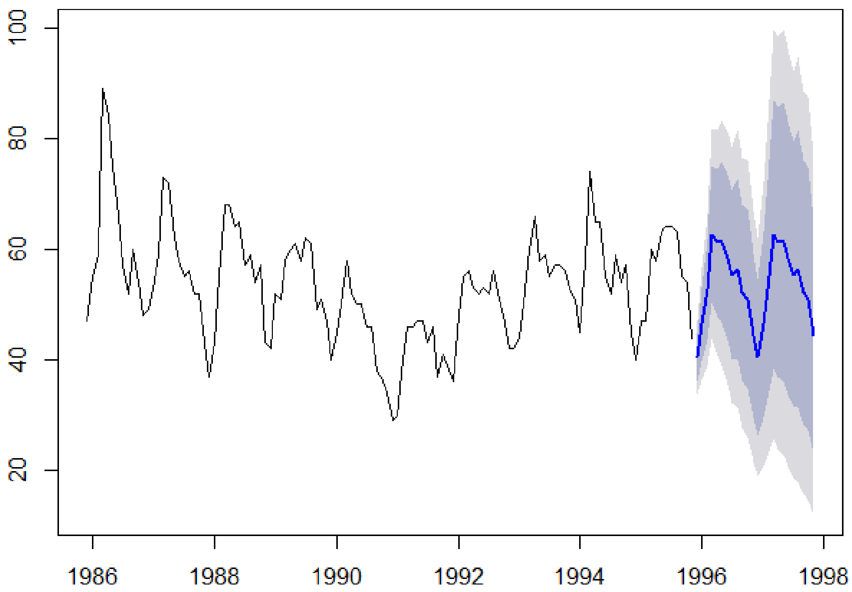
Consider a Gaussian $AR(p)$ time series,

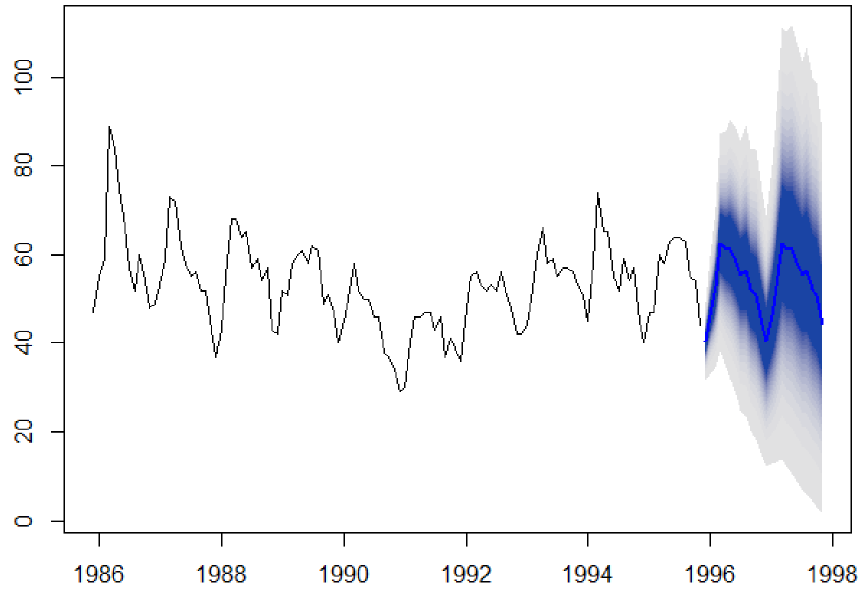
$$Y_t = c + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t, \text{ with } \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

then forecast with horizon 1 yields

$$F_t \sim \mathcal{N}({}_{t-1}\widehat{Y}_t, \sigma^2)$$

where ${}_{t-1}\widehat{Y}_t = c + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p}$.





Probabilistic Forecasts

Suppose that Y can be explained by covariates $\boldsymbol{x} = (x_1, \cdots, x_m)$. Consider some kernel based conditional density estimation

$$\widehat{p}(y|\boldsymbol{x}) = \frac{\widehat{p}(y,\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})} = \frac{\sum_{i=1}^n K_h(y-y_i)K_h(\boldsymbol{x}-\boldsymbol{x}_i)}{\sum_{i=1}^n K_h(\boldsymbol{x}-\boldsymbol{x}_i)}$$

In the case of a linear model, there exists $\boldsymbol{\theta}$ such that $\widehat{p}(y|\boldsymbol{x}) = \widehat{p}(y|\boldsymbol{\theta}^\top\boldsymbol{x})$, and

$$\widehat{p}(y|\boldsymbol{\theta}^\top\boldsymbol{x}=s) = \frac{\sum_{i=1}^n K_h(y-y_i)K_h(s-\boldsymbol{\theta}^\top\boldsymbol{x}_i)}{\sum_{i=1}^n K_h(s-\boldsymbol{\theta}^\top\boldsymbol{x}_i)}$$

Parameter $\boldsymbol{\theta}$ can be estimated using a proxy of the log-likelihood

$$\widehat{\boldsymbol{\theta}} = \text{argmax}\left\{\sum_{i=1}^n \log \widehat{p}(y_i|\boldsymbol{\theta}^\top\boldsymbol{x}_i)\right\}$$

#9 Additional Technical Stuff

≪ Confidence Bands

Also called variability bands for functions in Härdle (1990) Applied Nonparametric Regression.

From Collomb (1979) Condition nécessaires et suffisantes de convergence uniforme d'un estimateur de la régression, with Kernel regression (Nadarayah-Watson)

$$\sup\left\{|m(x)-\widehat{m}_h(x)|\right\}\sim C_1h^2+C_2\sqrt{\frac{\log n}{nh}}$$

$$\sup\left\{|m(\boldsymbol{x})-\widehat{m}_h(\boldsymbol{x})|\right\}\sim C_1h^2+C_2\sqrt{\frac{\log n}{nh^{\dim(\boldsymbol{x})}}}$$

« Confidence Bands

So far, we have mainly discussed pointwise convergence with

$$\sqrt{nh}\,(\widehat{m}_h(x) - m(x)) \xrightarrow{\mathcal{L}} \mathcal{N}(\mu_x, \sigma_x^2).$$

This asymptotic normality can be used to derive (pointwise) confidence intervals

$$\mathbb{P}(IC^-(x) \le m(x) \le IC^+(x)) = 1 - \alpha \;\; \forall x \in \mathcal{X}.$$

But we can also seek uniform convergence properties. We want to derive functions $IC^{\pm}$ such that

$$\mathbb{P}(IC^-(x) \le m(x) \le IC^+(x) \;\; \forall x \in \mathcal{X}) = 1 - \alpha.$$

≪ Confidence Bands

- Bonferroni's correction

Use a standard Gaussian (pointwise) confidence interval

$$IC^{\pm}_{\star}(x) = \widehat{m}(x) \pm \sqrt{nh}\widehat{\sigma}t_{1-\alpha/2}.$$

and take also into accound the regularity of m. Set

$$V(\eta) = \frac{1}{2}\left(\frac{2\eta+1}{n} + \frac{1}{n}\right)\|m'\|_{\infty,x}, \text{ for some } 0 < \eta < 1$$

where $\|\varphi'\|_{\infty,x}$ is on a neighborhood of x. Then consider

$$IC^{\pm}(x) = IC^{\pm}_{\star}(x) \pm V(\eta).$$

≪ Confidence Bands

- Use of Gaussian processes

Observe that $\sqrt{nh}\,(\widehat{m}_h(x) - m(x)) \xrightarrow{\mathcal{D}} G_x$ for some Gaussian process (G_x).
Confidence bands are derived from quantiles of $\sup\{G_x, x \in \mathcal{X}\}$.

If we use kernel k for smoothing, Johnston (1982) Probabilities of Maximal Deviations for Nonparametric Regression Function Estimates proved that

$$G_x = \int k(x-t)dW_t, \text{ for some standard } (W_t) \text{ Wiener process}$$

is then a Gaussian process with variance $\int k(x)k(t-x)dt$. And

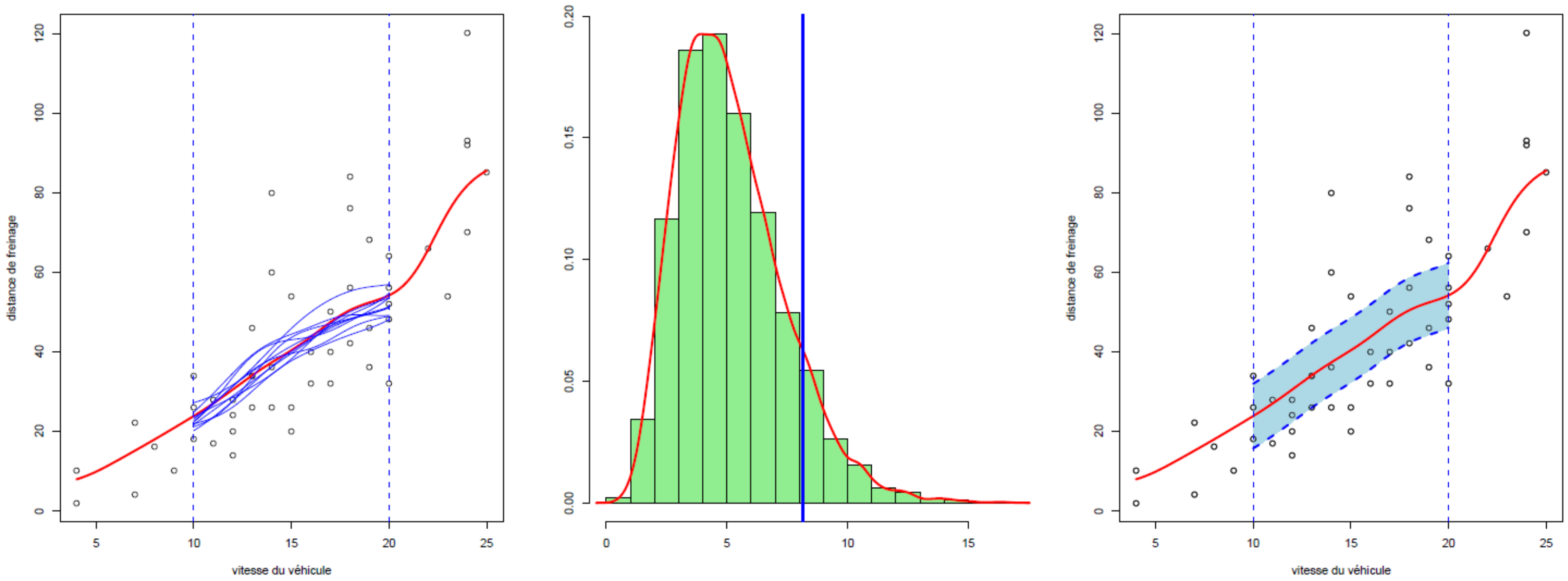
$$IC^{\pm}(x) = \widehat{\varphi}(x) \pm \left(\frac{q_\alpha}{\sqrt{2\log(1/h)}} + d_n\right)\frac{5}{7}\frac{\widehat{\sigma}^2}{\sqrt{nh}}$$

with $d_n = \sqrt{2\log h^{-1}} + \dfrac{1}{\sqrt{2\log h^{-1}}}\log\sqrt{\dfrac{3}{4\pi^2}}$, where $\exp(-2\exp(-q_\alpha)) = 1-\alpha$.

≪ Confidence Bands

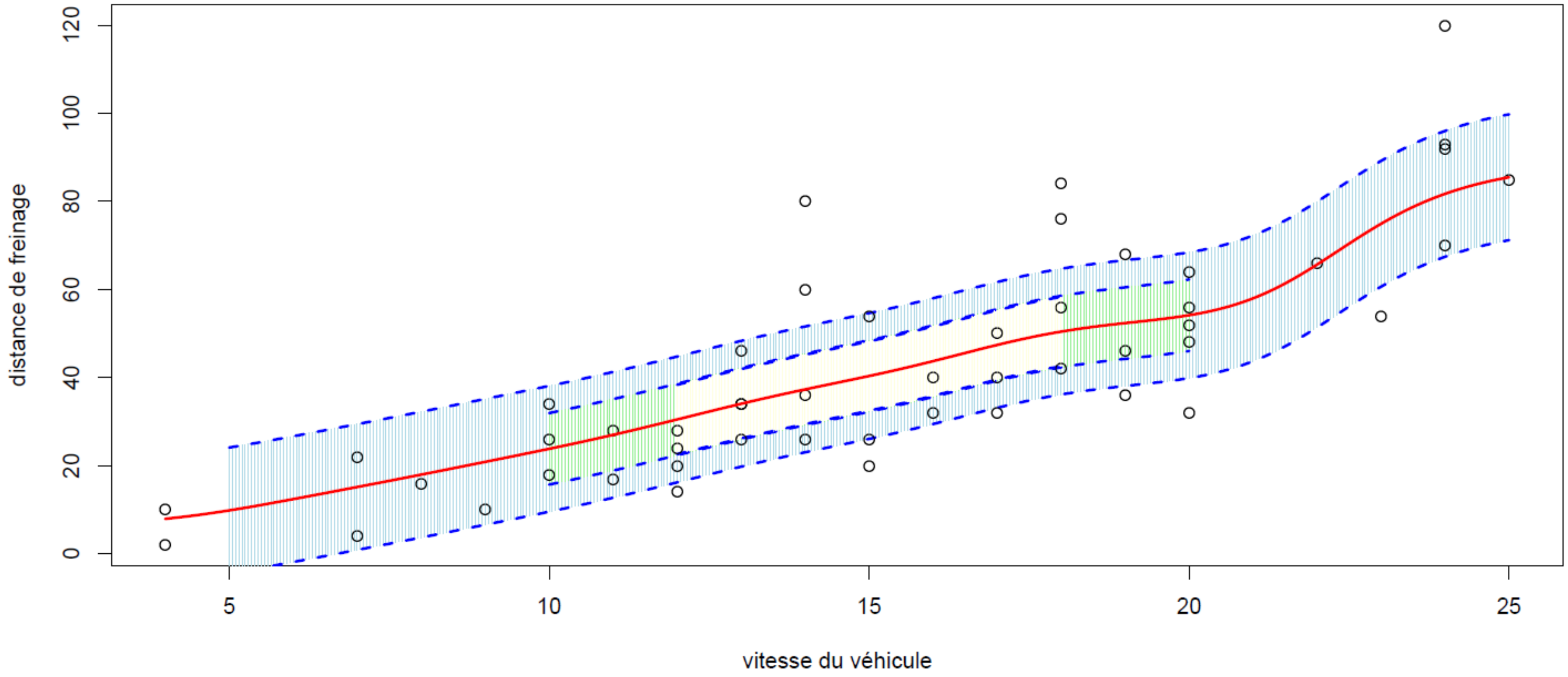
- Bootstrap (see #2)

Finally, McDonald (1986) Smoothing with Split Linear Fits suggested a bootstrap algorithm to approximate the distribution of $Z_n = \sup\{|\widehat{\varphi}(x) - \varphi(x)|, x \in \mathcal{X}\}$.



≪ Confidence Bands

Depending on the smoothing parameter h, we get different corrections



≪ Confidence Bands

Depending on the smoothing parameter h, we get different corrections