,,

Touqeer Ahmad's CV
An Analytical Presentation

Office, 211, 51 Rue Blaise Pascal, 35170, Bruz, France
+33 750-011-746
touqeer.ahmad8960@gmail.com; touqeer.ahmad@ensai.fr

Summary of the document

The document is an analytical presentation of Touqeer Ahmad's curriculum vitae, detailing his research interests, current role, and academic background, including visiting research periods and professional experience. It provides detailed information on the courses he has taught, along with his activities. Furthermore, the document includes a summary of his PhD thesis, as well as information about the reviewers and jury members. It also covers his publications, preprints, conferences, and seminars. Additionally, the document outlines his teaching and research activities during his PhD and postdoctoral period, along with a brief discussion of his future research plans. At the end, referee information is provided if needed.

Research interests

Extreme value theory, conditional extremes, classical and Bayesian distributional regression, times series model for extremes, financial extremes, graphical models for extremal dependence, dimension reduction in conditional extremes, spatial extremes statistics, statistical learning and imbalanced data problems, designs of experiment, modeling of extreme environmental phenomena.

Current position

CREST, ENSAI **Bruz, France**
Postdoctoral Researcher in Statistics *May-2023–April-2025*
Collaborators: François Portier & Gilles Stupfler

Project details

The project develops a methodology grounded in rigorous theory for the statistical analysis of conditional extreme values of a univariate random variable given auxiliary information of interest represented by a covariate based on massive data. Here, 'massive' means that the covariate's sample size or dimension is at least very large (and possibly both). The first goal of the project is to develop a statistical framework for the inference and prediction of extreme conditional quantiles given massive data. A particular objective will be to investigate rigorous theory (asymptotic and non-asymptotic) for estimating the gradient at a moderate level of extreme appearance and reduced dimension of the covariates set. The second goal of the project is the extension of the developed methodology to the case where there are only a few examples (extreme tail events), and one interest to predict those events correctly in order to provide a comprehensive picture of risk in a given context. The methods are implemented in open source code and applied to real sets of massive data.

Education

University of Padova **Padova, Italy**
PhD in Statistics (with doctor of Europeous label) *Jan-2020-June-2023*
Thesis: On the modeling of discrete extreme values
Supervisor: Carlo Gaetan; **Co-supervisor:** Philippe Naveau
Courses attended: Functional Analysis, Probability Theory, Spatial Statistics, Theory and Methods of Inference, Generalized Linear Mixed Models, Time Series Analysis, Applied Multivariate Techniques, Statistical Consulting, Programming with Python, Sampling Theory, Bayesian Data Analysis and Computation, Models for Categorical Ordinal Data: Standard Models and Recent Developments, Kalman Filter and State Space Models, Regression Modeling with Large Data Sets.

International Islamic University **Islamabad, Pakistan**
MS Statistics (with gold medal + distinction + position) *2015–2017*
Thesis: Rainfall frequency analysis in Pakistan using Bayesian approach
Supervisor: Ishfaq Ahmad
Courses attended: Linear Models, Multivariate Methods, Advance Experimental Designs, Survey Sampling, Statistical Inference, Advance Econometrics, Bayesian Statistics, Sampling and Sampling Distributions.

International Islamic University **Islamabad, Pakistan**
MSc Statistics *2013–2015*
Courses attended: Statistical Methods, Computer Language, Research Methodology, Advance Calculus, Design and Analysis of Experiments-I, Design and Analysis of Experiments-II, Regression and Econometrics-I, Regression and Econometrics-II, Statistical Inference-I (estimation), Statistical Inference-II, Nonparametric Statistics and Categorical Data, Time Series Analysis and Forecasting, Quality Control and Quality Management, Organization Behavior, Probability and Probability Distributions-I, Probability and Probability Distributions-II, Sampling Techniques-I, Sampling Techniques-II, Linear Algebra, Data Analysis and Statistical Packages, Multivariate Statistics, Numerical Techniques

Visiting periods

Le Laboratoire des Sciences du Climat et de l'Environnement (LSCE), **Paris, France**
Jointly worked with Philippe Naveau *2022*

Universit´e de Versailles Saint-Quentin-en-Yvelines - UVSQ, **Paris, France**
Jointly worked with Julien Worms *2022*

Research center for statistics, University of Geneva **Geneva, Switzerland**
Jointly worked with Sebastian Engelke *2023*

Teaching experience

Govt. of Punjab, Higher Education Department, **Rawalpindi, Pakistan**
Lecturer Statistics (Study Leave Dec-2019 to Dec-2023) *Feb-2018 to Dec-2023*

Course Taught (202.5 hours):

1. **Introductory statistics (45hours: Feb-2018 to June-2018):** This course coveres the basic concept of statistics to bachelor students. We covered the definitions of statistics, branches of statistics, charts, types of variables, data types, Use of Charts, central tendencies, measures of dispersion, and index numbers.
   Tutorials and Workshop: As for the tutorials and workshop, the aim was to illustrate the various concepts introduced in the course and especially to focus more on the practical aspect, namely the interpretation and use of the outputs. For this, we use the R software.

2. **Statistical Theory I & II (90hours: Feb-2018 to Dec-2018):** This course introduces the concept of inferential statistics. Students covered Point and interval estimations, unbiased estimator and its

properties, variance properties and inequalities, the Central limit theorem, likelihood estimation and method of moments, hypothesis testing, likelihood ration test, and ANOVA.
Tutorials and Workshop: Similar to the previous course, we conducted workshops and tutorials for real application aspects.

3. **Statistics and Probability (45hours: Jan-2019 to June-2019):** In this course, we introduce the concept of probability theory. The following topics were covered: definitions of probability, law of probabilities, discrete random variables, discrete distribution and their properties, continuous random variables, continuous distributions and their properties, joint and marginal distributions, Moment generating functions, expected value and variance, covariance and correlation, Central Limit Theorem and Law of large numbers.
Tutorials and Workshop: Similar to the previous course, we conducted workshops and tutorials to compare the different distribution behaviors.

4. **Business Statistics:(22.5hours Sep-2019 to Nov-2019):** I taught some introductory statistics topics to commerce students in this course.
Tutorials and Workshop: Similar to the previous courses, we do tutorials using R.

Department of Statistics, AIOU **Islamabad, Pakistan**
Teaching Assistant Lecturer *July-2016 to Jan-2018*

Courses Taught (225 hours):

1. **Statistical Methods (45 hours Autumn Semester 2016):** This course was designed to cover the basic concept of statistics for master's students. In this course, I introduced various statistical methods, branches of statistics, types of variables, data types, measures of central tendency, measures of dispersion, statistical inference, hypothesis testing, and nonparametric statistics.
Tutorials and Workshop: Similar to other courses. The workshops and tutorials were given using the R software.

2. **Regression Analysis (45 hours Autumn Semester 2016):** This course covers the fundamentals of correlation, causality, and regression. Further focuses on simple and multiple linear regression, where relationships between dependent and independent variables are analyzed using linear equations. Key considerations involve assumptions like linearity, independence, homoscedasticity, and normality, which ensure model validity. The least squares method estimates regression coefficients by minimizing errors, while measures like R^2 and adjusted R^2 assess goodness of fit. Hypothesis testing, including t-tests and F-tests, evaluates coefficient significance. Other essential topics include multicollinearity, residual analysis for model adequacy, logistic regression for categorical outcomes, and nonlinear regression for complex relationships.
Tutorials and Workshop: Similar to the previous courses, the tutorials were given to fit the models.

3. **Nonparametric Statistics (45 hours Spring semester 2017):** This course introduces the importance of nonparametric statistics in situations where parametric assumptions, such as normality and equal variances, do not hold. It covers various nonparametric methods and their theoretical foundations, emphasizing flexibility in statistical analysis. The course explores distribution-free methods, including rank-based tests like the Wilcoxon signed-rank test, Mann-Whitney U test, and Kruskal-Wallis test for comparing groups. Additionally, it introduces goodness-of-fit tests such as the Kolmogorov-Smirnov test and AD test, which assess how well data fit a theoretical distribution. These techniques provide robust alternatives to traditional parametric tests, making them valuable tools in statistical inference and real-world applications. **Tutorials and Workshop:** Similar to the previous course, I conducted workshops and tutorials.

4. **Econometrics (45 hours Spring semester 2017):** This covers topics of fundamental and advanced econometric concepts essential for empirical research. It begins with an introduction to the nature and scope of econometrics, followed by a review of statistical inference. Then, it explores simple and multiple regression analysis, including estimation, inference, and dummy variable regression. Key econometric issues such as multicollinearity, heteroscedasticity, and autocorrelation are discussed, as well as their detection and remedies. Model specification and diagnostic testing are introduced to ensure robust analysis. The course also covers simultaneous equation models, time series econometrics, including stationary and non-stationary processes, and panel data regression models. Additionally, it discusses qualitative response regression models, making it a comprehensive resource for applied econometrics..
   Tutorials and Workshop: Similar to the previous courses, we do tutorials using R.

5. **Research Methods (45 hours Autumn semester 2018):** In this course, I provide a comprehensive guide to conducting research systematically. It begins with an introduction to research methodology, followed by defining the research problem and designing an appropriate research framework. Then explore sampling design, measurement and scaling techniques, and various data collection methods. This course also covers data processing and analysis, including parametric and non-parametric hypothesis testing. It also discusses interpretation and report writing, highlighting the importance of effectively presenting research findings. The role of computers in research is also addressed, along with an introduction to word and latex.
   Tutorials and Workshop: Students collected survey data and wrote a report.

PhD Thesis

Title: On the modeling discrete extreme values

Thesis Advisors:

1. Prof. Carlo Gaetan, Professor of Statistics at the Dipartimento di Scienze Ambientali, Informatica e Statistica - DAIS Università Ca' Foscari - Venezia, Italy.

2. Dr. Philippe Naveau, Senior Researcher at CNRS, LSCE, ESTIMR, Paris, France.

Thesis Reviewrs:

1. Prof. Manuel Scotto, Professor in Statistics at the Mathematics Department of the Instituto Superior Técnico (IST), University of Lisbon, Portugal.

2. Prof. Valérie Chavez Demoulin, Professor of Statistics at the Department of Operations, Université de Lausanne, Switzerland.

Jury Members:

1. Prof. Ilaria Prosdocimi, Associate Professor of Statistics at the Dipartimento di Scienze Ambientali, Informatica e Statistica - DAIS Università Ca' Foscari - Venezia, Italy.

2. Prof. Mauro Bernardi, Associate Professor of Statistics, Department of Statistical Sciences, University of Padova, Italy.

3. Prof. Liliane Bel, Professor of Statistics at AgroParisTech, UMR518 MIA-Paris-Saclay AgroParisTech/INRAE, France.

Thesis Summary

The statistical modeling of integer-valued extremes has received less attention than its continuous counterparts in the extreme value theory (EVT) literature. In this dissertation, we mainly focus on two problems: one, how to introduce and deal with different kinds of dependence (either its simple or temporal)

behavior over the tail when one is working with discrete threshold exceedances, and second, how to model the entire range (i.e., low, moderate and extremes) of discrete extreme data.
Firstly, to describe simple or temporal dependence in discrete exceedances above a threshold. The modeling framework is executed in two steps. In the first step, discrete exceedances are modeled through a discrete generalized Pareto distribution (DGPD), which can be obtained by mixing a Geometric variable with a Gamma distribution. In the second step, a model for discrete extreme values is built by injecting Gamma random variables or latent Gamma process via hierarchical framework, which confirms that the marginal distribution is a DGPD, as expected from classical discrete EVT. In that construction, we obtained a bivariate distribution with DGPD marginals through the Laplace transform of multivariate Gamma distribution with Gamma marginals. In addition, we further developed a bivariate geometric distribution through Farlie-Gumbel-Morgenstern Copula, mixed it into bivariate Gamma distribution, and found a bivariate distribution with DGPD marginals. In this scenario, we have two dependence parameters: one is the copula dependence parameter, and the other is linked with the layer induced through Gamma random variables associated with the hierarchical setting.
Further, we employ four distinct underlying stationary Gamma processes, each producing a different temporal dependency structure, either asymptotic independence or asymptotic dependence. Through the use of pairwise likelihoods, the proposed model is applied to real discrete time series. Observations of both series over a finite threshold have shown asymptotic independent behavior. One can use a new model for the discrete-time series, which has asymptotic-dependent behavior over the tail. In both scenarios, the proposed model is more flexible.
Secondly, selecting the optimal threshold to define exceedances remains challenging when working with discrete extreme data. Moreover, within a regression framework, the treatment of the many data points (those below the chosen threshold) is either ignored or decoupled from extremes. One possibility is to model the bulk part (observation below the threshold) and tail part (observation above the threshold) by separate models with a mixture setting. Again, an optimal threshold is needed, and this framework is computationally burdensome. Based on these considerations, we propose to make sure EVT complies by using smooth transitions between the two tails (lower and upper). By extending Generalized Additive Models (GAM) to discrete extreme responses, we are able to incorporate covariates. A GAM model quantifies the parameters of the model as functions of covariates. We also develop models with an additional parameter representing the proportion of zero values in the data in the case of zero inflation. The maximum likelihood estimation procedure is implemented for estimation purposes. With the advantage of bypassing the threshold selection step, our findings indicate that the proposed models are more flexible and robust than competing models (i.e., DGPD, Poisson distribution, and negative binomial distribution).

Research publications and preprints

Published Articles

1. **Ahmad, T**. and Arshad I. A., (2025). New flexible versions of the extended generalized Pareto model for count data. Journal of Applied statistics (*To appear*)
   https://doi.org/10.48550/arXiv.2409.18719

2. **Ahmad, T**. and Sabir S., Arshad I. A., Hasan T., & Albalawi O., (2025). Estimating Extreme Drought Risk Through Classical and Bayesian Paradigms. *International Journal of Climatology*,1-15.
   https://rmets.onlinelibrary.wiley.com/doi/epdf/10.1002/joc.8705

3. **Ahmad, T**., Gaetan, C., & Naveau P., (2024). An extended generalized Pareto regression model for count data. *Statistical Modelling* 1471082X241266729. https://doi.org/10.1177/1471082X241266729

4. Ahmad, I., **Ahmad, T**., Rehman, S. U., Almanjahie, I. M., & Alshahrani, F. (2024). A detailed study on quantification and modeling of drought characteristics using different copula families. *Heliyon* **10(3)**.
   https://doi.org/10.1016/j.heliyon.2024.e25422

5. Ahmad, I., **Ahmad, T.,** Shahzad, U., Ameer, M. A., Emam, W., Tashkandy, Y., & Badar, Z. (2024). An estimation of regional and at-site quantiles of extreme winds under flood index procedure. *Heliyon* **10(1)**. https://doi.org/10.1016/j.heliyon.2023.e23388.

6. **Ahmad, T**., Ahmad, I., Arshad, I. A., & Almanjahie, I. M. (2023). An efficient Bayesian modelling of extreme winds in the favor of energy generation. *Energy Reports* **9(1)**, 2980–2992. https://doi.org/10.1016/j.egyr.2023.01.093

7. **Ahmad, T**., Ahmad, I., Arshad, I. A., & Bianco, N. (2022). A comprehensive study on the Bayesian modelling of extreme rainfall: a case study from Pakistan. *International Journal of Climatology*, **42(1)**, 208–224. https://doi.org/10.1002/joc.7240

8. Noor, F., Masood, S., Sabar, Y., Shah, S. B. H., **Ahmad, T.**, Abdollahi, A., & Sajid, A. (2021). Bayesian analysis of cancer data using a 4-component exponential mixture model. *Computational and Mathematical Methods in Medicine*, **2021(1).** https://doi.org/10.1155/2021/6289337/

9. Cheema, A. R., Firdous, S., **Ahmad, T.**, & Imran, M. (2021). Family planning and fertility reduction in Pakistan. *Ilkogretim Online*, **20(5)**, 3617–3627.https://ilkogretim-online.org/index.php/pub/article/view/5966

10. Ahmad, I., **Ahmad, T.**, & Almanjahie, I. M. (2019). Modelling of extreme rainfall in Punjab, Pakistan using Bayesian and frequentist approach. *Applied Ecology and Environmental Research*, **17(6)**, 13729-13748. https://doi.org/10.15666/aeer/1706_1372913748

Articles Preprints

1. **Ahmad, T**. and Portier F., & Stupfler G., (2024). Logistic lasso regression with nearest neighbors for gradient-based dimension reduction. https://doi.org/10.48550/arXiv.2407.08485

2. Rehman, S. U., **Ahmad, T**., Desheng, W D., & Karamoozian A., (2024). Analyzing selected cryptocurrencies spillover effects on global financial indices: Comparing risk measures using conventional and eGARCH-EVT-Copula approaches. https://doi.org/10.48550/arXiv.2407.15766

3. T. Hasan and **Ahmad, T**., (2024).Order of Addition in Orthogonally Blocked Mixture and Component-Amount Designs. https://doi.org/10.48550/arXiv.2410.22501

4. T. Hasan and **Ahmad, T**., (2024).Order of Addition in Mixture-Amount Experiments. https://doi.org/10.48550/arXiv.2410.04864

Articles in Progress

1. **Ahmad, T**., Gaetan, C., (2024). Extreme tail flexible novel nonstationary standardized precipitation index.

2. **Ahmad, T**., Gaetan, C., (2024). A latent process model for discrete temporal extremes.

3. **Ahmad, T**., Portier, F and Stupfler, G (2024). Novel Algorithms for imbalance classification problems.

4. Shafiq, U.R., **Ahmad, T** (2024).Modeling of financial risk through extreme value based neutral networks.

5. **Ahmad, T**, Saforah S., Shafiq, U.R. (2024). Bayesian modeling of drought extremes.

Conferences & Seminars

Invited Talks

1. **Ahmad, T**., Gaetan, C., & Naveau P., (2024). An extended generalized Pareto regression model for count data. *17th International Conference of the ERCIM Working Group on Computational and Methodological Statistics (CMStatistics2024)* King's College London, UK. Date 14-16 December 2024.

2. **Ahmad, T**., Hasan T., (2023). A flexible novel extension of discrete generalized Pareto distribution. *2nd International Conference on Recent Trends in Statistics & Data Analytics, National University of Science and Technology, Islamabad*. Date, 14-15 December 2023.

3. **Ahmad, T**., (2022). Modelling the entire range of discrete extreme data. *International Conference on Recent Trends in Statistics & Data Analytics, National University of Science and Technology, Islamabad*. Date, 23 September 2022.

Contributed Talks

1. **Ahmad, T**., & Portier, F., Stupfler, G., (2024). Local logistic regression for dimension reduction in classification. *International Symposium on Nonparametric Statistics (ISNPS 2024)*, Braga, Portugal. Date 25-29 June, 2024.

2. **Ahmad, T**., & Portier, F., Stupfler, G., (2024). Dimension reduction for binary classification problems. *Causality in Extremes Workshop and Mini-Courses*, University of Geneva, Geneva, Switzerland. Date 12-16 February 2024.

3. **Ahmad, T**., & Gaetan, C., (2023). A latent process model for discrete extremes. *13th International Conference of Extreme Value Analysis 2023 (EVA2023)*, Bocconi University, Milan, Italy. Date 26-30 June 2023.

4. **Ahmad, T**., Gaetan, C., & Naveau P., (2022). Modelling of discrete extremes through extended versions of discrete generalized Pareto distribution. *15th International Conference of the ERCIM Working Group on Computational and Methodological Statistics (CMStatistics)* King's College London, UK. Date 17-19 December 2022.

Seminars

1. **Ahmad, T**., (2022). Extreme value theory and its role in the modeling of rare events. *Department of Statistics, Allama Iqbal Open University, Islamabad, Pakistan*. Date, 28 Jan 2025.

2. **Ahmad, T**., (2022). Some new versions of discrete extreme models. *Laboratoire de Mathématiques de Versailles, Versailles, France*. Date, 19 April 2022.

3. **Ahmad, T**., (2022). Some new versions of discrete extreme models. *Department of Statistical Sciences, University of Padova, Italy*. Date, 17 February 2022.

Supervising experience

2024 (PhD co-supervision): Classical and Bayesian Modeling for Droughts Risk Assessment. Sumaira Perveen (IIU, Islamabad)

Awards & Grants

University of Padova — **Padova, Italy**
CARIPARO Research Grant for PhD — *Dec-2019 to Mar-2023*

International Islamic University — **Islamabad, Pakistan**
Awarded Gold Medal in MS — *March-2019*

International Islamic University — **Islamabad, Pakistan**
Awarded Laptop by Prime Minister Laptop Scheme — *August-2015*

Other

Languages: Urdu (native), Punjabi (native), English (advanced), Italian (basic), French (basic).

Technologies: R (advanced), Python (intermediate), C++ (advanced), Julia (basic), LaTex (advanced).

Teaching Activities

I gained teaching experience as a lecturer in Statistics at Allama Iqbal Open University (AIOU), Islamabad, and the Higher Education Department (HED), Govt. of Punjab, Pakistan. During this time, I taught various courses in Statistics, which allowed me to observe student reactions and behavior. Students often inquire about the "why and how," showing a preference for concrete examples, real-world applications, and connections to their future studies. They also seek a deeper understanding of research, expecting more precise explanations of the research process and its relevance. This insight shapes my teaching approach and methodology.

I worked for two years as a Teaching Assistant Lecturer in the Department of Statistics at AIOU, Islamabad, and the Higher Education Department. I taught courses such as Statistical Methods, Regression Analysis, Econometrics, Nonparametric Statistics, Research Methodology, Statistical Theory and Statistics and Probability, which helped me learn both course design and implementation. The tutorials and exercises, combining theory with practical applications, allowed me to develop skills in structuring lessons and managing student projects. These projects also introduced me to student supervision, blending pedagogical and mathematical guidance. Practical work in the computer lab revealed that while students are familiar with computers, they struggle with statistical software and programming, highlighting the challenges in applying statistical tools effectively.

Alongside my apprenticeship, I received pedagogical and theoretical training organized by the Higher Education Department (HED), Govt. of Punjab, Pakistan. During this training, I gained insight into the diversity and complexity of colleges and universities in Pakistan, as well as the various methods of evaluating both students and teachers. I also learned about the rights and responsibilities of teachers, as well as career development. Discussion sessions with mentors, supervised by professionals, provided valuable opportunities to address concrete challenges and exchange ideas on national education. Additionally, workshops allowed me to improve my skills as a teacher-researcher, focusing on public speaking and vocal range and promoting both my research and professional identity.

I gained significant pedagogical and theoretical knowledge during these teaching years. What stands out most is the incredible support I received from various teaching colleagues I worked with. They assisted me through challenges and shared their experiences in both teaching and research. This reinforced the importance of collaboration within a teaching team. These years also introduced me to the teacher-researcher profession as I worked later for my PhD thesis research. I learned to balance two seemingly distinct research and teaching careers, in fact, complement each other. Currently, I am a Postdoc at ENSAI; I may continue my teaching role as a lecturer in the future, where I may be responsible for managing courses for master's students in data science. This new experience offers further opportunities to expand my knowledge and develop my teaching independence.

Research Activities

My research in statistics focuses on extreme value modeling to observe different kinds of dependence structures (either simple or temporal). The extreme value modeling part concentrates mainly on the tail of the discrete extreme value distribution, allowing asymptotic independence and dependence. The research further focused on modeling the entire data when selecting the optimal threshold for tail exceedance, which is not easy. My postdoc research deals with dimension reduction in classification problems and imbalanced data problems when a high enough threshold is set in response to getting events of interest (for instance, the patient diagenesis cancer in medical data example). My work includes both theoretical and applied aspects.

Research activities during PhD

My research contributes to proposing a new discrete extreme value model that can account for different types of dependence in the extremes. We link this modeling framework with a hierarchical approach, ensuring that the marginal distribution of exceedances lies in the maximum domain of attraction of discrete generalized Pareto distribution (DGPD). Therefore, the classical extremal techniques are appreciated, and a wide range of dependence can be considered at the extreme level. We can accomplish this by representing the DGPD as a mixture of geometric and gamma random variables.

Let $\{Y_t\}$ be a stationary random sequence of discrete nature. We are interested in modeling the data points above the fixed high threshold u, which basically represents the tail behavior. Similar to GPD, the discrete observations of $\{Y_t\}$ above the high threshold u follow the DGPD.

For the purpose of subsequent constructions, we use DGPD(β, α) with re-parametrization of scale and shape parameters as $\beta = \sigma/\xi$ and $\alpha = 1/\xi$. After re-parametrization, the PMF of DGPD is

$$\Pr(Y = k) = \left(1 + \frac{k}{\beta}\right)^{-\alpha} - \left(1 + \frac{k+1}{\beta}\right)^{-\alpha}, \qquad k \in \mathbb{N}_0 \tag{1}$$

By following the construction of Buddana and Kozubowski (2014), the DGPD(β, α) can be written as a mixture of Geometric and Gamma distributions that is

$$\Pr(Y = k) = \int_0^\infty \Pr(Y = k|\Lambda = \lambda) f(\lambda) d\lambda \tag{2}$$

where $\Pr(Y = k|\Lambda = \lambda)$ follow a Geometric distribution, that is, Geo(q), where $q = 1 - e^{-\lambda/\beta}$, and λ follows Gamma distribution with pdf

$$f(\lambda, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}. \tag{3}$$

We aim to work in discrete extreme value theory and keep the DGPD as a marginal distribution of Y_t. Using the standard conditioning argument, the survival function (SF) of DGPD can be derived by mixing the composition of Geometric and Gamma distributions, that is

$$\Pr(Y > k) = S(k) = \int_0^\infty e^{-\lambda k/\beta} f(\lambda) d\lambda = L^{(1)}(s)|_{(s=k/\beta)}, \tag{4}$$

where $e^{-\lambda k/\beta}$ is the SF of the Geometric distribution, and $f(\lambda)$ is a probability density function of the standard Gamma distribution, i.e., Gamma$(\alpha, 1)$. $L^{(1)}(s)|_{(s=k/\beta)} = (\beta/(\beta + k))^\alpha$ is the LT of Gamma$(\alpha, 1)$ distribution. For $\alpha > 0$, the given consideration falls in D-MDA$_\alpha$ (Hitz et al. 2024).

The interesting feature of our representation is that the bivariate distributions with discrete Pareto-type margins are easily tractable. Moreover, the following propositions will clearly define our contribution regarding bivariate distributions.

Proposition 1. *Let $\Lambda = (\Lambda_1, \Lambda_2)$ have a bivariate distribution with Gamma margins. Suppose, given $\Lambda = (\lambda_1, \lambda_2)$, $Y_i, i = 1, 2$ are independent Geometric random variable with parameter $q_i = 1 - e^{-\lambda_i/\beta}, i = 1, 2$, where $\beta > 0$. Then $\mathbf{Y} = (Y_1, Y_2)$ follows a bivariate distribution with DGPD marginals and is defined through bivariate Laplace transforms of bivariate Gamma distributions. That is, the joint survival function is written as*

$$S(k_1, k_2) = \int_0^\infty \int_0^\infty e^{-k_1\lambda_1/\beta - k_2\lambda_2/\beta} f(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2 = L^{(2)}(s_1, s_2)|_{(s_1=\frac{k_1}{\beta}, s_2=\frac{k_2}{\beta})} \tag{5}$$

where $e^{-k_1\lambda_1/\beta}$ and $e^{-k_2\lambda_2/\beta}$ are the SFs of independent Geometric random variables Y_1 and Y_2. The expression $L^{(2)}(s_1, s_2)$ is the bivariate LT of $\mathbf{\Lambda} = (\Lambda_1, \Lambda_2)$.

We use four distinct stationary gamma processes to generate a temporal layer with different kinds of dependence behavior. Interestingly, we introduce temporal dependence in the proposed representation through the gamma process $\{\Lambda_t\}$. Thus, we focus on different stationary gamma processes, which are more flexible with Markov chains and have recursive forms that may lead to different extremal dependence structures. After inducing the temporal layer via $\{\Lambda_t\}$, the original variable $\{Y_t\}$ may have a non-Markovian nature model (Bortot and Gaetan, 2014). We use the following stationary gamma processes as

1. Gaver and Lewis process (GLP): This class of process was firstly introduce by Gaver and Lewis (1980). Let

$$\begin{aligned} \Lambda_{t-1} &\sim \text{Gamma}(\alpha,\beta) \\ P_t &\sim \text{Gamma}(\alpha,1) \\ X_t|P_t &\sim \text{Poisson}\left(P_t\frac{(1-\rho)}{\rho}\right) \\ V_t|X_t &\sim \text{Gamma}(X_t,\frac{\beta}{\rho}) \\ \Lambda_t &= \rho\Lambda_{t-1}+V_t \end{aligned} \tag{6}$$

with Λ_{t-1} independent of P_t, $X_t|P_t$ and $V_t|X_t$. ρ is the dependence parameter which is the correlation between Λ_{t-1} and Λ_t, it ranges $0 \leq \rho < 1$. Given $\{\Lambda_t\}$ is stationary gamma process whose marginal distribution is Gamma(α,β).

2. Warren process (WP): The Warren process, as introduced by Warren (1992), is defined as follows. Let

$$\begin{aligned} \Lambda_{t-1} &\sim \text{Gamma}(\alpha,\beta) \\ X_t|\Lambda_{t-1} &\sim \text{Poisson}\left(\frac{\rho\Lambda_{t-1}\beta)}{1-\rho}\right) \\ \Lambda_t|X_t &\sim \text{Gamma}\left(X_t+\alpha,\frac{1-\rho}{\beta}\right) \end{aligned} \tag{7}$$

The resulting $\{\Lambda_t\}$ again stationary Markov gamma process with recursive form and follow Gamma(α,β) margins with dependence parameter ρ, it ranges $0 \leq \rho < 1$.

3. Thinned gamma process (TGP): This class of process was introduced by Wolpert (2021) by using the thinning layer generated from Beta distribution. Let

$$\Lambda_{t-1} \sim \text{Gamma}(\alpha,\beta)$$

$$B_t \sim \text{Beta}(\alpha\rho,\alpha(1-\rho)) \tag{8}$$

$$V_t \sim \text{Gamma}(\alpha(1-\rho),\beta) \tag{9}$$

$$\Lambda_t = B_t\Lambda_{t-1}+V_t$$

where $\{\Lambda_{t-1}\}$ is independent of B_t and V_t. The $\{\Lambda_t\}$ is a Markov process with gamma univariate marginal distribution Gamma(α,β) with auto-correlation ρ, it ranges $0 \leq \rho < 1$. The process of passing from $\{\Lambda_t\}$ to $B_t\Lambda_{t-1}$ is called thinning (Wolpert, 2021). Hence, $\{\Lambda_t\}$ is called thinned gamma process.

4. The Markov change-point process (MCPP): Let $\{\xi_n : n \in \mathbb{Z}\} \overset{\text{iid}}{\sim} \text{Gamma}(\alpha,\beta)$ be independent and identically distributed gamma random variables and let P_t be a standard Poisson process index by $t \in \mathbb{R}$ (so $P_0 = 0$ and $(P_t - P_{s=t-1}), \forall -\infty < s < t < \infty$, independent increments), and let

$$\Lambda_t = \xi_n, \qquad n = P_{ut}, \tag{10}$$

then each $\Lambda_t \sim \text{Gamma}(\alpha,\beta)$ and, for $t-1, t \in \mathbb{R}$, $\{\Lambda_{t-1}\}$ and $\{\Lambda_t\}$ are either identical or independent reminiscent of Metropolis Markov Chain Monte Carlo chain. Once again $\{\Lambda_t\}$ have Gamma(α,β) marginal

distribution.
It should be noted that all the processes have the same correlation function, that is

$$\mathrm{Corr}(\Lambda_t, \Lambda_{t+j}) = \rho^{|j|}$$

In addition, we will work with LTs of Λ_t and Λ_{t+j}, because of Gamma(α, β) margins, the univariate Laplace transform of $\{\Lambda_t\}$ is defined earlier in (4). The bivariate LTs are defined as

$$L_j^{(2)}(k_1, k_2) = \mathrm{E}\left(e^{-\frac{k_1\Lambda_t}{\beta} - \frac{k_2\Lambda_{t+j}}{\beta}}\right),$$

which is given by

$$L_j^{(2)}(k_1, k_2) = \left[\frac{(\beta + \rho^j k_2)\beta}{(\beta + k_2)(\beta + k_1 + \rho^j k_2)}\right]^{\alpha}, \qquad 0 < \rho < 1, \tag{11}$$

for GLP and by

$$L_j^{(2)}(k_1, k_2) = \left[\frac{\beta^2}{(k_1 + \beta)(k_2 + \beta) - \rho^j k_1 k_2}\right]^{\alpha}, \qquad 0 < \rho < 1 \tag{12}$$

for WP, and by

$$L_j^{(2)}(k_1, k_2) = \left[\frac{\beta^{\alpha(2-\rho^j)}}{(\beta + k_1)^{\alpha(1-\rho^j)}(\beta + k_1 + k_2)^{\alpha\rho^j}(\beta + k_2)^{\alpha(1-\rho^j)}}\right], \quad 0 < \rho < 1 \tag{13}$$

for TGP, and by

$$L^{(2)}(k_1, k_2) = \left[\frac{\rho^j\beta^{\alpha}}{(\beta + k_1 + k_2)^{\alpha}} + \frac{(1-\rho^j)\beta^{2\alpha}}{(\beta + k_1)^{\alpha}(\beta + k_2)^{\alpha}}\right], \qquad 0 < \rho < 1, \tag{14}$$

for MCPP, respectively.
There is a common measure of dependence in extremes called extremogram, which describes the conditional probability that one random variable will be extreme when the other is extreme (Davis and Mikosch, 2009; Chavez-Demoulin and Davison, 2012). For a strictly stationary $\mathbb{N}_0^d$ integer-valued time series (Y_t), by following Davis et al. (2012) the extremogram is defined for two sets A and B bounded away from zero by following as

$$\rho_k^{A,B} = \lim_{n\to\infty} P(n^{-1}Y_{t+k} \in B | n^{-1}Y_t \in A), \qquad k = 0, 1, 2, \ldots \tag{15}$$

with the given limit exists. Since A and B are bounded away from zero, the events $n^{-1}Y_t \in A$ and $n^{-1}Y_{t+k} \in B$ are becoming extreme in a sense the probabilities of these events are converging to zero with $n \to \infty$. For the special choice in the $d = 1$ case of $A = B = (1, \infty)$, the extremogram reduces to the tail dependence coefficient (χ_k)

$$\chi_k = \lim_{n\to\infty} \Pr(Y_{t+k} > n | Y_t > n), \tag{16}$$

used in EVT and quantitative risk management (McNeil et al., 2015). The variables Y_t and Y_{t+k} are said to be asymptotically independent when $\chi_k = 0$ and dependent otherwise. Combining all four processes defined for $\{\Lambda_t\}$ with our model through Proposition 1, the proposed model leads to different dependence structures among Y_t. In response, GLP and MCPP induce asymptotic dependence for all $0 < \rho \leq 1$. The theoretical results of χ corresponding to GLP and MCPP are derived $[\rho^j/(1 + \rho^j)]^{\alpha}$ and $[\rho^j/2^{\alpha}]$, respectively. The non-zero χ indicates that the GLP and MCPP models have asymptotic-dependent

behavior. On the other hand, WP and TGP induce asymptotic Independence in Y_t. The theoretical result of the χ measure corresponding to WP and TGP converges to zero and clearly shows asymptotic independent behavior.
As for the practical implementation of the proposed procedure, we propose to apply the likelihood inference for hierarchical models requires approximating the n-fold integral, that is,

$$L_n(\theta) = \int \left[\prod_{t=1}^{n} \{(1 - \exp(-\lambda_t/\beta) \exp(-\lambda_t k_t/\beta)\} f(\lambda_1, \ldots, \lambda_n; \alpha, \beta, \rho) \right] d\lambda_1, \ldots, d\lambda_n \qquad (17)$$

where $\theta = (\alpha, \beta, \rho)$ and $f(\lambda_1, \ldots, \lambda_n; \alpha, \beta, \rho)$ is the joint density function of $\Lambda_1, \ldots, \Lambda_n$. The formula (17) may proceed further via the filtering algorithm. In addition to its drawbacks, the filtering algorithm propagates numerical errors through nested integrals (Pedeli and Varin, 2020). In light of this, evaluating $L_n(\theta)$ is not feasible because of the complex integral involved in (17). The following pairwise log-likelihood (PL) replaces the full likelihood in this situation:

$$pl_n(\theta) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{min(i+\Delta, n)} \log \Pr(k_i, k_j; \theta) \qquad (18)$$

where $\Pr(k_i, k_j)$ is the joint PMF of (Y_i, Y_j) and $1 \leq \Delta \leq n-1$ is the constant which defines the maximum lag. We will compute the pairwise likelihood for the Δ order using all pairs of observations with lag distances. Compared to the ordinary likelihood, the pairwise likelihood significantly reduces computational cost. Moreover, when the pairs (y_i, y_j) are treated as independent, the PL viewed an example of composite likelihood (Lindsay, 1988; Bortot and Gaetan (2014). Simulation studies and real applications prove that GLP and MCPP behave asymptotic-dependent over the tail, while WP and TGP show asymptotic Independence.

Modeling entire range of data:
As mentioned before the statistical modeling of discrete extremes has received less attention than their continuous counterparts in EVT literature. One approach to the transition from continuous to discrete extremes is the modeling of threshold exceedances of integer random variables by the discrete version of the generalized Pareto distribution. However, the optimal choice of thresholds defining exceedances remains a problematic issue. Moreover, in a regression framework, the treatment of the majority of non-extreme data below the selected threshold is either ignored or separated from the extremes. To tackle these issues, we expand on the concept of employing a smooth transition between the bulk and the upper tail of the distribution. In the case of zero inflation, we also develop models with an additional parameter. To incorporate possible predictors, we relate the parameters to additive smoothed predictors via an appropriate link, as in the generalized additive model (GAM) framework. A penalized maximum likelihood estimation procedure is implemented.

Modeling framework
The distribution of exceedances (i.e., the amount of data that appears over a given high threshold) is often approximated by the Generalized Pareto distribution (GPD) defined by its CDF as

$$F(z; \sigma, \xi) = \begin{cases} 1 - (1 + \xi z/\sigma)_+^{-1/\xi} & \xi \neq 0 \\ 1 - \exp(-z/\sigma) & \xi = 0 \end{cases}, \qquad (19)$$

where $(a)_+ = \max(a, 0)$. The $\sigma > 0$ and $-\infty < \xi < +\infty$ represent the scale and shape parameters of the distribution, respectively.
More precisely let X be a random variable taking values in $[0, x_F)$ where $x_F \in (0, \infty) \cup \{\infty\}$ Suppose that there exists a strictly positive sequence a_u such that the distribution of $a_u^{-1}(X - u)|X \geq u$ weakly

converge to a non-degenerate probability distribution on $[0, \infty)$ as $u \to x_F$, then this distribution is the GPD (Balkema and De Haan, 1974). Thus, for large u,

$$\Pr(X - u > x | X \geq u) = \Pr(a_u^{-1}(X - u) > a_u^{-1} x | X \geq u) \approx 1 - F(x; a_u\sigma, \xi).$$

The shape parameter, ξ, defines the tail behavior of the GPD. If $\xi < 0$, the upper tail is bounded. If $\xi = 0$, we have the exponential distribution, where all moments are finite. If $\xi > 0$, the upper tail is unbounded and the higher moments ultimately become infinite. The three defined cases are labeled "short-tailed", "light-tailed", and "heavy-tailed", respectively. These categorizations enhance the flexibility of the GPD and underscore its adaptability to various modeling scenarios.
Using the GPD to approximate the distribution tail for discrete data can be inappropriate. These authors proposed to approximate the distribution tail of a count random variable Y by discretizing the CDF defined by (19) and, for large u,

$$\Pr(Y - u = k | Y \geq u) = F(k + 1; \sigma, \xi) - F(k; \sigma, \xi), \qquad k \in \mathbb{N}_0, \tag{20}$$

with $\sigma > 0$ and $\xi \geq 0$. The distribution is called discrete GPD (DGPD).
A drawback of GPD in the continuous case is that it only models observations that occur above a certain high threshold. This imposes an artificial dichotomy in the data (i.e., observations are either below or above the threshold), and finding the optimal threshold remains complex for practitioners. The choice becomes more complicated when the observations feature a substantial number of ties.
We use the idea of the integral transformation to simulate GPD random draws, i.e. $F^{-1}_{\sigma,\xi}(U)$, where $U \sim \mathcal{U}(0, 1)$ represents a uniformly distributed random variable on $(0, 1)$ and $F^{-1}_{\sigma,\xi}$ denotes the inverse of the CDF (19). This leads to the family of distribution for the random variable

$$Z = F^{-1}_{\sigma,\xi}\left(G^{-1}(U)\right), \tag{21}$$

where G is a CDF on $[0, 1]$ and $U \sim \mathcal{U}(0, 1)$. The CDF of Z is $G(F(z; \sigma, \xi))$, which is called extended GPD. The key problem is finding a function G that preserves the upper tail behavior with shape parameter ξ and controls the lower tail behavior. Naveau et al. (2016) defined restrictions for the validity of G functions. For instance, the tail of G denoted by $\bar{G} = 1 - G$ has to satisfy

$$\lim_{u \to 0} \frac{\bar{G}(1 - u)}{u} = a, \text{ for some finite } a > 0 \text{ (upper tail behavior)}, \tag{22}$$

$$\lim_{u \to 0} \frac{G(u)}{u^\kappa} = c, \text{ for some finite } c > 0 \text{ (lower tail behavior)}. \tag{23}$$

Four parametric examples for G have been considered. We follow the same idea and define the probability mass function (pmf) for the count variable Y as

$$\Pr(Y = y) = G\left(F\left(y + 1; \sigma, \xi\right)\right) - G\left(F\left(y; \sigma, \xi\right)\right), \qquad y \in \mathbb{N}_0. \tag{24}$$

The distribution defined by (24) is referred to as the discrete extended generalized Pareto distribution (DEGPD). The explicit formula of CDF of DEGPD is developed as

$$\Pr(Y \leq y) = G\left(F\left(y + 1; \sigma, \xi\right)\right) \tag{25}$$

and the quantile function is derived as

$$q_p = \begin{cases} \left\lceil \frac{\sigma}{\xi} \left\{ \left(1 - G^{-1}(p)\right)^{-\xi} - 1 \right\} \right\rceil - 1, & \text{if } \xi > 0 \\ \left\lceil -\sigma \log\left(1 - G^{-1}(p)\right) \right\rceil - 1, & \text{if } \xi = 0 \end{cases} \tag{26}$$

with $0 < p < 1$. In this paper, we use four parametric expressions of $G(\cdot)$ (**?**), namely

Model (i): $G(u;\psi) = u^{\kappa}$, $\psi = \kappa > 0$;

Model (ii): $G(u;\psi) = 1 - D_{\delta}\{(1-u)^{\delta}\}$, $\psi = \delta > 0$ where D_{δ} is the CDF of a Beta random variable with parameters $1/\delta$ and 2, that is:

$$D_{\delta}(u) = \frac{1+\delta}{\delta} u^{1/\delta}\Big(1 - \frac{u}{1+\delta}\Big);$$

Model (iii): $G(u;\psi) = [1 - D_{\delta}\{(1-u)^{\delta}\}]^{\kappa/2}$, $\psi = (\delta, \kappa)$ with $\delta > 0$ and $\kappa > 0$;

Model (iv): $G(u;\psi) = pu^{\kappa_1} + (1-p)u^{\kappa_2}$, $\psi = (p, \kappa_1, \kappa_2)$ with $\kappa_2 \geq \kappa_1 > 0$ and $p \in (0,1)$.

Zero-inflation and regression modeling

As we have seen in many real data problems, many zeros can be found in various real data sets. In that case, the current model with a flexible lower and upper tail cannot be adjusted for the excessive zeros. We follow Lambert (1992) and change DEGPD's pmf to

$$\Pr(Y = y) = \begin{cases} \pi + (1-\pi)G(F\,(1,\sigma,\xi)\,;\psi) & y = 0 \\ (1-\pi)\,[G\,(F(y+1,\sigma,\xi);\psi) - G\,(F(y,\sigma,\xi);\psi))] & y \in \mathbb{N} \end{cases} \tag{27}$$

where $0 \leq \pi \leq 1$ is the mixing proportion, determining from which state Y is generated. In the following, we coin (27) as the ZIDEGPD model.

Suppose that $\boldsymbol{x} \in \mathbb{R}^q$ is a vector of covariates measured with Y. By allowing the parameters to depend on covariates, we extend the pmf (27) to the zero-inflated count regression setting. More specifically, we identify the vector of parameters (ξ, σ, ψ, π) with $\theta = (\theta_1, ..., \theta_d)$. The parameters of the distribution of Y can depend on the covariates $\boldsymbol{x}$, i.e. $\theta(\boldsymbol{x}) = (\theta_1(\boldsymbol{x}), \ldots, \theta_d(\boldsymbol{x}))$. To relate the distribution parameters $(\theta_1(\boldsymbol{x}), \ldots, \theta_d(\boldsymbol{x}))$ to the covariates, we consider additive predictors of the form

$$\eta_i(\boldsymbol{x}) = s_{i1}(\boldsymbol{x}) + \cdots + s_{iJ_i}(\boldsymbol{x}), \qquad i = 1, \ldots, d, \tag{28}$$

where $s_{i1}(\cdot), \ldots, s_{iJ_i}(\cdot)$ are smooth functions of the covariates $\boldsymbol{x}$. The predictors are linked to the parameters via known monotonic and twice differentiable link functions $h_i(\cdot)$.

$$\theta_i(\boldsymbol{x}) = h_i(\eta_i(\boldsymbol{x})), \qquad i = 1, \ldots, d. \tag{29}$$

For instance, we use the following linking functions for the model (i) associated with $G(u;\psi) = u^{\kappa}$. The parameters can be written as

$$\xi(\boldsymbol{x}) = \exp(\eta_{\xi}(\boldsymbol{x})),\ \sigma(\boldsymbol{x}) = \exp(\eta_{\sigma}(\boldsymbol{x})),\ \kappa(\boldsymbol{x}) = \exp(\eta_{\kappa}(\boldsymbol{x})),\ \pi(\boldsymbol{x}) = \exp\left(\frac{\eta_{\pi}(\boldsymbol{x})}{1+\eta_{\pi}(\boldsymbol{x})}\right).$$

The functions $s_{ij}(\cdot)$ in (28) are approximated by a set of K_{ij} basis functions $\{B_{k,ij}(\boldsymbol{x})\, k = 1, \ldots, K_{ij}\}$, namely

$$s_{ij}(\boldsymbol{x}) = \sum_{k=1}^{K_{ij}} \beta_{ij,k} B_k(\boldsymbol{x}). \tag{30}$$

The basis functions can be of different types. The basis function expansions can be written as $s_{ij}(\boldsymbol{x}) = \boldsymbol{t}_{ij}(\boldsymbol{x})^T \boldsymbol{\beta}_{ij}$ where $\boldsymbol{t}_{ij}(\boldsymbol{x})$ is still a vector of transformed covariates that depends on the basis functions and $\boldsymbol{\beta}_{ij} = (\beta_{ij,1}, \ldots, \beta_{ij,K_{ij}})^T$ is a parameter vector to be estimated.

The penalized maximum likelihood estimation (MLE) method is used to estimate the parameters of the proposed models. More precisely, let $y_1, \ldots, y_n$ be n independent observations from (24) and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ the related covariates. The log-likelihood function is given by

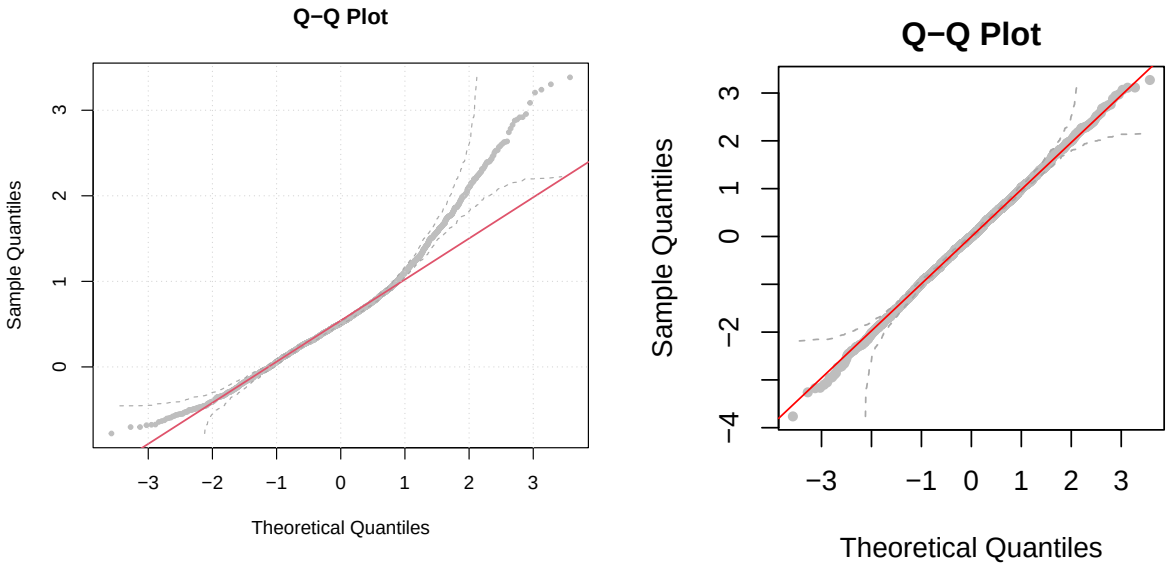
$$\begin{aligned} l(\boldsymbol{\beta}) = & \sum_{i=1}^{n} I_{\{0\}}(y_i) \log\left[\pi(\boldsymbol{x}_i) + (1-\pi(\boldsymbol{x}_i)) G(F(1; \sigma(\boldsymbol{x}_i), \xi(\boldsymbol{x}_i)); \psi(\boldsymbol{x}_i))\right] \\ & + \sum_{i=1}^{n} (1 - I_{\{0\}}(y_i)) \log(1-\pi(\boldsymbol{x}_i))) \times \\ & \left[G(F(y_i+1; \sigma(\boldsymbol{x}_i), \xi(\boldsymbol{x}_i)); \psi(\boldsymbol{x}_i)) - G(F(y_i; \sigma(\boldsymbol{x}_i), \xi(\boldsymbol{x}_i)); \psi(\boldsymbol{x}_i))\right], \end{aligned} \tag{31}$$

where $I_A(\cdot)$ is the indicator function of the set A. To ensure regularization of the functions $s_{ij}(\boldsymbol{x})$, so-called penalty terms are added to the objective log-likelihood function. Usually, the penalty for each function $s_{ij}(\boldsymbol{x})$ is a quadratic penalty $\boldsymbol{\beta}_{ij}^T \boldsymbol{P}_{ij}(\boldsymbol{\lambda}_{ij}) \boldsymbol{\beta}_{ij}$ where $\boldsymbol{P}_{ij}(\boldsymbol{\lambda}_{ij})$ is a known semi-definite matrix and the vector $\boldsymbol{\lambda}_{ij}$ regulates the amount of smoothing needed for the fit. A special case that we use in the real data application is when $\boldsymbol{P}_{ij}(\boldsymbol{\lambda}_{ij}) = \lambda_{ij} \boldsymbol{P}_{ij}$, for a scalar $\lambda_{ij} > 0$ and a semi-definite matrix $\boldsymbol{P}_{ij}$. The entries of the penalty matrix P_{ij} are the integrals of the products of the second derivatives of pairs of cubic spline functions, see Wood (2011) for more details. The penalized log-likelihood function for the latter models reads:

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{J_i} \lambda_{ij} \boldsymbol{\beta}_{ij}^T \boldsymbol{P}_{ij} \boldsymbol{\beta}_{ij}. \tag{32}$$

We apply the restricted maximum likelihood (REML) approach to estimate $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{\lambda}_{ij}$ following Wood (2011). Our current implementation exploits the R package `evgam` and adds two new families of distributions named `degpd` and `zidegpd`. Note that within the `degpd` family, all four models for $G(;\psi)$ have been implemented. By contrast, for the `zidegpd` family, only models (i), (ii), and (iii) have been implemented since the `evgam` package allows the simultaneous estimation of only five distribution parameters $\theta(\boldsymbol{x}) = (\theta_1(\boldsymbol{x}), \ldots, \theta_5(\boldsymbol{x}))^T$.

Results



The Model proposed in (27) corresponding to $G(u;\psi) = u^{\kappa}$, $\psi = \kappa > 0$ is fitted to the real counts data of Avalanches of French Alps with environmental covariates by using penalized likelihood defined in

equation (32). The left figure represents the existing Zero-Inflated Negative Binomial Regression (ZINBR), while the right figure corresponds to the proposed model. ZINBR fails to capture both tail behaviors accurately, whereas the proposed model effectively accounts for extreme tails and properly estimates zero inflation.

Research activities during Post Doc

This research proposes a new dimension reduction framework tailored to classification, handling dimension reduction subspaces not limited to a single projection and able to perform variable selection while dealing with the curse of dimensionality. Assume that Y satisfies $\pi(X) := \mathbb{P}(Y = 1|X) = g(\beta^T X)$, for some $\beta \in \mathbb{R}^{p\times d}$, $d \leq p$ and a measurable function $g : \mathbb{R}^d \to [0, 1]$. The function g, treated as a nuisance parameter, and the matrix β are both unknown. If we had $d = 1$ and $g(t) = \mathrm{expit}(t) := \exp(t)/(1+\exp(t))$, then we would have $\mathrm{logit}(\pi(X)) := \log(\pi(X)/(1 - \pi(X))) = \beta^T X$, and $\beta \in \mathbb{R}^p$ would be exactly the gradient of $\ell := \mathrm{logit}(\pi)$. Our main idea is that, if the conditional probability π is sufficiently regular at a given point x then $\ell(X)$ can be approximated by a linear function of $X - x$ in a neighborhood of x, whose intercept $a = a(x)$ is just $\ell(x)$ and whose gradient $b = b(x) = \nabla\ell(x)$ belongs to the vector space $\mathrm{span}(\beta)$ that is spanned by the columns of β. We then construct, at each point x, a nearest-neighbor, penalized, local logistic (and hence convex) loss function depending on two arguments $a \in \mathbb{R}$ and $b \in \mathbb{R}^p$. The weak convergence of the estimator of $(\ell(x), \nabla\ell(x))$ hence constructed follows, under very mild conditions, from a new general empirical process theory result on a family of nearest-neighbor estimators that is of independent interest. The obtained estimation rates for $\ell(x)$ and $\nabla\ell(x)$ achieve the minimax optimal rates of convergence. In particular, our theoretical results compare favorably, where the gradient is estimated using a *Reproducing Kernel Hilbert Spaces* (RKHS) technique (see below our Corollary 3 for a precise comparison).

Main results

Let $Y \in \{0, 1\}$ be a binary response variable, with random covariate $X \in \mathbb{R}^p$. The data is assumed to be made of independent copies (Y_i, X_i) of the random pair (Y, X), for $1 \leq i \leq n$. Suppose that for an x in the support S_X of X (assumed to have nonempty interior), $\pi(x) = \mathbb{P}(Y = 1|X = x) = g(\beta^T x)$, for some unknown $\beta \in \mathbb{R}^{p\times d}$, $d \leq p$ and an unknown measurable function $g : \mathbb{R}^d \to [0, 1]$. The Kullback-Leibler divergence of the Bernoulli distribution with parameter $\pi \in (0, 1)$ from the Bernoulli distribution with parameter $q \in (0, 1)$ is

$$D_{\mathrm{KL}}(\pi||q) = \pi \log\left(\frac{\pi}{q}\right) + (1 - \pi) \log\left(\frac{1 - \pi}{1 - q}\right).$$

Gibbs' theorem entails $D_{\mathrm{KL}}(\pi||q) \geq 0$, with equality if and only if $\pi = q$. Equivalently, the cross-entropy function

$$H(\pi, q) = D_{\mathrm{KL}}(\pi||q) - \{\pi \log(\pi) + (1 - \pi) \log(1 - \pi)\} = -\pi \log(q) - (1 - \pi) \log(1 - q),$$

obtained by adding to $D_{\mathrm{KL}}(\pi||q)$ the entropy of the Bernoulli distribution with parameter π, is minimal if and only if $\pi = q$. In the classification setup, where the random covariate X has distribution P_X, the natural loss function to consider is the integrated cross-entropy

$$\int_{x\in S_X} H(\pi(x), q(x))P_X(\mathrm{d}x) = -\mathbb{E}[Y \log(q(X)) + (1 - Y) \log(1 - q(X))]$$

viewed as a function of the map $x \mapsto q(x)$. When $g = g_0$ is known, one should search for a function q of the form $x \mapsto g_0(b^T X)$, for some $b \in \mathbb{R}^{p\times d}$, meaning that one should minimize $-\mathbb{E}[Y \log(g_0(b^T X)) + (1 - Y) \log(1 - g_0(b^T X))]$ as a function of b. This suggests the estimator

$$\arg\max_{b\in\mathbb{R}^{p\times d}} \sum_{i=1}^{n} Y_i \log(g_0(b^T X_i)) + (1 - Y_i) \log(1 - g_0(b^T X_i)).$$

For $d = 1$ and $g_0 = \mathrm{expit}$, this is the logistic regression estimator, and the empirical integrated cross-entropy loss is the negative conditional log-likelihood in the logistic model.
When g is unknown, the logistic regression estimator is not guaranteed to be a consistent estimator of β. However, if π is differentiable at x, then the gradient $\nabla\pi(x)$ of π at x satisfies $\nabla\pi(x) = \beta\nabla g(\beta^T x)$. As such, $\nabla\pi(x) \in \mathrm{span}(\beta)$. If moreover $0 < \pi(x) < 1$, then likewise

$$\nabla\ell(x) = \frac{1}{\pi(x)(1-\pi(x))}\nabla\pi(x) \in \mathrm{span}(\beta), \text{ where } \ell = \mathrm{logit}\,\pi.$$

To recover $\mathrm{span}(\beta)$, it then suffices to estimate enough gradients of the form $\nabla\ell(x_j)$. Now a Taylor expansion of ℓ around x yields $\pi(y) \approx \mathrm{expit}(a + b^T(y-x))$, with $a = \mathrm{logit}(\pi(x)) = \ell(x)$ and $b = \nabla\,\mathrm{logit}(\pi(x)) = \nabla\ell(x)$, for y close to x. This suggests that the logistic regression estimator can still be used to produce an estimator of $\nabla\ell(x)$ if it is restricted to data points close enough to x. Namely, one should expect the local integrated cross-entropy

$$-\mathbb{E}[Y\log(\mathrm{expit}(a + b^T(X-x))) + (1-Y)\log(1-\mathrm{expit}(a+b^T(X-x)))\,|\,X \in B(x,\varepsilon)]$$

to be minimized at $(a_\varepsilon(x), b_\varepsilon(x)) \approx (\ell(x), \nabla\ell(x))$ as $\varepsilon \downarrow 0$, where $B(x,\varepsilon)$ is the closed Euclidean ball with center x and radius ε in $\mathbb{R}^p$. This is the cornerstone of our methodology.

Nearest-neighbor penalized local logistic estimator

The localization induced by conditioning upon covariate values belonging to $B(x,\varepsilon)$ is reproduced empirically using the well known nearest-neighbor procedure. Fix $x \in \mathbb{R}^p$ and let $N_k(x) \subset \{1,\dots,n\}$ be the set gathering the indices of the k-nearest neighbors X_i of the point x; we shall assume in the theory below that the distribution of X has a density w.r.t. Lebesgue measure, so that ties will not happen with probability 1 and $N_k(x)$ is well-defined. The empirical counterpart of the local integrated cross-entropy is then

$$-\frac{1}{k}\sum_{i\in N_k(x)} Y_i\log(\mathrm{expit}(a+b^T(X_i-x))) + (1-Y_i)\log(1-\mathrm{expit}(a+b^T(X_i-x))).$$

Introducing the LASSO penalty $\lambda\|b\|_1$, where $\|\cdot\|_1$ is the ℓ_1-norm on $\mathbb{R}^p$, and rescaling, we naturally obtain a nearest-neighbor, penalized, local logistic estimator as

$$(\widehat{a}_n(x), \widehat{b}_n(x)) = \underset{(a,b)\in\mathbb{R}\times\mathbb{R}^p}{\arg\max}\,\{L_n(a,b) - \lambda\|b\|_1\} \tag{33}$$

with

$$\begin{aligned} L_n(a,b) &= \sum_{i\in N_k(x)} Y_i\log(\mathrm{expit}(a+b^T(X_i-x))) + (1-Y_i)\log(1-\mathrm{expit}(a+b^T(X_i-x))) \\ &= \sum_{i\in N_k(x)} Y_i(a+b^T(X_i-x)) - \log(1+\exp(a+b^T(X_i-x))). \end{aligned} \tag{34}$$

Our main theoretical result is that one can obtain the asymptotic distribution of this pair of estimators under very weak assumptions on the distribution of X and the conditional distribution of $Y|X=x$. We spell out these conditions and their interpretation below.

(A1) The distribution of X has a continuous density f_X with respect to the Lebesgue measure on $\mathbb{R}^p$ and $f_X(x) > 0$.

(A2) The function $\pi : \mathbb{R}^p \to [0,1]$ is twice differentiable with continuous second order derivatives at x and such that $\pi(x) \in (0,1)$.

Assumption (A1) ensures that there are enough points around x for the nearest-neighbor procedure to work. It also ensures the good probabilistic behavior of the bandwidth

$$\widehat{\tau}_{n,k}(x) := \inf\left\{\tau \geq 0 : \sum_{i=1}^{n} \mathbb{1}_{B(x,\tau)}(X_i) \geq k\right\}$$

corresponding to the smallest radius $\tau \geq 0$ such that the ball $B(x,\tau)$ contains at least k points from the sample. Actually, the fact that X has a continuous distribution w.r.t. Lebesgue measure yields

$$\begin{aligned} &L_n(a,b) \\ &= \sum_{i=1}^{n} [Y_i \log(\operatorname{expit}(a + b^T(X_i - x))) + (1 - Y_i)\log(1 - \operatorname{expit}(a + b^T(X_i - x)))]\mathbb{1}_{B(x,\widehat{\tau}_{n,k}(x))}(X_i). \end{aligned}$$

It also turns out that if $k = k_n \to \infty$ with $k/n \to 0$, then under (A1), $\widehat{\tau}_{n,k}(x)/\tau_{n,k}(x) \to 1$ in probability, where, if V_p denotes the volume of the Euclidean unit ball in $\mathbb{R}^p$,

$$\tau_{n,k}(x) = \left(\frac{k}{n}\frac{1}{f_X(x)V_p}\right)^{1/p}.$$

See Lemma 1 in Portier (2021). It is then reasonable to write, for n large enough,

$$\begin{aligned} &L_n(a,b) \approx \overline{L}_n(a,b) \\ &= \sum_{i=1}^{n} [Y_i \log(\operatorname{expit}(a + b^T(X_i - x))) + (1 - Y_i)\log(1 - \operatorname{expit}(a + b^T(X_i - x)))]\mathbb{1}_{B(x,\tau_{n,k}(x))}(X_i). \end{aligned}$$

The asymptotic behavior of $\overline{L}_n(a,b)$ is much easier to study than that of $L_n(a,b)$, since it is a sum of independent and identically distributed random variables; like $L_n(a,b)$, it defines a concave objective function and therefore one should expect that (up to technical details) the asymptotic behavior of $(\widehat{a}_n(x), \widehat{b}_n(x))$ will follow from the pointwise convergence of $\overline{L}_n(a,b)$. The key result in order to make this intuition rigorous.

Because of the LASSO penalty term in (33), the asymptotic distribution of $\widehat{b}_n(x) - \nabla\ell(x)$ will depend on the so-called local active set associated to $\nabla\ell(x)$, defined as the set of indices j such that $\nabla\ell_j(x) \neq 0$. This is in line with the asymptotic distribution obtained for the standard least squares LASSO regression estimator. Let, for any real number t, the quantity $\operatorname{sgn}(t) = \mathbb{1}_{[0,\infty)}(t) - \mathbb{1}_{(-\infty,0)}(t)$ be the sign of t, that is, $\operatorname{sgn}(t) = 1$ when $t \geq 0$ and -1 otherwise. Finally, define

$$\Gamma(x) = \pi(x)(1 - \pi(x)) \begin{pmatrix} 1 & 0_p^T \\ 0_p & \frac{1}{p+2} I_p \end{pmatrix}$$

where 0_p is the zero vector in $\mathbb{R}^p$ and I_p denotes the identity matrix of order p. Our first main result provides the limiting distribution of the pair $(\widehat{a}_n(x), \widehat{b}_n(x))$. In this result, assumption (A2) ensures that the gradient of $\ell = \operatorname{logit}\pi$ is well-defined, with the second order derivatives of π and hence of ℓ coming into play when evaluating the bias term incurred when localizing. Denote by $\Delta\pi(x)$ the Laplacian of π at x and by $\|\cdot\|_2$ the Euclidean norm.

Theorem 2 (Convergence of nearest-neighbor penalized local logistic regression estimators)**.** *Suppose that (A1) and (A2) are fulfilled. If $k := k_n \to \infty$ and $\lambda := \lambda_n$ are such that $k^{1+p/2}/n \to \infty$, $k^{1+p/4}/n$ is*

bounded and $n\lambda^p/k^{1+p/2} \to c \in [0,\infty)$ *then we have*

$$\begin{pmatrix} \sqrt{k}(\widehat{a}_n(x) - \ell(x)) \\ \tau_{n,k}(x)\sqrt{k}(\widehat{b}_n(x) - \nabla\ell(x)) \end{pmatrix}$$
$$\stackrel{\mathrm{d}}{=} \underset{u=(u_0,u_1,\ldots,u_p)^T\in\mathbb{R}^{p+1}}{\arg\max} \left\{ u^T(W_n(x) + T_n(x)) - \frac{1}{2}u^T\Gamma(x)u \right.$$
$$\left. - (cf_X(x)V_p)^{1/p} \left(\sum_{j=1}^{p} \mathrm{sgn}(\nabla\ell_j(x))u_j \mathbb{1}_{\{\nabla\ell_j(x)\neq 0\}} + |u_j|\mathbb{1}_{\{\nabla\ell_j(x)=0\}} \right) \right\} + o_{\mathbb{P}}(1)$$

with $W_n(x) \xrightarrow{\mathrm{d}} \mathcal{N}(0,\Gamma(x))$ *and*

$$T_n(x) = \tau_{n,k}^2(x)\sqrt{k}\left(\frac{1}{2(p+2)} \left(\Delta\pi(x) - \frac{1-2\pi(x)}{\pi(x)(1-\pi(x))}\|\nabla\pi(x)\|_2^2 \right) \begin{pmatrix} 1 \\ 0_p \end{pmatrix} + o_{\mathbb{P}}(1) \right).$$

The conditions on k and λ in Theorem 2 are equivalent to assuming that $\tau_{n,k}(x)\sqrt{k} \to \infty$, $\tau_{n,k}^2(x)\sqrt{k}$ is bounded and $\lambda/(\tau_{n,k}(x)\sqrt{k})$ converges to the finite constant $(cf_X(x)V_p)^{1/p}$. Condition $\tau_{n,k}(x)\sqrt{k} \to \infty$ is necessary in order to be able to write a Taylor expansion of the penalized component in the loss function. In pointwise results on local linear kernel quasi-maximum likelihood estimation with one-dimensional covariates, Theorem 1a in Fan et al. (1995) requires $\sqrt{nh^3} = h\sqrt{nh} \to \infty$, where h is the kernel bandwidth; note that h and $\tau_{n,k}(x)$ play the same role and that, for $p = 1$, the kernel regression analogue of k is a quantity proportional to nh, so that conditions $h\sqrt{nh} \to \infty$ and $\tau_{n,k}(x)\sqrt{k} \to \infty$ are indeed analogous, and constitute the assumptions required in order to ensure consistency of the local linear estimators of ℓ and its gradient. Of course, condition $\tau_{n,k}(x)\sqrt{k} \to \infty$ is automatically satisfied if $\tau_{n,k}^2(x)\sqrt{k}$ converges to a finite positive limit, namely, when the nonparametric bias-variance tradeoff is achieved and the optimal rate of convergence of the local linear estimator is found. Also note that condition $k/n \to 0$, which is standard in nearest-neighbor estimation, follows from assuming that $k \to \infty$ and $\tau_{n,k}^2(x)\sqrt{k}$ is bounded. It follows that, under conditions (A1) and (A2) plus classical conditions linking the number of nearest neighbors and the penalizing constant λ, the two estimators $\widehat{a}_n(x)$ and $\widehat{b}_n(x)$ converge respectively at the rate $1/\sqrt{k}$ and $1/(\tau_{n,k}(x)\sqrt{k})$. Note that for non-penalized local logistic regression ($\lambda = 0$ and thus $c = 0$), the result of Theorem 2 is just

$$\begin{pmatrix} \sqrt{k}(\widehat{a}_n(x) - \ell(x)) \\ \tau_{n,k}(x)\sqrt{k}(\widehat{b}_n(x) - \nabla\ell(x)) \end{pmatrix} \stackrel{\mathrm{d}}{=} \mathcal{N}(0,\Gamma^{-1}(x)) + O_{\mathbb{P}}(\tau_{n,k}^2(x)\sqrt{k}) + o_{\mathbb{P}}(1).$$

In particular, if $k\tau_{n,k}^4(x) \to 0$, a straightforward application of the delta-method yields

$$\sqrt{k}(\mathrm{expit}(\widehat{a}_n(x)) - \pi(x)) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \pi(x)(1-\pi(x))\right)$$

as expected from standard maximum likelihood theory when the logistic regression model is valid. Condition $k\tau_{n,k}^4(x) \to 0$, which makes the bias term $T_n(x)$ vanish asymptotically, again has a straightforward analogue in local linear kernel quasi-maximum likelihood estimation; for $p = 1$, the corresponding condition is $nh^5 \to 0$, which is exactly the bias condition necessary to eliminate the smoothing bias term in Theorem 1a in Fan et al. (1995).

Another immediate corollary of Theorem 2 can also be given on the estimation rate of the gradient $\nabla\ell(x)$ by $\widehat{b}_n(x)$.

Corollary 3 (Rate of convergence of the gradient estimator)**.** *Under the conditions of Theorem 2,* $\widehat{b}_n(x)$ *is a consistent estimator of* $\nabla\ell(x)$*, and*

$$\widehat{b}_n(x) - \nabla\ell(x) = O_{\mathbb{P}}\left(\frac{1}{\tau_{n,k}(x)\sqrt{k}} \right) + O_{\mathbb{P}}(\tau_{n,k}(x)).$$

These results will be useful in order to find statistical guarantees on the dimension reduction procedure. The best achievable rate of convergence, obtained for $k = n^{4/(p+4)}$, is $n^{-1/(p+4)}$ for the estimation of the gradient vector. This rate is optimal, see Stone (1982), and is better than the rate $n^{-1/(6(p+4))}$ obtained in Kang and Shin (2022). A similar (non-asymptotic) bound is obtained in Ausset et al. (2021) in the simpler case of a local least squares estimator of the gradient which is not appropriate in the classification problem.
The above results only require regularity conditions at the point x. As a consequence, the best that can be hoped for is a pointwise asymptotic convergence result in the spirit of Theorem 2. Stronger results, such as uniform convergence results for $\widehat{b}_n(x)$ over compact subsets of the support of X, could be obtained under much more restrictive conditions.
Let us point out that from the estimator $(\widehat{a}_n(x), \widehat{b}_n(x))$, one can easily construct, using the standard plug-in rule, an estimator of $(\pi(x), \nabla\pi(x))$. This estimator will inherit the good statistical properties of $(\widehat{a}_n(x), \widehat{b}_n(x))$ established before.

Aggregating the directions
The optimization procedure (33) allows to estimate $\nabla\ell(x)$, which belongs to the central subspace for each $x \in S_X$. This subspace can be recovered by estimating such directions at different points x. To this aim, given a probability measure μ supported on S_X, let

$$M = \int_{\mathbb{R}^p} \nabla\ell(x)\nabla\ell(x)^T \mu(\mathrm{d}x) = \mathbb{E}_{X^*\sim\mu}[\nabla\ell(X^*)\nabla\ell(X^*)^T].$$

To estimate the matrix M, generate $X_i^* \sim \mu$, $i = 1, \ldots, m$ independently and compute

$$\widehat{M} = \frac{1}{m}\sum_{i=1}^{m} \widehat{b}_n(X_i^*)\widehat{b}_n(X_i^*)^T. \tag{35}$$

One can then define $(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ as the set of orthogonal eigenvectors of $\widehat{M}$, ordered according to their eigenvalues (in decreasing order). Finally, given $d \in \{1, \ldots, p\}$ (which is chosen in practice by cross-validation, as we shall explain next), the projection matrix

$$\widehat{P}_{\widehat{\beta}} = \widehat{P}_{\widehat{\beta},d} = \sum_{k=1}^{d} \widehat{\beta}_k \widehat{\beta}_k^T$$

defines an estimator of the projection on the central subspace of interest. A corollary from our main results can now be stated on the estimation of M.

Corollary 4. *Suppose that μ is a finitely supported measure on S_X and that for each $x \in \mathrm{supp}(\mu)$, assumptions (A1) and (A2) are satisfied. If $k := k_n \to \infty$ and $\lambda := \lambda_n$ are such that $k^{1+p/2}/n \to \infty$, $k^{1+p/4}/n$ is bounded and $n\lambda^p/k^{1+p/2}$ converges to a finite constant, then*

$$\widehat{M} - M = O_{\mathbb{P}}\left(\frac{1}{\tau_{n,k}(x)\sqrt{k}}\right) + O_{\mathbb{P}}(\tau_{n,k}(x)) + O_{\mathbb{P}}\left(\frac{1}{\sqrt{m}}\right).$$

The eigenprojector $\widehat{P}_{\widehat{\beta}}$ has the same rate of convergence. Also note that a natural choice for μ, which is also the one made in our numerical experiments, is the empirical measure of $X_1, \ldots, X_n$. This is, of course, a random measure whose number of atoms is not bounded with respect to n. From a theoretical perspective, we conjecture, following results given by Hristache et al. (2001) and Dalalyan et al. (2008), that such a choice would lead to a different rate of convergence compared to the one given in the above corollary under substantially stronger regularity assumptions than ours.

Numerical experiments

We now explain how the algorithms for the proposed methods work to estimate the dimension reduction matrix M. We then describe the different competitors and evaluation metrics that we shall use. We next analyze synthetic data examples and finally consider real data examples in order to showcase the benefits of our methodology in classification tasks.

The algorithm: Estimation of the dimension reduction matrix

For a given choice of m (the number of b's used to estimate M), a given value of λ, the penalization parameter in the optimization problem, and k the number of neighbors, the computation of the dimension reduction matrix is straightforward, see Algorithm 1 below (the optimization uses the `R` function `glmnet` from the package of the same name).

Algorithm 1 Estimation of M

Input: $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^p \times \{0, 1\}$, $\lambda > 0$, $k \in \{1, \dots, n\}$ and $m \in \{1, \dots, n\}$
Output: Dimension reduction matrix $\widehat{M}$
Draw uniformly a list $\mathcal{X}_m$ of m observations among $\mathcal{X} = (X_1, \dots, X_n)$ without replacement
for each $x \in \mathcal{X}_m$ **do**
    Compute $N_k(x)$, the index of the k nearest neighbors to x among $\mathcal{X}$
    Compute $\widehat{a}_n(x)$ and $\widehat{b}_n(x)$ according to (33) using gradient descent
end for
Return $\widehat{M} = \frac{1}{m} \sum_{x \in \mathcal{X}_m} \widehat{b}_n(x) \widehat{b}_n(x)^T$

We now discuss the choice of the hyperparameters. Ideally, one should set $m = n$ so that the gradient is estimated at each data point; we set here $m = n/4$ to save computing time, as in our experience this choice does not adversely affect finite-sample results substantially. We also set $k = \lfloor \sqrt{n} \rfloor$. Furthermore, to mitigate biases that may arise from imbalanced class distributions, we exclude samples where after finding the closest neighbors, either class 0 or class 1 is rare, defined here as having fewer than 5 points within one of the two classes.

To estimate the directions featured in (35), we use either the pure nearest-neighbor logistic log-likelihood without penalization or its penalized version. For the latter, we require an optimal choice of λ which balances the regularization and goodness-of-fit. We propose to select the same parameter λ for all $x \in \mathcal{X}_m$ in order to alleviate the computational burden. We therefore select λ by 10-fold cross-validation at the average point $x = \overline{x} = \frac{1}{n} \sum_{i=1}^n X_i$. In other words, we divide the data into 10 randomly selected subsets of equal size, each subset serving as a validation set while the remaining subsets are used for model fitting. The fitted model is assessed on the validation set using the misclassification error, that is, the proportion of observations whose label is not correctly predicted, which is the empirical counterpart of the misclassification risk $\mathcal{R}(g) = \mathbb{P}(g(X) \neq Y)$ for a given classifier g; in this cross-validation procedure, an observation is labeled as 1 if and only if its predicted probability of success by the nearest-neighbor logistic log-likelihood estimator $\widehat{\pi}_n(\overline{x}) = \mathrm{expit}(\widehat{a}_n(\overline{x}))$ exceeds $1/2$. This evaluation of the quality of the fitted model is then done across a sequence of λ values: more precisely, we use the `R` function `cv.glmnet` from the `glmnet` package with `type.measure=class`, which computes the average misclassification error across all validation sets, and the regularization parameter λ selected is the one minimizing this error.

Selecting the dimension of the reduction subspace

Unlike some dimension reduction approaches, we estimate the dimension of the reduction subspace using cross-validation on the underlying prediction problem, and not from a test on the eigenvalues of the outer product of gradients $\widehat{M}$. Our procedure first divides the data into a training set $(X_{\mathrm{train}}, Y_{\mathrm{train}})$ and a testing set $(X_{\mathrm{test}}, Y_{\mathrm{test}})$. The matrix $\widehat{M}$ and its p orthogonal eigenvectors $\widehat{\beta}_1, \dots, \widehat{\beta}_p$, in decreasing order according to their eigenvalues, are estimated from the training set following the procedure described in Algorithm 1. For every $d \in \{1, \dots, p\}$, the first d eigenvectors are then gathered in a matrix $\widehat{\beta}_{(1:d)} \in \mathbb{R}^{p \times d}$, the sets of

covariates X_{train} and X_{test} are projected onto the sets of lower-dimensional covariates $X_{\text{train}}\widehat{\beta}_{(1:d)}$ and $X_{\text{test}}\widehat{\beta}_{(1:d)}$, and a classifier is learned based on the training set $(X_{\text{train}}\widehat{\beta}_{(1:d)}, Y_{\text{train}})$. We use here mainly the `knn` nearest-neighbor classifier, that is, at a given point x, the result of the majority vote among the nearest neighbors of x within the space of projections of the covariates in the training set (with ties broken at random). For this classifier, the training step merely consists of storing the covariates in the training set along with their labels, which will form the basis for the vote at each point. In the real data analysis, we shall also compare our results with the Random Forest classifier, whose training step is nontrivial. The performance of the chosen classifier is then evaluated on the test set $(X_{\text{test}}\widehat{\beta}_{(1:d)}, Y_{\text{test}})$, and the dimension retained is the one for which the classifier has the lowest misclassification risk.
This forward iterative procedure, summarized in Algorithm 2, strike a balance between dimension reduction and the preservation of relevant information for classification. We shall compare the results obtained with the situation where the correct dimension of the reduction subspace is known in order to assess the influence of the dimension selection step.

Algorithm 2 Estimation of the dimension d

Input: Dataset (X, Y) with $X \in \mathbb{R}^{n\times p}$ and $Y \in \{0,1\}^n$, classification algorithm g (kNN, random forest...), and parameters $\lambda > 0$, $k \in \{1,\dots,n\}$, $m \in \{1,\dots,n\}$ and $K \geq 2$
Output: Dimension of reduction subspace
Estimate $\widehat{M}$ using Algorithm 1 and compute the eigenvectors $\widehat{\beta}_1,\dots,\widehat{\beta}_p$ of $\widehat{M}$
Split (X, Y) into K folds $(X_{(j)}, Y_{(j)})_{j=1,\dots,K}$
for each $d \in \{1,\dots,p\}$ **do**
    Define $\widehat{\beta}_{(1:d)} = [\widehat{\beta}_1 \cdots \widehat{\beta}_d]$
    for each $j \in \{1,\dots,K\}$ **do**
        Define $(X_{\text{train}}, Y_{\text{train}}) = (X, Y)\backslash(X_{(j)}, Y_{(j)})$
        Train the classification rule g on data $(X_{\text{train}}\widehat{\beta}_{(1:d)}, Y_{\text{train}})$
        Evaluate its misclassification risk $R_{j,d}$ using $(X_{(j)}\widehat{\beta}_{(1:d)}, Y_{(j)})$
    end for
    Compute $R_d = \frac{1}{K}\sum_{j=1}^{K} R_{j,d}$
end for
Return: d minimizing R_d

Real data analyses

We apply the proposed methodology to three real data sets, all freely available from the UCI repository and on file with the authors:

- The Hill-Valley (HV) dataset[1]. Each data point is made of 100 real numbers $\boldsymbol{x}_i = (x_{i,j})_{1\leq j\leq 100}$ which create a curve in the two-dimensional plane that features a hill (a "bump" in the curve) or a valley (a "dip" in the curve). The data consists of the $n = 1212$ pairs $(Y_i, \boldsymbol{x}_i) \in \{0,1\} \times \mathbb{R}^{100}$, where $Y_i = 1$ if and only if the curve features a hill.

- The Mice Protein Expression (MPE) dataset[2]. After data cleaning, the dataset contains $n = 1047$ observations with $p = 71$ attributes, consisting of healthy mice and mice diagnosed with Down's syndrome.

- The Wisconsin Diagnostic Breast Cancer (WDBC) dataset[3]. A total of $n = 569$ subjects are diagnosed with breast tumors, either benign or malignant. Ten features of breast cell nuclei are measured for each subject, with the mean, standard error, and largest values recorded for each feature, leading to $p = 30$ predictors in total.

https://archive.ics.uci.edu/dataset/166/hill+valley
https://archive.ics.uci.edu/dataset/342/mice+protein+expression
https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Each dataset is divided at random into a training set and a testing set, approximately made of 70% and 30% of the original data, respectively. Table compares the performance of the complete workflow (dimension reduction and selection and then classification) using the full space of covariates, the covariates projected on the dimension reduction subspace provided by the proposed LLO$(\lambda > 0)$ method, and its version obtained using the non-penalized version LLO$(\lambda = 0)$, paired with either the `knn` or the `RandomForest` classifier, when applied to the testing set. The classification procedure using LLO$(\lambda > 0)$ generally has a comparable or lower misclassification risk, a comparable or higher AUC with comparable or lower computing time with respect to the non-penalized version, and always improves substantially upon the classifier not featuring dimension reduction.

Classifier	Random Forest			knn		
	Miscl. risk	AUC	Est. time	Miscl. risk	AUC	Est. time
Hill-Valley (HV)						
No dimension reduction	0.437	0.563	3.95	0.481	0.516	1.48
LLO$(\lambda = 0)$	0.212	0.855	3.16	0.429	0.597	**0.73**
LLO$(\lambda > 0)$	**0.115**	**0.952**	**1.44**	**0.126**	**0.953**	**0.73**
Mice Protein Expression (MPE)						
No dimension reduction	0.019	0.998	1.78	0.105	0.962	1.04
LLO$(\lambda = 0)$	**0.013**	**0.999**	1.35	**0.07**	0.98	0.94
LLO$(\lambda > 0)$	**0.013**	**0.999**	**0.62**	0.08	**0.985**	**0.41**
Wisconsin Diagnostic Breast Cancer (WDBC)						
No dimension reduction	0.064	0.979	0.88	0.058	0.972	0.41
LLO$(\lambda = 0)$	0.053	0.987	0.8	0.058	**0.985**	**0.28**
LLO$(\lambda > 0)$	**0.035**	**0.99**	**0.69**	**0.029**	0.982	0.32

Future Research Plans

Building on my research interest in machine learning and enhanced time series models, I aim to integrate flexible methodologies that further improve inference and prediction in complex stochastic systems. In addition to exploring neural network-based distributional regression models with deep Bayesian neural networks for one-step-ahead extreme event forecasting. I am also interested in hybrid Neural-GAM models that enhance interpretability in time series analysis, spatio-temporal Bayesian deep learning for environmental predictions, and functional data analysis using deep learning for continuous time series extremes. My research will also investigate nonparametric machine learning approaches, causal inference in time series and dimension reduction scenarios, and deep generative models to address challenges such as missing data and irregular structures in environmental datasets. Moreover, I plan to extend my work on uncertainty quantification in time series by incorporating bootstrap saive confidence intervals and Bayesian methods to provide robust prediction intervals for high-dimensional environmental applications. These directions will enable the development of more reliable, interpretable, and flexible models for forecasting in complex data structures.

References

- Buddana, A., and Kozubowski, T. J. (2014). Discrete Pareto distributions. *Economic Quality Control*, 29(2), 143–156.
- Hitz, A. S., Davis, R. A., & Samorodnitsky, G. (2024). Discrete Extremes. *Journal of Data Science*, 22(4).
- Gaver, D. P., and Lewis, P. (1980). First-order autoregressive gamma sequences and point processes. *Advances in Applied Probability*, 12(3), 727–745.

- Warren, D. (1992). "A multivariate gamma distribution arising from a Markov model. *Stochastic Hydrology and Hydraulics*, 6(3), 183–190.
- Wolpert, R. L. (2021). Lecture Notes on Stationary Gamma Processes. arXiv preprint arXiv:2106.00087.
- Chavez-Demoulin, V., and Davison, A. (2012). Modelling time series extremes. *REVSTATStatistical Journal*, 10(1), 109–133.
- Davis, R. A., and Mikosch, T. (2009). The extremogram: A correlogram for extreme events. *Bernoulli*, 15(4), 977–1009.
- McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools-revised edition*, New Jersey: Princeton University Press.
- Pedeli, X., and Varin, C. (2020). Pairwise likelihood estimation of latent autoregressive count models. *Statistical Methods in Medical Research*, 29(11), 3278–3293.
- Lindsay, B. G. (1988).Composite likelihood methods. *Contemporary Mathematics*, 80(1), 221– 239.
- Bortot, P., and Gaetan, C. (2014). A latent process model for temporal extremes. *Scandinavian Journal of Statistics*, 41(3), 606–621.
- Naveau, P., Huser, R., Ribereau, P., and Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52, 2753–2769.
- Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1–14.
- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 3–36.
- Fan, J., N. E. Heckman, and M. P. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* 90(429), 141–150.
- Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* 10(4), 1040–1053.
- Kang, J. and S. J. Shin (2022). A forward approach for sufficient dimension reduction in binary classification. *Journal of Machine Learning Research* 23(199), 1–31.
- Ausset, G., S. Clémençon, and F. Portier (2021). Nearest neighbour based estimates of gradients: Sharp nonasymptotic bounds and applications. *In International Conference on Artificial Intelligence and Statistics*, pp. 532–540. PMLR
- Hristache, M., A. Juditsky, and V. Spokoiny (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics* 29(3), 595–623.
- Dalalyan, A. S., A. Juditsky, and V. Spokoiny (2008). A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research* 9(53), 1648–1678.

Referees

1. **Prof. Carlo Gaetan**
   Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Italy.

   E-mail: gaetan@unive.it **Phone:**(+39)-412-348-404

2. **Prof. François Portier**
   Department of Statistics, CREST, ENSAI, France.

   E-mail: francois.portier@gmail.com **Phone:**(+33)-766-138-179

3. **Prof. Irshad Ahmad Arshad**
   Department of Statistics, Allama Iqbal Open University, Islamabad, Pakistan.

   E-mail: rshad.ahmad@aiou.edu.pk **Phone:** (+92)-333-518-180-7