Cadre Théorique de l'Apprentissage Statistique pour la Régression Distributionnelle utilisant le Continuous Ranked Probability Score

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

[1] *Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon, UMR CNRS 6623, 25000 Besançon, France. E-mails :* romain.pic@univ-fcomte.fr, clement.dombry@univ-fcomte.fr

[2] *Université Paris-Saclay, Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, 91191 Gif-sur-Yvette, France. E-mail :* philippe.naveau@lsce.ipsl.fr

[3] *Météo France, Centre National de Recherches Météorologique, UMR 3589, 31057 Toulouse, France. E-mail :* maxime.taillardat@meteo.fr

Résumé. La régression distributionnelle répond à un besoin fondamental de l'analyse statistique : permettre de faire des prévisions tout en quantifiant leur incertitude. Cette approche surmonte les limites de la régression classique qui estime uniquement l'espérance conditionnellement aux covariables en fournissant un estimateur de l'intégralité de la loi conditionnelle. Cette méthodologie, dite de prédiction probabiliste, est largement adoptée dans de nombreux domaines tels que la météorologie et la production d'énergie, mais ses aspects théoriques restent peu développés. Par analogie avec la théorie classique de l'apprentissage statistique, nous définissons un cadre où le prédicteur est une loi de probabilité, dite loi prédictive, et où la fonction de perte est donnée par un score strictement propre au sens de Gneiting et Raftery (2007). Le prédicteur de Bayes coïncide alors avec la loi conditionnelle. Dans le cas du CRPS, nous étudions ensuite la vitesse minimax de convergence et montrons en particulier qu'en dimension supérieure ou égale à 2, l'algorithme des k plus proches voisins pour la régression distributionnelle atteint le taux minimax optimal.

Mots-clés. régression distributionnelle, apprentissage statistique, prédicteur de Bayes, taux minimax, méthode des plus proches voisins

Abstract. The distributional regression fulfills a fundamental need of statistical analysis : being able to make forecasts and quantify their uncertainty. This approach overcomes the limits of classical regression which estimates only the conditional mean by estimation the whole distribution law. This methodology, called probabilistic forecast, is widely used in numerous fields such as meteorology and energy production, but its theoretical aspects have not been studied. By analogy with the classical theory of statistical learning, we define a framework where the predictor is a law of probability, called prediction law, and where the loss function is given by a strictly proper scoring rule in the sense of Gneiting and Raftery (2007). Bayes predictor is then the conditional law. In the case of the Continuous Ranked Probability Score, we study then the minimax rate of convergence

and show that, in particular in dimension higher or equal to 2, the k nearest neighbor algorithm for the distributional regression reaches the optimal rate of convergence.

Keywords. distributional regression, statistical learning, Bayes predictor, minimax rate, nearest neighbors method

1 Introduction

De nombreux domaines nécessitent la capacité de prédire la valeur future de variables cibles. Dès 1906, W. Ernest Cook met en avant l'importance de la caractérisation de l'incertitude liée à une prédiction dans *Monthly Weather Review*. Ainsi, la globalité des méthodes qui ont été utilisées et développées depuis estiment simultanément la prédiction de la variable cible et l'incertitude qui lui est associée comme expliqué dans Kneib et al. (2021).

La régression distributionnelle est largement utilisée en post-traitement météorologique avec une approche visant à minimiser le risque empirique d'un algorithme à partir du Continuous Ranked Probability Score (CRPS). Par exemple, la méthode Ensemble Model Ouput Statistics (Gneiting et al., 2005) et les méthodes basées sur les réseaux de neurones (Schulz et Lerch, 2021), utilisent cette minimisation du CRPS.

Dans les travaux présentés ci-après, nous cherchons à expliquer une approche statistique (Györfi et al., 2002) minimisant le CRPS. Nous commençons par introduire les différentes notions liées à la régression distributionnelle et à l'apprentissage statistique. Puis, nous présentons un résultat que nous avons obtenus par analogie avec l'apprentissage statistique classique.

2 Régression Distributionnelle et Apprentissage Statistique

Les modèles de régression distributionnelle permettent de surmonter les limites de l'approche traditionnelle qui estime l'espérence de la variable cible conditionnée par des covariables. Ces modèles estiment l'intégralité de la distribution conditionnelle de la variable cible et cela permet une approche complète du phénomène étudié et d'évaluer l'incertitude associé à l'estimation. Ce développement récent était motivé par des enjeux concrets dans de nombreux domaines tels que la prévision météorologique, la prédiction de risques sismiques et la gestion de risques économique et financier comme présenté dans Gneiting et Katzfuss (2014).

Nous considérons $Y \in \mathbb{R}$ la variable aléatoire cible (e.g. les précipitations) et $X \in \mathbb{R}^d$ la variable aléatoire représentant les covariables (e.g. température, humidité, pression,

etc...). L'objectif de la régression distributionnelle est d'estimer la distribution conditionnelle de Y sachant $X = x$, notée $\mathbb{P}_{Y|X=x}(\mathrm{d}y) = F_x(\mathrm{d}y)$.

Définition 1. *Un prédicteur est une application mesurable $F\colon \mathbb{R}^d \to \mathcal{P}$, avec $\mathcal{P} \subseteq \mathcal{M}(\mathbb{R})$, avec $\mathcal{M}(\mathbb{R})$ qui est l'ensemble des distributions de $\mathbb{R}$.*

L'apprentissage statistique fait intervenir une fonction de perte qui compare la prédiction avec des observations. Dans le cadre de la prédiction probabiliste, les fonctions de perte sont appelées *scores* et comparent la distribution prédictive F avec une observation (x, y).

Définition 2. *Un score est une fonction à valeurs réelles étendues $S : \mathcal{P} \times \mathbb{R} \to [-\infty, +\infty]$, telle que $S(F, .)$ est $\mathcal{P}$-quasi-intégrable pour tout $F \in \mathcal{P}$.*

La fonction de perte est utilisée pour évaluer la qualité d'un prédicteur grâce au *risque théorique*, correspondant à l'espérance du score. Le risque dépend de P, la distribution de (X, Y).

Définition 3. *Le risque d'un prédicteur F est défini par $R_P(F) = \mathbb{E}[S(F(X), Y)]$.*

La notion de score propre est essentielle en théorie et en pratique. Nous introduisons le score moyen d'une distribution prédictive donnée F lorsque les observations suivent une loi G par

$$\bar{S}(F, G) = \int S(F, y)G(\mathrm{d}y), \quad F, G \in \mathcal{P} \subseteq \mathcal{M}(\mathbb{R})$$

Définition 4. *Soit S un score et $\mathcal{P}$ un sous-ensemble de $\mathcal{M}(\mathbb{R})$. S est dit propre sur $\mathcal{P}$ s'il vérifie la propriété suivante :*

$$\bar{S}(G, G) \leq \bar{S}(F, G) \quad \textit{pour tout } F, G \in \mathcal{P}. \tag{1}$$

On dit que S est strictement propre si l'égalité dans (1) tient si et seulement si $F = G$.

Définition 5. *Le **risque de Bayes** est défini par :*

$$R_P^* = \inf_{F \in \mathcal{F}(\mathbb{R}^d, \mathcal{M}(\mathbb{R}))} \mathbb{E}_{(X,Y)\sim P}[S(F_X, Y)]$$

*avec F_X la fonction de répartition prédite de Y sachant X. De plus, si l'infimum est atteint, le prédicteur associé F^* est appelé **prédicteur de Bayes**.*

Le risque de Bayes correspond au risque théorique minimal. Ainsi, l'écart entre le risque d'un prédicteur et le risque de Bayes est une quantité d'intérêt lorsque l'on étudie les performances d'une méthode de prédiction.

Dans notre étude, nous nous sommes placés dans le cadre de l'apprentissage statistique qui correspond au cas pratique où l'on dispose d'un nombre n d'observations de la variable cible et des covariables et où l'on cherche à prédire la distribution de Y sachant les covariables X. On note l'échantillon de n observation

$$D_n = \{(X_i, Y_i), i \in [\![1; n]\!]\}.$$

Définition 6. *Un algorithme de prédiction est une application mesurable*

$$\hat{F} : (\mathbb{R}^d \times \mathbb{R})^n \to \mathcal{F}(\mathbb{R}^d, \mathcal{P}), \text{ avec } \mathcal{P} \subseteq \mathcal{M}(\mathbb{R}).$$
$$d_n \mapsto \hat{F}_n$$

$\hat{F}$ associe un prédicteur $\hat{F}_n$ à un échantillon $d_n = \{(x_i, y_i), i \in [\![1; n]\!]\}$.

De plus, nous considérons le Continuous Ranked Probability Score (CRPS) qui est un score strictement propre largement utilisé pour les prévisions météorologiques. Le CRPS a été défini dans Matheson et Winkler (1976) comme suit :

$$S(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}_{y \leq z})^2 \mathrm{d}z$$

3 Taux de Convergence Optimal

3.1 Définitions

Dans ce contexte théorique, nous nous intéressons au taux de convergence du risque d'un algorithme de prédiction vers le risque de Bayes. Nous étudions donc la grandeur suivante :

$$\bar{S}(F, G) - \bar{S}(G, G) = \mathbb{E}_{Y \sim G}[S(F, Y) - S(G, Y)] = \int |F(z) - G(z)|^2 \mathrm{d}z$$

Pour étudier le taux de convergence des algorithmes, nous nous intéressons à la notion de *taux de convergence minimax minimal* et au *taux de convergence optimal*.

Définition 7. *La suite de nombres positifs a_n est appelée taux de convergence minimax minimal pour la classe $\mathcal{D}$ si*

$$\liminf_{n \to \infty} \inf_{\hat{F}_n} \sup_{(X,Y) \in \mathcal{D}} \frac{\mathbb{E}[S(\hat{F}_{n,X}(Y), Y)] - \mathbb{E}[S(F^*_X(Y), Y)]}{a_n} = C_1 > 0$$

De plus, cette suite a_n est appelée taux de convergence optimal pour la classe $\mathcal{D}$ s'il existe un algorithme $\hat{F}_n$ tel que :

$$\limsup_{n \to \infty} \sup_{(X,Y) \in \mathcal{D}} \frac{\mathbb{E}[S(\hat{F}_{n,X}(Y), Y)] - \mathbb{E}[S(F^*_X(Y), Y)]}{a_n} = C_0 < \infty$$

La notion de taux de convergence minimal minimal est lié au fait de vouloir minimiser l'erreur maximal au sein d'une classe de distribution.

Nous introduisons l'algorithme des k plus proches voisins (k-NN) adapté à la régression distributionnelle. La méthode des k-NN estime la distribution conditionnelle de Y sachant $X = x$ en moyennant des distributions de Dirac centrées sur la valeur de la variable cible prise par les k plus proches voisins de x parmi l'échantillon utilisé pour l'apprentissage d_n. Il peut être formulé des deux manières suivantes :

$$\hat{F}_k(x) = \frac{1}{k} \sum_{i \in knn(x)} \delta_{y_i} = \frac{1}{k} \sum_{i=1}^{k} \delta_{y_{(i)}}$$

où l'indexation par (i) correspond au i-ème plus proche voisin de x parmi d_n.

3.2 Résultats

Par analogie avec des résultats en apprentissage statistique classique (Biau et Devroye (2015), Bobkov et Ledoux (2019), Györfi et al. (2002)), nous avons obtenus des résultats concernant le taux de convergence optimal pour la classe de distribution $\mathcal{D}^{(H,C,M)}$ lorsque la dimension des covariables d est supérieure à 2 et ce taux optimal est atteint pour l'algorithme des k-NN.

Définition 8. *Soit $\mathcal{D}^{(H,C,M)}$ une classe de distribution de (X, Y) telle que :*

i) $X \in [0,1]^d$;

ii) Pour tout x, F_x^* vérifie $\int_{\mathbb{R}} F_x^*(z)(1 - F_x^*(z))\mathrm{d}z < M$; et

iii) $\|F_{x'}^* - F_x^*\|_{L^2} \leq C\|x' - x\|^H$ pour tout $x, x' \in [0,1]^d$,

où F_x^* est la fonction de répartition de Y sachant $X = x$, $M \geq 0$, $0 < H \leq 1$ et $C > 0$.

Théorème 1. *La suite $a_n = n^{-\frac{2H}{2H+d}}$ est le taux de convergence optimal de la classe $\mathcal{D}^{(H,C,M)}$ pour $d \geq 2$.*

Autrement dit, a_n vérifie les propriétés suivantes :

$$\liminf_{n\to\infty} \inf_{\hat{F}_n} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbb{E}[S(\hat{F}_{n,X}(Y), Y)] - \mathbb{E}[S(F_X^*(Y), Y)]}{n^{-\frac{2H}{2H+d}}} = C_1 > 0$$

et, pour $\hat{F}_n$ l'algorithme des k_n plus proche voisins avec $k_n = \left(\frac{Md}{4HCc_d^H}\right)^{\frac{d}{2H+d}} n^{\frac{2H}{2H+d}}$, on a :

$$\limsup_{n\to\infty} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbb{E}[S(\hat{F}_{n,X}(Y), Y)] - \mathbb{E}[S(F_X^*(Y), Y)]}{a_n} = C_0 < \infty$$

avec $C_1 > 0$ dépendant de C, H, M et d et $c_d = \frac{2^{3+\frac{2}{d}}(1+\sqrt{d})^2}{V_d^{2/d}}$, avec V_d le volume de la boule euclidienne unitaire dans $\mathbb{R}^d$.

Pour obtenir ce résultat, nous avons minoré le taux de convergence minimax minimal de la classe $\mathcal{D}^{(H,C,M)}$ en considéré une sous-classe dont on peut calculer facilement le taux minimax et cela permet de minorer le taux de convergence minimax minimal de $\mathcal{D}^{(H,C,M)}$ comme le fait Györfi et al. (2002) pour la régression classique. Ensuite, nous montrons que l'algorithme des k-NN réalise ce taux de convergence, ce qui en fait un taux de convergence optimal.

Références

[1] Gérard Biau et Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015.

[2] Sergey Bobkov et Michel Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*. Memoirs of the American Mathematical Society, 2019. doi : 10.1090/memo/1259.

[3] Tilmann Gneiting et Matthiass Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and its Applications*, 2014. doi : 10.1146/annurev-statistics-062713-085831.

[4] Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld III, et Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. 133 :1098–1118, 2005. doi : 10.1175/MWR2904.1.

[5] Lászlò Györfi, Michael Kohler, Adam Krzyżak, et Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, 2002.

[6] Thomas Kneib, Alexander Silbersdorff, et Benjamin Säfken. Rage against the mean – a review of distributional regression approaches. 2021. doi : 10.1016/j.ecosta.2021.07.006.

[7] James E. Matheson et Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, pages 1087–1096, 1976. doi : 10.1287/mnsc.22.10.1087.

[8] Benedikt Schulz et Sebastian Lerch. Machine learning methods for postprocessing ensemble forecasts of wind gusts : A systematic comparison. 2021. doi : 10.1175/MWR-D-21-0150.1.