Supervisory Committee Meeting

Xiaomeng Ju

UBC

18/08/2021

Overview

- Thesis scope
- Contents of the thesis
- Projects and progress
- Challenges
- Timeline

Thesis Scope

- Title: Robust boosting for complex data
- Goal: estimate F in a regression model, $Y \in \mathbb{R}$

$$Y = F(X) + \epsilon \tag{1}$$

- Complex data:
    - (a) $X \in \mathbb{R}^p$, Y contains outliers → Chapter 2
    - (b) $X \in$ Hilbert space, Y follows (1) → Chapter 3
    - (c) $X \in$ Hilbert space, Y contains outliers → Chapter 4
    - (d) $X \in L^2(\mathcal{I})$ and evaluated on a sparse grid, possibly contaminated, Y follows (1) → Chapter 5

  For (b), (c), and (d), X may also contain real-valued predictors $\in \mathbb{R}^d$

Robust gradient boosting: MM-estimators

- Motivated by MM-estimator for regression
  - S-estimator: highly robust

    $$\hat{F}_S = \operatorname*{argmin}_F \hat{\sigma}_S(F)\,,$$

    where $\hat{\sigma}_S(F)$ is a robust scale of residuals
  - M-estimator: highly efficient, initialized at $\hat{F}_S$

    $$\hat{F}_M = \operatorname*{argmin}_F \sum_{i \in \mathcal{I}_{\text{train}}} \rho\left(\frac{y_i - F(\mathbf{x}_i)}{\hat{\sigma}_S(\hat{F}_S)}\right)$$

  - MM-estimator: highly robust and highly efficient

Robust gradient boosting: methodology

RRBoost

- **Stage 1** : compute an S-type boosting estimator $\hat{F}_S$ with high robustness but possibly low efficiency (S-scale as L)
- **Stage 2**: compute an M-type boosting estimator initialized at the function estimator ($\hat{F}_S$) and scale estimator $\hat{\sigma}_S$ obtained in Stage 1. (bonded loss as L)

Boosting for functional regression: progress

- Ju, Xiaomeng, and Matías Salibián-Barrera. "Robust boosting for regression problems." Computational Statistics & Data Analysis 153 (2021): 107065.
- RRBoost package on CRAN
- Repo: `https://github.com/xmengju/RRBoost`

Boosting for functional regression: model

- Goal: estimate $F : X \to Y$ in order to make predictions for future observations
- Multi-index model:

$$F(X) = r(\langle X, \alpha_1 \rangle, ..., \langle X, \alpha_p \rangle)$$

  Fit complex functions; capture iterations between indices

- Approximation:

$$F(X) \approx r_1(\langle X, \beta_{1,1} \rangle, ..., \langle X, \beta_{1,k} \rangle) + ... + r_T(\langle X, \beta_{T,1} \rangle, ..., \langle X, \beta_{T,K} \rangle),$$

  where each $r_j(\langle X, \beta_{j,1} \rangle, ..., \langle X, \beta_{j,K} \rangle)$ is fitted by a functional multi-index tree.

Boosting for functional regression: methodology

- Propose a boosting algorithm: TFBoost
- Input data: $(\mathbf{x}_i, y_i)$, $i \in \{\mathcal{I}_{\text{train}} \cup \mathcal{I}_{\text{val}}\}$
- Loss function: $L(y_i, F(x_i))$
- Every boosting iteration:
  calculate negative gradient → fit base learner (functional multi-index tree) → find step size (α_t) → update function

$$\hat{F}(x_i) = \sum_{t=1}^{T} \alpha_t \hat{r}_t(\langle x_i, \hat{\beta}_{t,1} \rangle, ..., \langle x_i, \hat{\beta}_{t,k} \rangle)$$

- Multi-index tree
  - Type A tree: optimal indices for the whole tree
  - Type B tree: optimal index for each split (fast calculation)

Boosting for functional regression: progress

- Paper draft soon to be submitted (in August)
- TFBoost package completed
- Repo: `https://github.com/xmengju/TFBoost`

Robust TFBoost: problem description

- Extend TFBoost to data with outliers
- Most proposals are for linear models
- Types of outliers (include a figure):
  - Shape outliers
  - Magnitude outliers (curve, point, or interval)
  - Vertical outliers

Robust TFBoost: methodology

- TFBoost(LAD): TFBoost with L1 loss
- TFBoost(LAD-M): TFBoost(LAD) $\rightarrow$ residual scale $\rightarrow$ M-type TFBoost
- TFBoost(RR): S-type TFBoost $\rightarrow$ M-type TFBoost

Robust TFBoost: progress

- Simulation results comparing 3 proposals with competing robust functional regression methods in the literature
- Technical report that describes the methodology and the simulation study
- Deadline: mid September

Sparse TFBoost: methodology

- Problem: difficulty to calculate the inner product with sparsely observed functions
- Idea: borrow strength across functions

$$\tilde{X}(t) = \sum_{j=1}^{K} \xi_j \phi_j(t),$$

  where $\phi_j(t)$ are FPC and $\xi_j = \langle (X(t) - \mu(t)), \phi_j(t) \rangle$

- Methods:
  - Sparse X with possible measurement errors: PACE
  - Sparse X with functional outliers: ROB

Sparse TFBoost: progress

- Have reviewed literature and got familiar with the software of PACE and ROB
- Most of this project remains to be done.

Challenges

Past:

- Late completion of the comprehensive exam
- Limited computational resources (solved)
- Time management

Future:

- Writing speed
- Additional time to revise the TFBoost paper after submission

Timeline

- 2021/08:
  - TFBoost: submit paper and package
  - Thesis: draft the RRBoost and TFBoost chapters
  - Robust TFBoost: simulation and real example
- 2021/09:
  - Thesis: draft Robust TFBoost chapter
  - Sparse TFBoost: simulation
  - Revision: RRBoost and TFBoost chapters
- 2021/10:
  - Thesis: draft Sparse TFBoost chapter
  - Sparse TFBoost: simulation and real example
  - Revision: RRBoost and TFBoost chapters
- 2021/11:
  - Thesis: draft Sparse TFBoost chapter, conclusion and future work
  - Revision: Robust TFBoost chapter
- 2021/12:
  - Thesis: draft conclusion and future work
  - Revision: Sparse TFBoost chapter

- 2022/01:
  - Revision: Conclusion and future work
  - Complete the first revised draft
- 2022/02:
  - Complete the second revised draft
  - Send the draft to committee members
- 2022/06:
  - Thesis defense

Contents of the Thesis

- Chapter 1: Introduction
  - 1.1 The regression problem
  - 1.2 Gradient boosting
  - 1.3 Complex data
  - 1.4 Outline of the thesis
- Chapter 2: Robust gradient boosting
  - 2.1 Robust regression
    - 2.1.1 Robust linear regression
    - 2.1.2 Robust nonparametric regression
  - 2.2 Related work
  - 2.3 SBoost
  - 2.4 RRBoost
  - 2.5 Robust variable importance
  - 2.6 Simulation studies
  - 2.7 Empirical studies
  - 2.8 Software

Contents of the Thesis

- Chapter 3: Boosting for functional regression
  - 3.1 Functional regression
    - 3.1.1 Functional linear regression
    - 3.1.2 Functional non-parametric regression
    - 3.1.3 Functional single-index regression
  - 3.2 Related work
  - 3.3 Tree-based functional boosting (TFBoost)
  - 3.4 Functional multi-index tree
    - 3.4.1 Type A tree
    - 3.4.2 Type B tree
  - 3.5 Simulation studies
  - 3.6 German electricity data
  - 3.7 Software
- Chapter 4: Robust boosting for functional regression
  - 4.1 Related work
  - 4.2 Robust TFBoost
  - 4.3 Simulation studies
  - 4.4 Empirical studies
  - 4.5 Software

Contents of the Thesis

- Chapter 5: Boosting for functional regression with sparsely observed functional explanatory variables
  - 5.1 Functional principal component analysis
    - 5.1.1 Principal analysis by conditional estimation (PACE)
    - 5.1.2 Robust functional principal components analysis (ROB)
  - 5.2 Related work
  - 5.3 Sparse TFBoost
  - 5.4 Simulation studies
  - 5.5 Empirical studies
  - 5.6 Software
- Chapter 6: Conclusion remarks and future work
  - 6.1 Conclusion
  - 6.2 Future work