IAA AAI
INSTITUT DES ACTUAIRES
SECTIONS VIRTUAL COLLOQUIUM | 2020

Generalized Pareto Regression Trees for extreme claim prediction

joint work with O. Lopez

Maud Thomas,
ISUP - LPSM/ Sorbonne Université

May 11th – May 15th 2020

About the speaker

SECTIONS VIRTUAL COLLOQUIUM 2020

Maud Thomas

- Assistant professor at Sorbonne Université
- Co-chair of the Actuarial Master Degree of ISUP
- Associate member of the French Institute of Actuaries

Sorbonne Université

- Institut statistique de l’Université de Paris (ISUP)
- Laboratoire de Probabilités, Statistique et Modélisation

Actuarial modelling

- X characteristics of a policyholder
- N number of claims ($\mathbb{E}[N \mid X]$ =frequency)
- Y cost of a claim ($\mathbb{E}[Y \mid X]$ =severity)

Pricing principle = balance (in average) the cost of a policyholder and the commitments of the insurer

$$\pi(X) = E[N \mid X]E[Y \mid X]$$

- $\pi(X)$ = premium of the insurance contract of a policyholder with characteristics X
- Common assumption: Y and N are independent given X

Reserving = Need to estimate the whole conditional distribution of N and Y given X

Extreme claims

- Risk management
- Extreme event: some value exceeds a (high) threshold
- Lack of data and/or historical information
- Present some heterogeneity

⇒ Evaluating the potential cost of extreme risks is a challenging task

Objectives of the presentation

Main goals

1. Study extreme claims
2. Gain further insight on their heterogeneity
3. Analyse the impact of characteristics on extreme claims

Focus on

- Tail of the distribution
- Severity of extreme claims

⇒ Two statistical tools :

1. Extreme value theory
2. Regression and classification trees

Statistical tools

Extreme Value theory

Goals of Extreme Value Theory

1. Estimate extreme quantiles
2. Estimate the occurrence probability of an event more extreme than previously observed

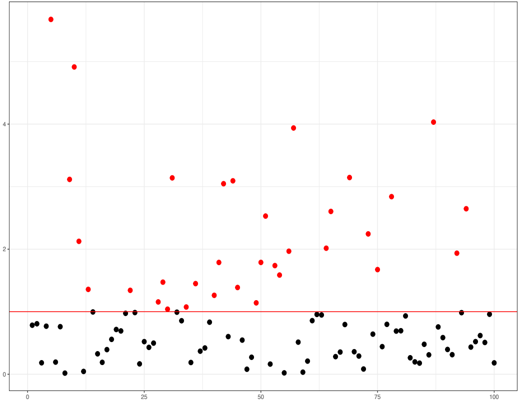
$\Rightarrow$ Inference outside of the range of the data

Extreme value theory

Peaks over threshold method

- $Y_1, Y_2, \ldots$ series of i.i.d. random variables
- Fix a (high) threshold u
- Extreme event = Y_i exceeds u

  → Given that $Y_i > u$, define the excess $X_i = Y_i - u$



Extreme value theory

Peaks over threshold method

- $Y_1, Y_2, \ldots$ series of i.i.d. random variables
- Fix a (high) threshold u
- Extreme event = Y_i exceeds u

  → Given that $Y_i > u$, define the excess $X_i = Y_i - u$

Balkema and de Haan (1974)

If there exist $(a_u) > 0$, (b_u) and a non-degenerated distribution function H such that,

$$\mathbb{P}[Y_i - u \geqslant a_u x + b_u \mid Y_i > u] \xrightarrow[u\to\infty]{d} 1 - H(x)\,,$$

then H is necessarily of the form

$$H_{\sigma,\gamma}(x) = \begin{cases} 1 - \left(1 + \frac{\gamma}{\sigma}x\right)^{-1/\gamma} & \text{if } \gamma \neq 0 \\ 1 - \exp\left(-\frac{x}{\sigma}\right) & \text{if } \gamma = 0 \end{cases}$$

- Possible limits of excesses = Parametric family of distributions

  ↪ Generalized Pareto Distributions

Extreme value theory and regression models

- Semi-parametric approaches
  - Exponenial regression model (Beirlant et al., 2003)
  - Smoothing splines (Chavez-Demoulin et al., 2015)
- Non parametric approach (Beirlant and Goegebeur, 2004)
  - Local polynomial maximum likelihood
  - Only for continuous covariates

Statistical tools

CART algorithm

Classification And Regression Trees (CART)

SECTIONS VIRTUAL COLLOQUIUM 2020

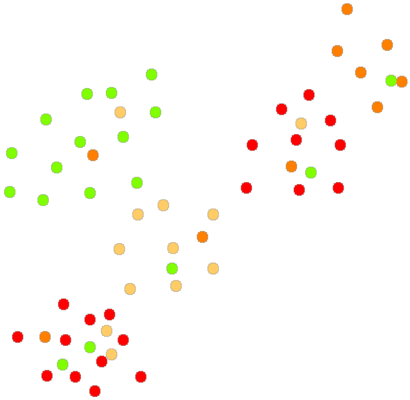
Regression tree (Breiman et al., 1984)

$$m^* = \arg\min_{m \in \mathscr{M}} \mathbb{E}[\phi(Y, m(\mathbf{X}))],$$

- Y is a response variable (the cost of a cyber claim in our case)
- $\mathbf{X} \in \mathscr{X} \subset \mathbb{R}^d$ is a set of covariates
- $\mathscr{M}$ is a class of target functions on $\mathbb{R}^d$
- ϕ is a loss function that depends on the quantity we wish to estimate

Growing phase

SECTIONS VIRTUAL COLLOQUIUM 2020



CART : Step 0

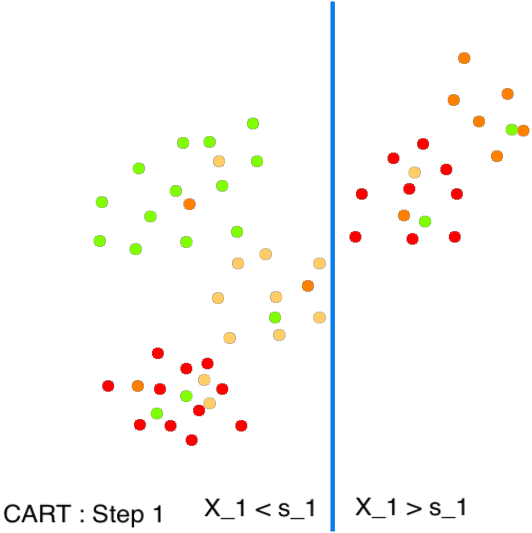
SECTIONS VIRTUAL COLLOQUIUM 2020

Splitting rules

$$\mathbf{x} = (x^{(1)}, \dots, x^{(d)}) \longrightarrow R_j(\mathbf{x})$$

with

$$\begin{cases} R_j(\mathbf{x}) & = 0 \text{ ou } 1 \\ R_j(\mathbf{x}) R_{j'}(\mathbf{x}) & = 0 \text{ for } j \neq j' \\ \sum_j R_j(\mathbf{x}) & = 1 \end{cases}$$



Growing phase

1. **Step 0 :** $R_0(\mathbf{x}) = 1$ and $n_1 = 1$ (root)
2. **Step k + 1**
    - $(R_0, \ldots, R_{n_k})$ rules obtained at step **k**. For $j = 1, \ldots, n_k$
    - If all observations s.t. $R_j(\mathbf{X}_i) = 1$ have the same characteristics. Keep R_j
    - else, R_j is replaced by two new rules R_{j_1} and R_{j_2}

      → For each component $X^{(l)}$ of $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$, define $x_\star^{(l)}$

$$x_\star^{(l)} = \arg\min_{x^{(l)}} \Phi(R_j, x^{(l)})$$

$\Phi(R_j, x^{(l)})$ = an empirical version of $\mathbb{E}[\phi(Y_i, \mathbf{X}_i)]$ computed on each sub-group

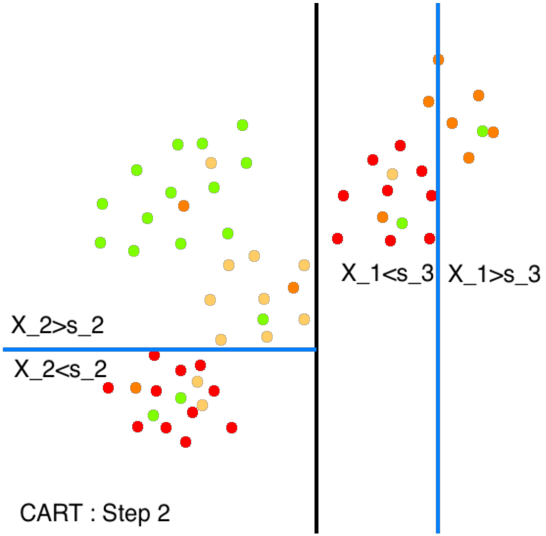
→ Select the best component index

$$\hat{l} = \arg\min_{l} \Phi(R_j, x_\star^{(l)})$$

→ Define

$$R_{j_1}(\mathbf{x}) = R_j(\mathbf{x}) \mathbb{1}_{x^{(\hat{l})} \leq x_\star^{(\hat{l})}} \quad \text{and} \quad R_{j_2}(\mathbf{x}) = R_j(\mathbf{x}) \mathbb{1}_{x^{(\hat{l})} > x_\star^{(\hat{l})}}$$

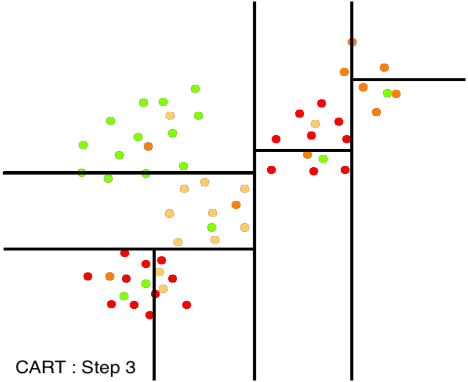
Growing phase



Growing phase

Regression estimator $\hat{m}^{\mathscr{R}}(\mathbf{x})$ of m^* given by

$$\hat{m}^{\mathscr{R}}(\mathbf{x}) = \sum_{j=1}^{s} \hat{m}(R_j) R_j(\mathbf{x}) \;\text{ where }\; \hat{m}(R_j) = \arg\min_{m \in \mathscr{M}} \sum_{i=1}^{n} \phi(Y_i, \mathbf{X}_i) R_j(\mathbf{X}_i)$$



The splitting rule and loss functions

- Quadratic loss → Mean regression

$$\phi(y, m(\mathbf{x})) = (y - m(\mathbf{x}))^2$$

$\hookrightarrow m^*(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$

- Absolute loss → Median regression

$$\phi(y, m(\mathbf{x})) = |y - m(\mathbf{x})|$$

$\hookrightarrow m^*(\mathbf{x}) =$ conditional median

- Log-likelihood loss, here GPD

$$\phi(y, m(\mathbf{x})) = -\log(\sigma(\mathbf{x})) - \left(\frac{1}{\gamma(\mathbf{x})} + 1\right) \log\left(1 + \frac{y\gamma(\mathbf{x})}{\sigma(\mathbf{x})}\right),$$

$\hookrightarrow m^*(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$

The splitting rule and loss functions

- Quadratic loss → Mean regression

$$\phi(y, m(\mathbf{x})) = (y - m(\mathbf{x}))^2$$

$\hookrightarrow m^*(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$

- Absolute loss → Median regression

$$\phi(y, m(\mathbf{x})) = |y - m(\mathbf{x})|$$

$\hookrightarrow m^*(\mathbf{x}) =$ conditional median

- Log-likelihood loss, here GPD

$$\phi(y, m(\mathbf{x})) = -\log(\sigma(\mathbf{x})) - \left(\frac{1}{\gamma(\mathbf{x})} + 1\right) \log\left(1 + \frac{y\gamma(\mathbf{x})}{\sigma(\mathbf{x})}\right),$$

$\hookrightarrow m^*(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$

Pruning step: model selection

- Let $T_{\max}$ be the maximal tree obtained in the first phase and $K_{\max}$ the number of its leaves
- Consists in the extraction of a subtree from $T_{\max}$
- Standard way to proceed = use a penalized approach
  $\rightarrow$ Disadvantage the trees with large numbers of leaves

- Subtree S associated with a set of rules $\mathscr{R}^S = \left(R_1^S, \dots, R_{n_S}^S\right)$
- Select the subtree $\widehat{S}(\alpha)$ that minimizes, among all subtrees of $T_{\max}$ the criterion

$$C_\alpha(S) = \sum_{i=1}^{n} \phi(Y_i, m^{\mathscr{R}^S}(\mathbf{X}_i)) + \alpha n_S$$

- $\alpha > 0$ is chosen by cross-validation
- Denote $\widehat{T}_{\widehat{K}}$ the selected tree and $\widehat{K}$ the number of its leaves

Consistency of the algorithm

- Let $\widehat{T}_K$ any subtree of $T_{\max}$ with K leaves
- Let T_K^* be the optimal tree among all trees with K leaves

Consistency of the tree

Under certain conditions, for all $K = 0, \ldots, K_{\max}$

$$\mathbb{E}\left[\|\widehat{T}_K - T_K^*\|_2^2\right] \leq C\frac{(\log n)^2 \log(n/k_n)}{k_n}$$

- Let T^* be the optimal tree and K_0 the number of its leaves

Consistency of the pruning step

Under certain conditions

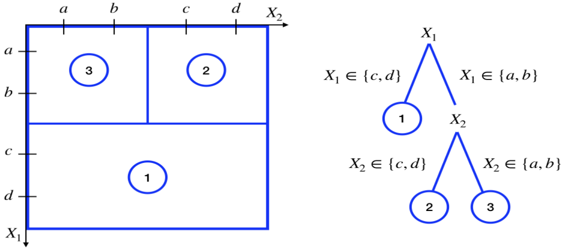
$$\mathbb{E}\left[\|\widehat{T}_{\widehat{K}} - T^*\|_2^2\right] \leq C' K_0 \frac{(\log n)^2 \log(n/k_n)}{k_n}$$

Numerical expirements

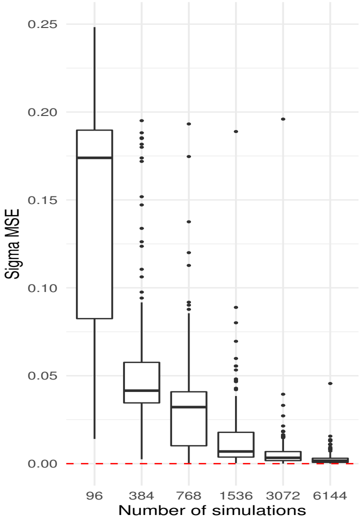
Simulated data

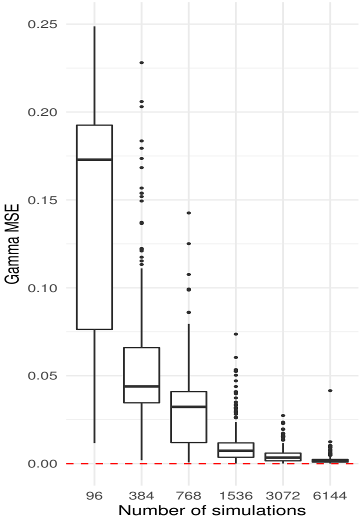
- $\mathbf{X} = (X_1, X_2, X_3)$ 3 discrete covariates taking values in $\{a, b, c, d\}$
- $Y \sim \text{GPD}(\sigma(\mathbf{X}), \gamma(\mathbf{X}))$ distributed according to a toy model
- 2 splits on X_1 and X_2
- 3 terminal leaves
- $(\sigma_1, \sigma_2, \sigma_3) = (\gamma_1, \gamma_2, \gamma_3) = (0.5, 1, 1.5)$

- Simulate $Y_1, \ldots, Y_N$
- N =96, 384, 768, 1536, 3072, 6144

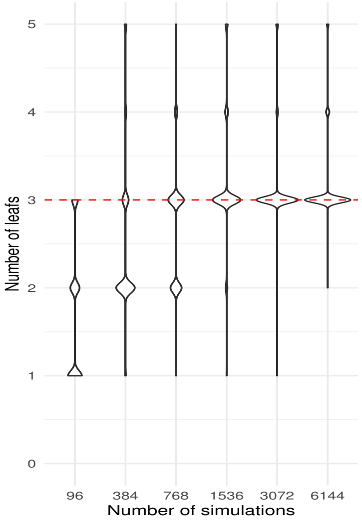


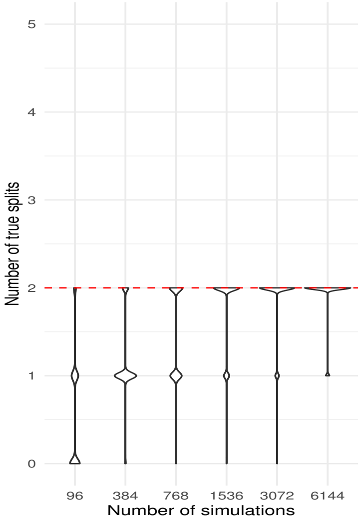
Simulated data





Simulated data





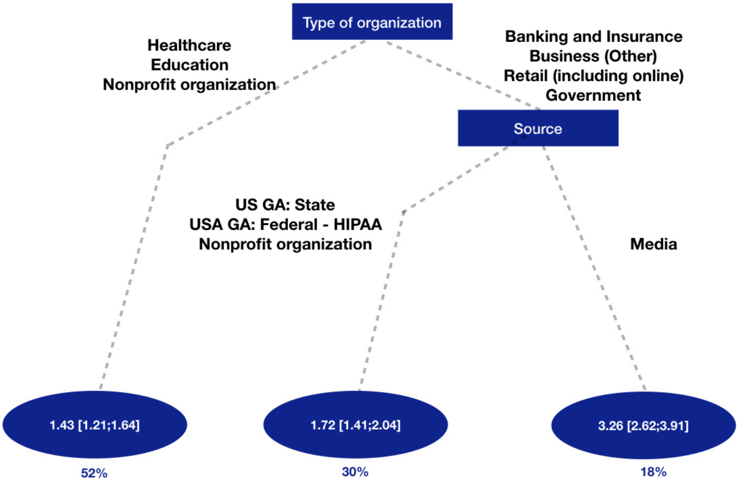
Application to real data: cyber-claims

(Farkas et al, 2020)

- Privacy Rights Clearinghouse (nonprofit association)
- Founded in 1992
- Publicly available
- Benchmark for Cyber event analysis
- Aim at raising awareness about privacy issues.
- Chronology of data breaches maintained from 2005.
- Gathering events information from multiple sources:
  - US Government Agencies (Federal level–HIPAA): Health domain, obligation to declare any breach that affects more than 500 individuals
  - US Government Agencies (State level): since 2018, each state has a specific legislation related to data breaches
  - Media
  - Non profit organizations.
- Focus on the Tail of the distribution
  - Consider only the number of affected records above 27 000
  - Fit a GPD CART

Application to real data: cyber-claims

Farkas et al, 2020



Conclusion

- Propose a methodology to study extreme claims by taking into account
  - heterogeneity,
  - impact of the covariates
  - evolution through time
- Give theoritical guarantees

- Advantage: interpretation.
- Drawbacks: the robustness of the tree structure and the estimator.

- Future works: consider random forest

Thank you for your attention

SECTIONS VIRTUAL COLLOQUIUM 2020

Contact details :

Maud Thomas

ISUP - LPSM, Sorbonne Université
4 place Jussieu 75005 Paris

maud.thomas@sorbonne.universite.fr

https://www.actuarialcolloquium2020.com/

SECTIONS VIRTUAL COLLOQUIUM 2020

Disclaimer: