Thesis proposal

4 avril 2024

1 General information

Supervisors : Clément DOMBRY, Davit VARRON, Ahmed ZAOUI, Laboratoire de Mathématique de Besançon.
Titre : "Empirical process methods in distributional regression".
Important : unlike the majority of thesis subjects offered by the doctoral school, **this subject will not be accompanied by any funding** for the doctoral student. This means that the candidate must be able to support themselves financially through another means, such as full/part-time employment. The duration of the doctorate will therefore be adjusted accordingly.

2 Introduction

The statistical framework of this thesis is the distributional regression problem, namely : how to properly model and estimate the conditional cumulative distribution functions (c.d.f.)

$$F_x^*(y) = \mathbb{P}\left(Y \leq y | X = x\right)$$

given n independent, identically distributed copies (X_i, Y_i) of (X, Y). It is crucial to emphasize that this topic is more intricate and challenging than what is typically encountered in classical regression framework : instead of building point-to-point predictions for $\mathbb{E}\big(Y \mid X = x\big)$ we rather predict the **entire conditional distribution** $\mathbb{P}\big(Y \in \cdot \mid X = x\big)$.
Distributional regression is of relevance in various fields, such as meteorology [7, 4], distributional random forest [1], probabilistic forecasting [2], quantile regression [6], or structured additive distributional regression [5].
The distributional regression network (DRN) method has been developed by [8] as an evolution of the classical Ensemble Model Output Statistics (EMOS) framework, adding significantly more flexibility through the use of neural networks. More precisely, given

$$\big\{F_{\mu,\sigma}, \mu \in \mathbb{R},\ \sigma > 0\big\},$$

a parametric collection of c.d.f (i.e, a parametric model for the laws of Y given $X = x$), it is assumed that

$$F_x^*(y) = F_{\mu_{\theta_1}(x), \sigma_{\theta_2}(x)}(y),$$

where both $\mu_{\theta_1}(\cdot)$ and $\log \sigma_{\theta_2}(\cdot)$ are regression neural networks. Therefore, the DRN aims to enhance the model's capability to represent complex phenomena by employing approaches better suited to the nonlinear structures of meteorological data, as opposed to traditional methods relying on linear and fixed links. This approach provides greater flexibility and adaptability to model relationships between predictive variables and parameters in a more realistic manner.

In this thesis we will investigate the estimation procedure based on the empirical risk minimisation principle, with a risk function that is derived from the concept of *Continuous Ranked Probability Score* (CRPS).

3 Background on distributional regression and main goals

3.1 Probabilistic forecast and its evaluation with scoring rules

We first consider the simple setting of probabilistic forecasting without covariate where a a future observation $Y \sim G$ is predictive by a probability distribution F, called predictive distribution. Proper scoring rules are used in order to compare the predictive distribution F and the materializing obervation y which are objects of different nature.

Let $\mathcal{D} \subset \mathcal{P}(\mathbb{R})$ denote a subset of the set of all probability measures on $\mathbb{R}$, often identified with their c.d.f. A scoring rule [1] on $\mathcal{D}$ is a function $S\colon \mathcal{D} \times \mathbb{R} \to [0, +\infty)$. The quantity $S(F, y)$ is interpreted as the error between the predictive distribution F and the materializing observation y. The mean error when $Y \sim G$ is denoted by

$$\bar{S}(F, G) = \mathbb{E}_{Y \sim G}[S(F, Y)].$$

The following notion of proper and strictly proper scoring rule is central in the theory.

Définition 3.1 *The scoring rule S is said proper on $\mathcal{D}$ when*

$$\bar{S}(F, G) \geq \bar{S}(G, G), \quad \textit{for all } F, G \in \mathcal{D}. \tag{3.1}$$

It is said strictly proper when equality in Eq. (3.1) *implies the equality $F = G$.*

Stated differently, the scoring rule S is strictly proper on $\mathcal{D}$ when

$$\arg\min_{F \in \mathcal{L}} \bar{S}(F, G) = \{G\}, \quad \text{for all } G \in \mathcal{D}.$$

The interpretation is that, in order to minimize its mean error, the forecaster has to predict the "true" observation distribution G.

In this thesis, we will consider the Continuous Ranked Probability Score (CRPS, [7]). This scoring rule is defined by the formula

$$S(F, y) = \int_{\mathbb{R}} \left(\mathbb{1}_{\{y \leq z\}} - F(z)\right)^2 \mathrm{d}z \tag{3.2}$$

1. For the sake of simplicity, we consider only the case of real-valued observation Y and of non-negative scoring rule. A more general definition can be found in [3].

for all F finite absolute moment, that is the Wasserstein space

$$\mathcal{P}_1(\mathbb{R}) = \Big\{ F \in \mathcal{P}(\mathbb{R}) \colon m_1(F) = \int_{\mathbb{R}} |y| F(\mathrm{d}y) < \infty \Big\}. \tag{3.3}$$

One can easily check from this definition that

$$\bar{S}(F, G) = \int_{\mathbb{R}} G(z)\,(1 - G(z))\,\mathrm{d}z + \int_{\mathbb{R}} (F(z) - G(z))^2\,\mathrm{d}z,$$

which implies

$$\bar{S}(F, G) - \bar{S}(G, G) = \int_{\mathbb{R}} (F(z) - G(z))^2\,\mathrm{d}z.$$

This quantity is nonnegative and vanishes if and only if $F = G$, ensuring that the CRPS is a strictly proper scoring rule on the Wasserstein space $\mathcal{P}_1(\mathbb{R})$.

The following formula, widely used in practice, provides a closed form expression for the CRPS when the predictive distribution is the Gaussian distribution $F = \mathcal{N}(m, \sigma^2)$: then

$$S(F, y) = \sigma \left(z(2\Phi(z) - 1) + 2\phi(z) - 1/\sqrt{\pi} \right),$$

with $z = (y - \mu)/\sigma$ and Φ (resp. ϕ) denoting the cdf (resp. df) of the standard normal distribution.

3.2 Model fitting, model selection and model aggregation

We now present the methods and aims of this thesis.

3.2.1 Theoretical risk in distributional regression

In a regression framework, we observe a sample $\mathcal{D}_n = \{(X_i, Y_i),\ 1 \leq i \leq n\}$ of independent copies of $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$. Distributional regression aims at estimating the conditional distribution $Y|X = x$ characterized by

$$F_x^*(y) = \mathbb{P}\left(Y \leq y \mid X = x\right), x \in \mathbb{R}^d.$$

The forecaster uses the trainging sample $\mathcal{D}_n$ and some algorithm to build a *functional* estimator $\hat{F}_n \colon\ : x \mapsto \hat{F}_{n,x}$ of the map $F^* \colon x \mapsto F_x^*$. The accuracy of this estimator is here measured by its theoretical risk

$$\mathcal{R}(\hat{F}_n) = \mathbb{E}\left[S(\hat{F}_{n,X}, Y) \right].$$

where expectation is taken with respect to the joint law of (X, Y). This quantity can be seen as the counterpart of the mean squared error in point regression. The excess risk of $\hat{F}_n$ is defined as

$$\begin{aligned} \mathcal{R}(\hat{F}_n) - \mathcal{R}(F^*) &= \mathbb{E}\left[S(\hat{F}_{n,X}, Y) - S(F_X^*, Y) \right] \\ &= \mathbb{E}\left[\bar{S}(\hat{F}_{n,X}, F_X^*) - S(F_X^*, F_X^*) \right] \geq 0. \end{aligned}$$

The nonnegativity is ensured by the fact that S is a proper scoring rule according to Definition 3.1. If S is strictly proper, the excess risk is equal to 0 if and only if $\hat{F}_{n,x} = F_x^*$ almost everywhere (with respect to $P_X(\mathrm{d}x)$).

For the CRPS, the excess risk is rewritten as

$$\mathcal{R}(\hat{F}_n) - \mathcal{R}(F^*) = \mathbb{E}\left[\int_{\mathbb{R}} \left|\hat{F}_{n,X}(u) - F_X^*(u)\right|^2 du\right].$$

3.2.2 Model fitting

Our first interest lies in model fitting by empirical risk minimization. Here we consider a parametric family $(F_\theta)_{\theta\in\Theta}$, $\Theta \subset \mathbb{R}^K$, where $F_\theta \colon x \in \mathbb{R} \mapsto F_{\theta,x} \in \mathcal{D} \subset \mathcal{P}(\mathbb{R})$. The empirical risk associated with F_θ is computed on the training sample $\mathcal{D}_n$ by

$$\hat{\mathcal{R}}_n(F_\theta) = \frac{1}{n}\sum_{i=1}^{n} S(F_{\theta,X_i}, Y_i)$$

and is an empirical counterpart of the theoretical risk $\mathcal{R}(F_\theta)$. Empirical risk minimization consists in finding

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \hat{\mathcal{R}}_n(F_\theta) \tag{3.4}$$

and proposing the estimator $F_{\hat{\theta}_n}$ which is thought as almost optimal within the family $(F_\theta)_{\theta\in\Theta}$. A classical decomposition of the excess risk of the corresponding estimator is given by

$$\mathcal{R}(F_{\hat{\theta}_n}) - \mathcal{R}(F^*) = \Big(\mathcal{R}(F_{\hat{\theta}_n}) - \inf_{\theta\in\Theta} \mathcal{R}(F_\theta)\Big) + \Big(\inf_{\theta\in\Theta} \mathcal{R}(F_\theta) - \mathcal{R}(F^*)\Big). \tag{3.5}$$

where the two terms are called the estimation error and the approximation error respectively. The approximation error is deterministic and depends on the ability of the family $(F_\theta)_{\theta\in\Theta}$ to approximate F^*. The estimation error depends on the training sample $\mathcal{D}_n$ and on the complexity of the familiy $(F_\theta)_{\theta\in\Theta}$. This decomposition is akin to the decomposition of the mean squared error into squared bias and variance in point regression. Our first goal is the following :

Goal 1 : provide non asymptotic estimates for the estimation error $\mathcal{R}(F_{\hat{\theta}_n}) - \inf_{\theta\in\Theta} \mathcal{R}(F_\theta)$.

3.2.3 Model selection

Our second interest lies in model selection via validation error minimization. Here we suppose that a validation sample $\mathcal{D}'_N = \{(X'_i, Y'_i), \quad 1 \le i \le N\}$ is available, which is assumed independent of the training sample $\mathcal{D}_n$.

A common situation in machine learning is that we have M algorithms at hand that are trained on $\mathcal{D}_n$, resulting in models $\hat{F}_n^1, \ldots, \hat{F}_n^M$. In order to select the best model, we compute the empirical risks on the validation sample

$$\hat{\mathcal{R}}'_N(\hat{F}_n^m) = \frac{1}{N}\sum_{i=1}^{N} S(\hat{F}^m_{n,X'_i}, Y'_i)$$

and select the model

$$\hat{m} = \underset{1\leq m\leq M}{\arg\min}\, \hat{\mathcal{R}}'_N(\hat{F}_n^m).$$

An oracle having access to the theoretical risk would have selected

$$m^* = \underset{1\leq m\leq M}{\arg\min}\, \mathcal{R}(\hat{F}_n^m),$$

leading to the definition of the regret

$$\mathcal{R}(\hat{F}_n^{\hat{m}}) - \mathcal{R}(\hat{F}_n^{m^*}) = \mathcal{R}(\hat{F}_n^{\hat{m}}) - \min_{1\leq m\leq M} \mathcal{R}(\hat{F}_n^m).$$

Goal 2 : provide non asymptotic estimates for the regret $\mathcal{R}(\hat{F}_n^{\hat{m}}) - \min_{1\leq m\leq M} \mathcal{R}(\hat{F}_n^m)$.

3.2.4 Model aggregation

Our third interest lies in convex aggregation of models. With the same notation as in Section 3.2.3, we define the convex aggregation of the models $\hat{F}_n^1, \ldots, \hat{F}_n^M$ with the weights $\lambda_1, \ldots, \lambda_M$ by

$$\hat{F}_{n,x}^{\lambda} = \sum_{m=1}^{M} \lambda_m \hat{F}_{n,x}^m,$$

with $\lambda = (\lambda_1, \ldots, \lambda_M)$ an element of the simplex $\Lambda = \{\lambda\colon \lambda_m \geq 0, \sum_{1\leq m\leq M} \lambda_m = 1\}$. Note that $\hat{F}_n^{\lambda}$ is a valid estimator of F^* because the space $\mathcal{P}(\mathbb{R})$ of probability measures is convex.

Similarly as in model selection, the best weights for convex agregation are obtained by minimization of the validation error, that is we define

$$\hat{\lambda} = \underset{\lambda\in\Lambda}{\arg\min}\, \hat{\mathcal{R}}'_N(\hat{F}_n^{\lambda}).$$

Goal 3 : provide non asymptotic estimates for the regret $\mathcal{R}(\hat{F}_n^{\hat{\lambda}}) - \min_{\lambda\in\Lambda} \mathcal{R}(\hat{F}_n^{\lambda})$.

4 The use of empirical processes theory

This thesis will be focused on exploring how to achieve **goals 1,2,3** by using techniques from empirical processes theory (see, e.g., [9] for a comprehensive monograph on the topic). Indeed, it is clear that a crucial step is the control of

$$\sup_{\theta\in\Theta} \mid \hat{\mathcal{R}}_n(F_\theta) - \mathcal{R}(F_\theta) \mid, \tag{4.1}$$

such as obtaining sharp upper bounds for its expectation or its probabilities of exceeding $\epsilon > 0$.

Note that (4.1) can be expressed as

$$\sup_{\theta\in\Theta} \Big|\hat{\mathcal{R}}_n(F_\theta) - \mathcal{R}(F_\theta)\Big| = \sup_{g\in\mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}(g(Z_1))\Big|. \tag{4.2}$$

Here we wrote $Z_i := (X_i, Y_i)$ and

$$\mathcal{G} := \big\{ g_\theta,\ \theta \in \Theta \big\},\ \text{with} \tag{4.3}$$

$$g_\theta : (x, y) \to CRPS(F_{\theta,x}, y), \tag{4.4}$$

and Θ is either a finite of infinite set depending on **goal 1,2,3**.
It is hence clear that it is worth exploring the properties of the class $\mathcal{G}$ which arises depending on **goals 1,2,3**. Such properties have been intensively investigated since 1980 :

1. Controlling bracketing numbers of $\mathcal{G}$. For example through the regularity properties of the elements of $\mathcal{G}$)
2. Controlling uniform entropy numbers. For example using the Vapnik-Chervonenkis combinatorial theory.

Once this empirical approach will be well understood and proven successfull in the present framework, the student will have the opportunity to extend it to other choices of scoring rules.

Références

[1] D. ?evid, L. Michel, J. Näf, P. Bühlmann, and N. Meinshausen. Distributional random forests : Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333) :1–79, 2022.

[2] T. Gneiting and M. Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1) :125–151, 2014.

[3] T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477) :359–378, 2007.

[4] T. Gneiting, A.E. Raftery, Westveld A.H., and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5) :1098–1118, 2005.

[5] N. Klein, T. Kneib, and S. Lang. Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, 110(509) :405–419, 2015.

[6] R. Koenker. *Quantile Regression*. Cambridge MA : Cambridge University Press, 2005.

[7] James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10) :1087–1096, 1976.

[8] S. Rasp and S. Lerch. Neural networks for post-processing ensemble weather forecasts. *Monthly Weather Review*, 146 :3885–3900, 2018.

[9] A.W. Van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer, 2023.