Meeting Items

July 12, 2021

- Paper revision: Chapters 1-3
  - Eigenfunctions: used a large sample to calculate
- Robust TFBoost:
  - Different shrinkage for L2, LAD, and RR TFBoost
  - Code (NAs initialize RFPLM)

1 Robust TFBoost: Simulations

1.1 Data generation

- We generated data sets $D = \{(x_i, y_i), i = 1, ..., N\}$, consisting of a predictor $x_i \in \mathcal{L}_2$ and a scalar response y_i that follow the model:

$$y_i = r(x_i) + \rho\epsilon_i, \tag{1}$$

where the errors ϵ_i are i.i.d, r is the regression function, and $\rho > 0$ is a constant that controls the signal-to-noise ratio (SNR):

$$\text{SNR} = \frac{\text{Var}(r(X))}{\text{Var}(\rho\epsilon)}.$$

- To sample the functional predictors x_i, we considered the model:

$$x_i(t) = \mu(t) + \sum_{p=1}^{4} \sqrt{\lambda_j}\xi_{ij}\phi_j(t), \tag{2}$$

where $\mu(t) = 2\text{sin}(t\pi)\text{exp}(1-t)$, $\lambda_1 = 0.8, \lambda_2 = 0.3, \lambda_3 = 0.2$, and $\lambda_4 = 0.1$, $\xi_{ij} \sim N(0, 1)$, and ϕ_j are the first four eigenfunctions of the "Mattern" covariance function $\gamma(s, t)$ with parameters $\rho = 3, \sigma = 1, \nu = 1/3$:

$$\gamma(s,t) = C\left(\frac{\sqrt{2\nu}|s-t|}{\rho}\right),\ C(u) = \frac{\sigma^2 2^{1-\nu}}{\Gamma(\nu)}u^{\nu}K_{\nu}(u),$$

where $\Gamma(.)$ is the Gamma function and K_ν is the modified Bessel function of the second kind. For each subject i, we evaluate x_i on a dense and regular grid $t_1, ..., t_{100}$ equally spaced in $\mathcal{I} = [0, 1]$.

- We considered five regression functions:
  - $r_1(X) = \int_{\mathcal{I}} \left(\sin\left(\frac{3}{2}\pi t\right) + \sin\left(\frac{1}{2}\pi t\right)\right) X(t) dt$,
  - $r_2(X) = 5\exp\left(-\frac{1}{2}\left|\int_{\mathcal{I}} x(t)\log(|x(t)|)dt\right|\right)$,
  - $r_3(X) = (\xi_1 + \xi_2)^{1/3}$, where $\xi_1 = \int_{\mathcal{I}}(X(t) - \mu(t))\psi_1(t)dt$ and $\xi_2 = \int_{\mathcal{I}}(X(t) - \mu(t))\psi_2(t)dt$ are projections onto the first two FPCs (ψ_1 and ψ_2) of X with mean $\mu(t) = E(X(t))$,
  - $r_4(X) = 5\text{sigmoid}\left(\int_{\mathcal{I}} X(t)^2 \sin(2\pi t)dt\right)$, where $\text{sigmoid}(u) = 1/(1 + \exp(-u))$, and
  - $r_5(X) = 5\left(\sqrt{\left|\int_{\mathcal{I}_1} \cos(2\pi t^2)X(t)dt\right|} + \sqrt{\left|\int_{\mathcal{I}_2} \sin(X(t))dt\right|}\right)$, where $\mathcal{I}_1 = [0, 0.5]$ and $\mathcal{I}_2 = (0.5, 1]$.

- For clean data (C_0), we generated ϵ_i in (1) from $N(0, 1)$ and selected ρ that corresponds to SNR $= 5$.

  For contaminated data, we sampled 10% training samples as outliers and let the set of their indices be I_{o}. The outliers belong to one of the five types introduced below. For $j \in I_{\text{o}}$,

  - C_1: *Shape outliers*

    In (1), $\epsilon_j \sim N(10, 0.25)$
    In (2), $\xi_{j,2} \sim N(10, 0.25)$ and the other parameters stay the same.

  - C_2: *Magnitude outliers*

    $x_j = 2\tilde{x}_j, y_j = 4\tilde{y}_j$, where $(\tilde{x}_j, \tilde{y}_j)$ were generated as clean data.

  - C_3: *Point-type measurement error outliers*

    Randomly sample 10 points form $t_1, ..., t_{100}$ and denote them as $t_{j,o_1}, ..., t_{j,o_{10}}$. For $k = 1, ..., 10$,

    $$x_j(t_{j,o_k}) = \tilde{x}_j(t_{j,o_k}) + \eta_{j,o_k},$$

    where $\eta_{j,o_k} \sim 0.5N(10, 0.25) + 0.5N(-10, 0.25)$, $y_j = \tilde{y}_j$, and $(\tilde{x}_j, \tilde{y}_j)$ were generated as clean data.

  - C_4: *Interval-type measurement error outliers*

    Randomly sample one interval from intervals $[t_1, ..., t_{10}]$, ...,$[t_{91}, ..., t_{100}]$, and denote the interval as $t_{j,o}, ..., t_{j,o+9}$
    For $k = 0, ..., 9$,

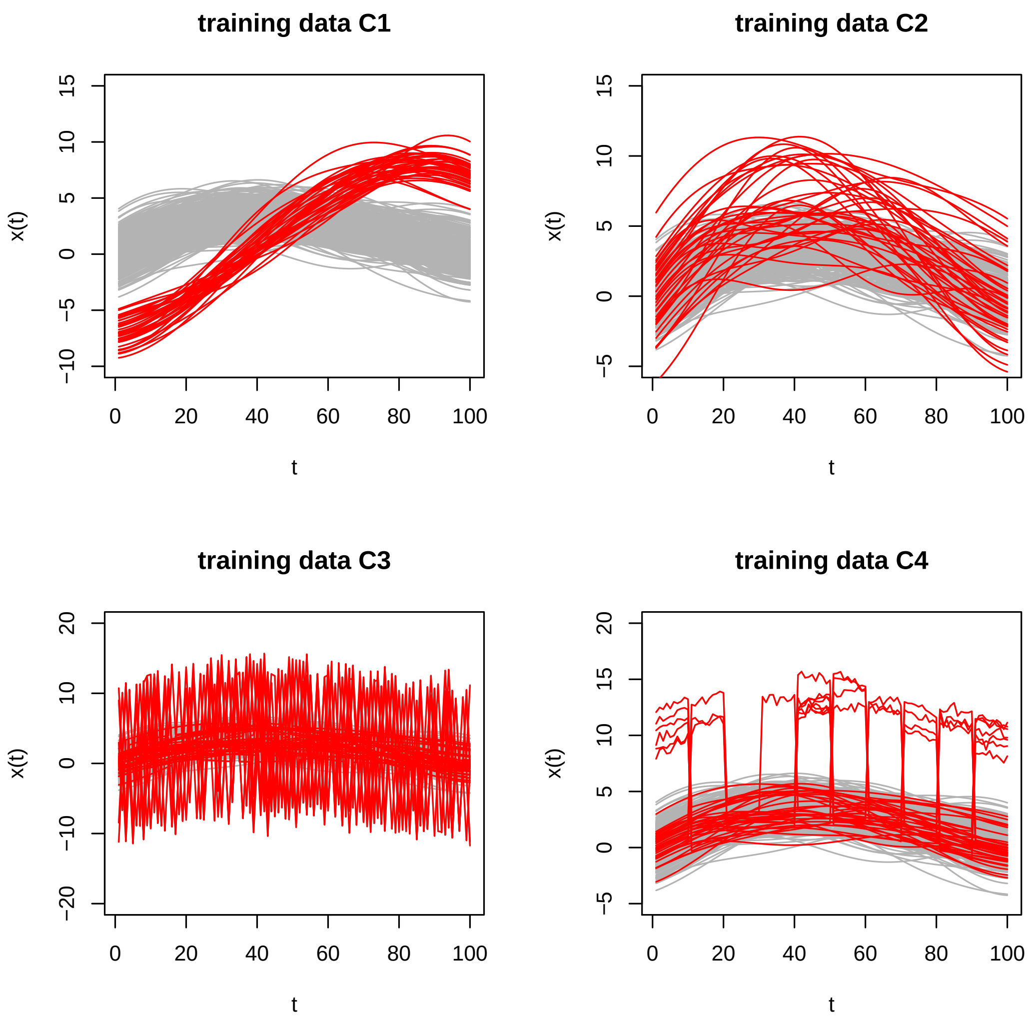
    $$x_j(t_{j,o+k}) = \tilde{x}_j(t_{j,o+k}) + \eta_{j,o+k},$$

where $\eta_{j,o+k} \sim N(10, 0.25)$, $y_j = \tilde{y}_j$, and $(\tilde{x}_j, \tilde{y}_j)$ were generated as clean data.

- C_5: *Pure vertical outliers*

$$\epsilon_j \sim N(10, 0.25)$$

1.2 Visualize the outliers



1.3 Model comparison

For each setting, we used 100 independently generated datasets and compared the performance of the following methods:

- `FPPR`: functional projection pursuit regression (Ferraty et al., 2013),
- `FGAM`: functional generalized additive models (McLean et al., 2014),
- `MFLM`: Sieve M-estimator for a semi-functional linear model Huang et al. (2015)
- `RFSIR`: robust functional sliced inverse regression (Wang et al., 2017)
- `RFPLM`: robust estimation for semi-functional linear regression models (Boente et al., 2020)
- `RTFBoost`: robust tree-based functional boosting

1.4 Timeline

- 2021/07:
  - TFBoost: revise paper (submit?)
  - Robust TFBoost: simulation
  - thesis: draft the background chapter
- 2021/08:
  - TFBoost: submit paper and package
  - thesis: draft the background, RRBoost and TFBoost chapters
  - Robust TFBoost: simulation and real example
  - record JSM presentation
- 2021/09:
  - thesis: draft Robust TFBoost chapter
  - Sparse TFBoost: simulation
- 2021/09:
  - thesis: draft robust TFBoost, Sparse TFBoost chapters
  - Sparse TFBoost: simulation and real example
- 2021/10:
  - thesis: draft Sparse TFBoost chapter, conclusion and future work
- 2021/11:
  - thesis: first draft complete, start revising

- 2021/12 (end of year):
  - thesis: second draft
- Before 2022/04:
  - thesis defence

References

Boente, G., Salibian-Barrera, M., and Vena, P. (2020). Robust estimation for semi-functional linear regression models. *Computational Statistics & Data Analysis*, 152:107041.

Ferraty, F., Goia, A., Salinelli, E., and Vieu, P. (2013). Functional projection pursuit regression. *Test*, 22(2):293–320.

Huang, L., Wang, H., Cui, H., and Wang, S. (2015). Sieve m-estimator for a semi-functional linear model. *Science China Mathematics*, 58(11):2421–2434.

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., and Ruppert, D. (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):249–269.

Wang, G., Zhou, J., Wu, W., and Chen, M. (2017). Robust functional sliced inverse regression. *Statistical papers*, 58(1):227–245.