www.onera.fr

ONERA
THE FRENCH AEROSPACE LAB

PROPOSITION DE SUJET DE THESE

Intitulé : Extreme quantile estimation enriched by tranfer learning

Référence : **TIS-DTIS-2024-44**
(à rappeler dans toute correspondance)

Début de la thèse : October 2024	**Date limite de candidature :** May 2024

Mots clés : Transfer learning, rare event estimation, data enrichment, extreme value theory, multi-fidelity

Profil et compétences recherchées

Applied mathematics, Probability / statistics, Python programming.

Présentation du projet doctoral, contexte et objectif

Learning to play violin can be seen as a difficult task. However, if one knows how to play guitar or piano, the knowledge already acquired on these instruments can be used to make the learning procedure of violin easier. This is the philosophy of transfer learning [1]. In practice, we consider a target task for which we have a sample not large enough to get a satisfying estimation. Moreover, we have one (or more) larger sample of data (call source) which are close to the target task. The question arising are the following:

1) How can we use the information contained in the large sample to improve the estimation of the target task?
2) Can we quantify the pertinence of the use of the large sample? Indeed, if the second sample is too far from the target task, the information introduced may lead to increase the error.

The aim of this thesis is the application of transfer learning methodologies in extreme value theory (EVT). This latest is the part of statistics which focus on the estimation of rare and intense events. One of the main goals of EVT is the estimation of quantiles beyond the range of the observed data. Such estimations need tail extrapolations. Classical results show that it is impossible to extrapolate too far from the maximum of the dataset without making important error [3]. The distance to which it is possible to extrapolate is governed by the number of extreme events present in the dataset. In this situation the use of a larger dataset via transfer learning, with more extremal events, should allows us to extrapolate farer in the tail. This way one is able to provide estimation of higher quantile (i.e rarer events) than classical methods.

The advancements outlined in this thesis will be applied in a given data context for aerospace codes with multi-fidelity characteristics. In such codes, the computational expense and precision grow in tandem with fidelity levels. Multi-fidelity modeling, in this context, facilitates precise estimations of important parameters by harmonizing results from inexpensive, less precise models with a limited number of highly accurate observations. This approach proves especially advantageous when there are strong connections between the low-fidelity (call source) and high-fidelity (target task) models, resulting in substantial computational efficiencies compared to relying solely on high-fidelity models [4]. Only two levels of fidelity will be considered in the thesis.

After a review of the literature, the first stage of this thesis will be to investigate the case of a linear link between the target and the source [5,6]. In this setup one tries to quantify the improvement, with respect to the rate of convergence, of the utilization of transfer learning in quantile estimation. Moreover, in view of the recent work [7], we can address the question of the risk bound to quantify the error made by such approaches. Secondly, we can extend the framework of quantile estimation to

quantile regression where both target and source tasks relied on covariate. In a final step, we can perform a budget constrained sensitivity analysis to determine where data enrichment should be applied between low or high fidelity sample to reduce the quantile estimation uncertainty.

References :

[1] Bozinovski, S & Fulgosi, A (1976) The influence of pattern similarity and transfer learning upon the training of a base perceptron B2. Proceedings of Symposium Informatica, 3-121-5, bled.

[2] Weissmann, I. (1978) Estimation or parameters and large quantiles based on k largest observation. Journal of the American Statistical Association, 73, 812-825

[3] Einmahl, J. H. J &De Haan, L. & Zhou, C. (2014) Statistics of heteroscedastic extremes. Journal of the Royal Statistical Society Serie B, 78.

[4] Perdikaris, P., Raissi, M., Damianou, A., Lawrence, N. D., & Karniadakis, G. E. (2017). Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 473(2198), 20160751.

[5] Obst, D. & Ghattas, B & Cugliari, J & Oppenheim, G & Claudel, S & Goude, Y. (2021) Transfer Learning for Linear Regression: a Statistical Test of Gain. Preprint.

[6] Chen, A & Owen, A & Shi, M (2015) Data enriched linear regression, Electronic Journal of Statistics, 9, 1078-1112.

[7] Tony Cai, T & Pu, H (2023) Tranfer learning for nonparametric regression : non asymptotic minimax and adaptative procedure. Preprint.

Collaborations envisagées : ISAE (Benjamin Bobbia, Benjamin.BOBBIA@isae-supaero.fr)

Laboratoire d'accueil à l'ONERA	**Directeur de thèse**
Département : Traitement de l'information et systèmes	Nom : Jérôme Morio
Lieu (centre ONERA) : Toulouse	Laboratoire : ONERA\DTIS
Contact : Jérôme Morio	Tél.: 05 62 25 26 63
Tél. : 05 62 25 26 63 Email : jerome.morio@onera.fr	Email : jerome.morio@onera.fr

Pour plus d'informations : https://www.onera.fr/rejoindre-onera/la-formation-par-la-recherche