Mathematical Properties of Continuous Ranked Probability Score Forecasting

Romain Pic [1] Clément Dombry [1] Philippe Naveau [2] Maxime Taillardat [3]

[1]LMB (CNRS/UBFC) [2]LSCE (CEA/CNRS/UVSQ) [3]CNRM (Météo-France/CNRS)

1. Introduction

Context - Probabilistic forecasting and its evaluation with scoring rules are widely used in fields such as weather forecasting, finance and sustainable energies. In meteorological forecasting, statistical postprocessing techniques are essential to improve forecasts made by physical models. Numerous state-of-the-art statistical postprocessing techniques are based on distributional regression evaluated with the Continuous Ranked Probability Score (CRPS).
Goal - Obtain theoretical properties of such minimization of the CRPS within a conditional framework and finite sample sizes.

2. What is Distributional Regression?

- **Aim** : Estimate the conditional distribution of a variable of interest Y (e.g. 1h cumulated precipitation) given covariables X (e.g. pressure, temperature, wind speed; see Fig. 1)
- The conditional distribution can be estimated in various ways such as a **cumulative distribution function** (cdf), a quantile function or an ensemble of members.

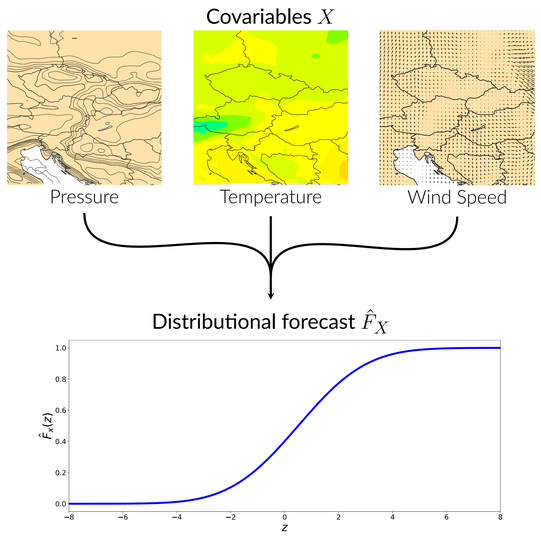

Figure 1. Example of distributional forecast.

Evaluate the prediction : **need a loss function able to compare a distribution and an observation.**

3. Evaluation with the Continuous Ranked Probability Score

- Scoring rule introduced by Matheson and Winkler (1976, [3]). See Fig. 2.
- Strictly proper scoring rule $\rightarrow$ minimizing the CRPS leads to the true distribution of Y given X.

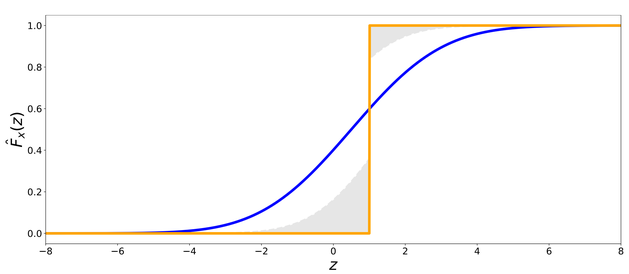

Figure 2. The CRPS (gray area) is the square of the difference between the cdf of the forecast (blue) and the empirical cdf of the observation (orange) integrated over $\mathbb{R}$.

4. Statistical Learning

- **Statistical Learning** : methods using n previous observations of (X, Y) to learn to estimate the conditional distribution of Y given X.
- Assumption : the training data $(X_i, Y_i)_{i\in[\![1,n]\!]}$ and the test observation are independent and from the same distribution denoted P.
- Any statistical learning algorithm $\hat{F}_n$ can be linked to a theoretical risk in terms of CRPS. This is the expected CRPS averaged over the training data.

$$R_P(\hat{F}_n) = \mathbb{E}_{(X_i,Y_i)\sim P}\mathbb{E}_{(X,Y)\sim P}\left[\mathrm{CRPS}(\hat{F}_{n,X}, Y)\right]$$

- The conditional distribution of Y given X, denoted F^*_X is the best forecast possible. The risk $R_P(F^*)$ associated is **minimal**, this is the **Bayesian risk**.
- **Naive kernel method** (see Fig. 3)

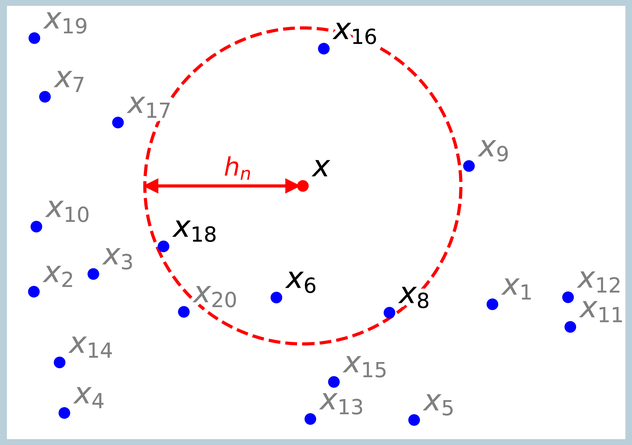

Figure 3. The naive kernel method is the average over the empirical cdf of the training observations that are within the ball of radius h_n in the covariates space.

$$\hat{F}_{n,X}(z) = \frac{\sum_{i=1}^n \mathbb{1}_{\|X-X_i\|\leq h_n}\mathbb{1}_{Y_i\leq z}}{\sum_{i=1}^n \mathbb{1}_{\|X-X_i\|\leq h_n}}$$

- **Remark.** *The index X denotes a dependence w.r.t the covariables and the index n a dependence w.r.t. the training data.*

5. Convergence

- **Choice of convergence** : Convergence of the maximal error on a given class. (minimax error)

Definition. A sequence of positive numbers (a_n) is called an **optimal minimax rate of convergence** on the class $\mathcal{D}$ if

- (**lower bound**) any algorithm $\hat{F}_n$ satisfies $\sup_{P\in\mathcal{D}}(R_P(\hat{F}_n) - R_P(F^*)) \geq \epsilon a_n$ for $\epsilon > 0$ and n large enough.
- (upper bound) there exists an algorithm $\hat{F}_n$ satisfying $\sup_{P\in\mathcal{D}}(R_P(\hat{F}_n) - R_P(F^*)) < \epsilon^{-1} a_n$ for $\epsilon > 0$ and n large enough.

- **Class of distributions** $\mathcal{D}^{(h,C,M)}$:
  - i) Covariables need to be on a **compact set** (e.g. $[0,1]^d$).
  - ii) The **dispersion** of Y given X needs to be majored.

$$\forall x \in [0,1]^d, \int_{\mathbb{R}} F^*_x(z)(1 - F^*_x(z))\mathrm{d}z \leq M$$

  - iii) The **regularity** of F^*_x needs to be controlled in order to extrapolate from training data.

$$\|F^*_x - F^*_{x'}\|_{L^2} \leq C\|x - x'\|^h,\ \forall x, x' \in [0,1]^d$$

6. Results

- (**lower bound**) — Use a subclass of distributions with a binary response to obtain a **lower minimax rate of convergence** : $a_n = n^{-\frac{2h}{2h+d}}$ (Györfi et al. 2006 [2]).
- (upper bound) for the naive kernel method

Proposition. Assume $P \in \mathcal{D}^{(h,C,M)}$ and let $\hat{F}_n$ be the naive kernel model. Then,

$$R_P(\hat{F}_n) - R_P(F^*) \leq \tilde{c}_d \frac{2M + Cd^{h/2} + \frac{M}{n}}{nh_n^d} + C^2 h_n^{2h}$$

where $\tilde{c}_d$ only depends on $d \geq 1$.

- **Optimal minimax rate of convergence** — Minimizing the upper bound with respect to h_n leads to the same dependence in n as the lower bound.

Theorem. For any d, the optimal minimax rate of convergence on the class $\mathcal{D}^{(h,C,M)}$ is $a_n = n^{-\frac{2h}{2h+d}}$. Moreover, the naive kernel algorithm reaches the optimal rate of convergence for $h_n = \left(\frac{\tilde{c}_d d(M + Cd^{h/2} + \frac{M}{n})}{2hC^2}\right)^{\frac{1}{2h+d}} n^{-\frac{1}{2h+d}}$.

- Remarks :
  - With regard to the sample size, distributional regression evaluated with the CRPS converges at the **same rate as point regression** even though the distributional estimate carries more information regarding the prediction of the underlying process.
  - The naive kernel method is related to the **Analog Method** (Delle Monache et al., 2013 [4]).
  - The results on naive kernel methods can be used to obtain convergence rate on a broader type of kernel methods : **boxed kernel** methods.

Conclusion

- Obtain the optimal minimax rate of convergence for any d.
- Upper bound on the convergence rate for kernel methods (and k-NN, see the associated article) at fixed n.
- Extension to usual weighted CRPSs (not shown but available upon request).
- *Perspectives* - Looking at the convergence of state-of-the-art techniques such as Quantile Regression Forests (Taillardat et al., 2016 [5]).
- *Related work* - Stone's theorem for distributional regression in Wasserstein distance, Dombry et al. (2023, [1]).

Sharing is encouraged

Associated article

[1] C. Dombry et al. "Stone's theorem for distributional regression in Wasserstein distance". In: (2023). arXiv: 2302.00975 [math.ST].

[2] L. Györfi et al. *A Distribution-Free Theory of Nonparametric Regression*. Springer New York, 2006. 650 pp. isbn: 9780387224428.

[3] J. E. Matheson and R. L. Winkler. "Scoring Rules for Continuous Probability Distributions". In: *Management Science* 22.10 (1976).

[4] L. D. Monache et al. "Probabilistic Weather Prediction with an Analog Ensemble". In: *Monthly Weather Review* 141.10 (2013).

[5] M. Taillardat et al. "Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics". In: *Monthly Weather Review* 144.6 (2016).