# Simulation and Timeline

July 15, 2021

## 1 Robust TFBoost: Simulation

### 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 $\epsilon_i$ 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 $\phi_j$ 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 $K_\nu$ 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:
  - $r_1(X) = \int_{\mathcal{I}} \left(\sin\left(\frac{3}{2}\pi t\right) + \sin\left(\frac{1}{2}\pi t\right)\right) X(t)dt,$

- $r_2(X) = (\xi_1 + \xi_2)^{1/3}$, 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 ($\psi_1$ and $\psi_2$) of $X$ with mean $\mu(t) = E(X(t))$,
- $r_3(X) = 5\text{exp}\left(-\frac{1}{2}\left|\int_{\mathcal{I}} x(t)\log(|x(t)|)dt\right|\right)$,
- $r_4(X) = 5\text{sigmoid}\left(\int_{\mathcal{I}} X(t)^2 \sin(2\pi t)dt\right)$, 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 $\epsilon_i$ in (1) from $N(0, 1)$ and selected $\rho$ 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 $j \in I_o$,

  - $C_1$: *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 $k = 0, ..., 9$,

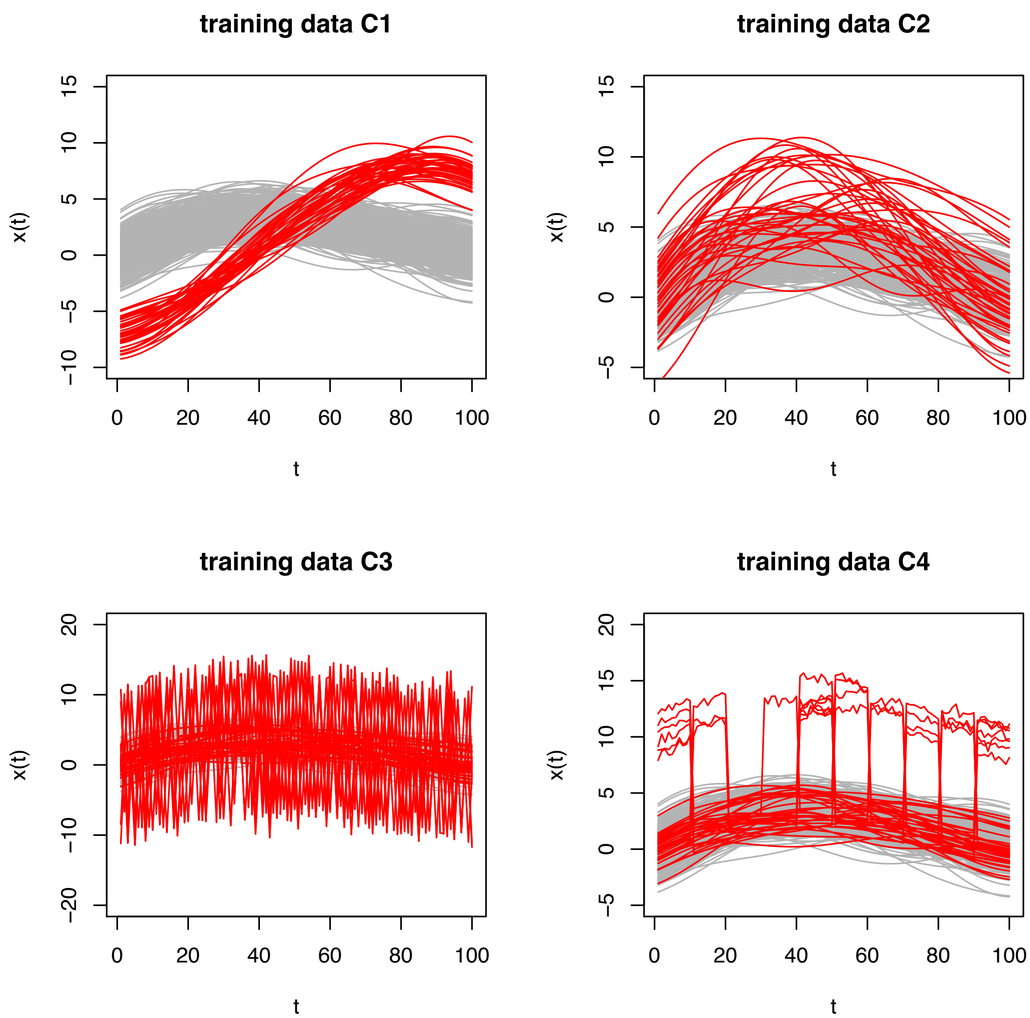
    $$x_j(t_{j,o+k}) = \tilde{x}_j(t_{j,o+k}) + \eta_{j,o+k},$$

    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:

- `TFBoost(L2)`: tree-based functional boosting with L2 loss
- `TFBoost(LAD)`: tree-based functional boosting with LAD loss
- `TFBoost(RR)`: tree-based functional boosting modified to follow the framework of RRBoost

- `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)

## 1.4 Results

| | $C_0$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|---|---|
| TFBoost(L2) | 0.144 (0.006) | 0.151 (0.009) | 0.206 (0.030) | 0.144 (0.006) | 0.146 (0.006) | 1.690 (0.271) |
| TFBoost(LAD) | 0.151 (0.010) | 0.202 (0.027) | 0.205 (0.030) | 0.150 (0.008) | 0.151 (0.008) | **0.158** (0.012) |
| TFBoost(RR) | 0.162 (0.017) | 0.193 (0.136) | 0.215 (0.110) | 0.158 (0.014) | 0.157 (0.012) | 0.159 (0.015) |
| FPPR | 0.137 (0.007) | 0.202 (0.076) | 0.164 (0.050) | 0.137 (0.007) | 0.149 (0.013) | 1.845 (0.517) |
| FGAM | **0.130** (0.005) | **0.143** (0.008) | **0.153** (0.016) | **0.130** (0.005) | **0.133** (0.005) | 1.205 (0.080) |
| RFPLM | **0.130** (0.006) | **0.130** (0.006) | **0.130** (0.006) | **0.130** (0.006) | **0.131** (0.006) | **0.130** (0.006) |
| MFLM | **0.129** (0.006) | 0.761 (0.054) | 0.269 (0.033) | **0.130** (0.006) | 0.138 (0.006) | 0.166 (0.014) |
| RFSIR | 0.137 (0.008) | 0.145 (0.014) | 0.157 (0.025) | 0.138 (0.007) | 0.142 (0.006) | 1.727 (0.587) |

Table 1: Summary statistics of test errors for data generated from $r_1$; displayed in the form of mean (sd).

| | $C_0$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|---|---|
| TFBoost(L2) | **0.181** (0.008) | **0.193** (0.010) | 0.223 (0.023) | **0.183** (0.009) | **0.184** (0.010) | 1.789 (0.347) |
| TFBoost(LAD) | 0.186 (0.010) | 0.257 (0.056) | 0.208 (0.020) | 0.188 (0.011) | **0.188** (0.011) | **0.195** (0.011) |
| TFBoost(RR) | 0.196 (0.014) | 0.225 (0.063) | **0.198** (0.019) | 0.202 (0.021) | 0.198 (0.023) | **0.203** (0.020) |
| FPPR | **0.181** (0.009) | 0.347 (0.133) | 0.288 (0.058) | **0.183** (0.011) | 0.196 (0.024) | 1.886 (0.545) |
| FGAM | 0.226 (0.012) | 0.243 (0.015) | 0.276 (0.027) | 0.233 (0.013) | 0.233 (0.012) | 1.343 (0.094) |
| RFPLM | 0.286 (0.014) | 0.286 (0.014) | 0.290 (0.016) | 0.287 (0.014) | 0.288 (0.017) | 0.286 (0.014) |
| MFLM | 0.285 (0.014) | 2.032 (0.099) | 0.325 (0.028) | 0.389 (0.023) | 0.626 (0.032) | 0.344 (0.024) |
| RFSIR | 0.183 (0.009) | **0.218** (0.061) | **0.202** (0.015) | 0.185 (0.010) | 0.193 (0.013) | 1.551 (0.569) |

Table 2: Summary statistics of test errors for data generated from $r_2$; displayed in the form of mean (sd).

| | $C_0$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|---|---|
| TFBoost(L2) | **0.305** (0.016) | **0.314** (0.020) | 0.518 (0.090) | **0.306** (0.015) | **0.308** (0.016) | 1.968 (0.360) |
| TFBoost(LAD) | 0.319 (0.019) | 0.382 (0.036) | 0.383 (0.049) | 0.317 (0.018) | 0.318 (0.016) | **0.326** (0.021) |
| TFBoost(RR) | 0.333 (0.032) | 0.370 (0.059) | **0.337** (0.041) | 0.337 (0.040) | 0.328 (0.028) | **0.335** (0.027) |
| FPPR | **0.303** (0.018) | 0.446 (0.105) | 0.606 (0.360) | 0.313 (0.022) | 0.318 (0.022) | 1.845 (0.453) |
| FGAM | 0.319 (0.017) | 0.331 (0.017) | 0.442 (0.061) | 0.321 (0.016) | 0.319 (0.017) | 1.445 (0.112) |
| RFPLM | 0.380 (0.018) | 0.379 (0.019) | 0.381 (0.018) | 0.379 (0.019) | 0.382 (0.019) | 0.379 (0.019) |
| MFLM | 0.377 (0.018) | 1.365 (0.080) | 0.485 (0.046) | 0.886 (0.063) | 2.165 (0.129) | 0.445 (0.028) |
| RFSIR | 0.310 (0.019) | **0.310** (0.019) | **0.337** (0.030) | **0.311** (0.020) | **0.311** (0.017) | 1.825 (0.677) |

Table 3: Summary statistics of test errors for data generated from $r_3$; displayed in the form of mean (sd).

| | $C_0$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|---|---|
| TFBoost(L2) | **0.321** (0.015) | **0.333** (0.015) | 0.681 (0.322) | **0.324** (0.014) | **0.328** (0.014) | 2.037 (0.267) |
| TFBoost(LAD) | **0.338** (0.017) | 0.404 (0.026) | 0.552 (0.161) | **0.340** (0.014) | **0.345** (0.017) | **0.361** (0.026) |
| TFBoost(RR) | 0.347 (0.036) | 0.490 (0.273) | 0.591 (0.667) | 0.365 (0.036) | 0.374 (0.048) | **0.360** (0.040) |
| FPPR | 0.362 (0.029) | 0.417 (0.045) | 0.538 (0.271) | 0.384 (0.040) | 0.415 (0.043) | 1.960 (0.386) |
| FGAM | 0.408 (0.019) | 0.417 (0.018) | **0.491** (0.054) | 0.415 (0.017) | 0.411 (0.019) | 1.659 (0.151) |
| RFPLM | 0.544 (0.032) | 0.544 (0.031) | 0.544 (0.032) | 0.555 (0.040) | 0.561 (0.039) | 0.543 (0.032) |
| MFLM | 0.538 (0.030) | 0.544 (0.032) | 0.826 (0.104) | 0.645 (0.037) | 0.827 (0.048) | 0.630 (0.048) |
| RFSIR | 0.341 (0.020) | **0.363** (0.021) | **0.421** (0.154) | 0.348 (0.018) | 0.354 (0.024) | 2.415 (0.512) |

Table 4: Summary statistics of test errors for data generated from $r_4$; displayed in the form of mean (sd).

| | $C_0$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|---|---|
| TFBoost(L2) | **0.583** (0.032) | **0.628** (0.045) | 1.725 (0.678) | **0.593** (0.033) | **0.590** (0.036) | 2.322 (0.278) |
| TFBoost(LAD) | 0.622 (0.034) | 0.694 (0.055) | 1.292 (0.315) | **0.633** (0.039) | **0.634** (0.030) | **0.677** (0.066) |
| TFBoost(RR) | 0.694 (0.092) | 0.869 (0.307) | 1.280 (1.596) | 0.703 (0.083) | 0.723 (0.084) | **0.686** (0.077) |
| FPPR | **0.608** (0.057) | 0.718 (0.175) | 0.967 (0.911) | 0.638 (0.049) | 0.673 (0.073) | 2.364 (0.482) |
| FGAM | 0.610 (0.041) | **0.670** (0.070) | **0.776** (0.100) | 0.643 (0.046) | 0.641 (0.048) | 1.909 (0.127) |
| RFPLM | 0.891 (0.045) | 1.045 (0.078) | 0.890 (0.045) | 0.889 (0.046) | 0.895 (0.047) | 0.888 (0.045) |
| MFLM | 0.881 (0.039) | 1.125 (0.067) | 1.698 (0.189) | 1.421 (0.080) | 2.527 (0.118) | 0.998 (0.052) |
| RFSIR | 0.677 (0.052) | 0.690 (0.063) | **0.821** (0.183) | 0.672 (0.052) | 0.650 (0.053) | 2.379 (0.518) |

Table 5: Summary statistics of test errors for data generated from $r_5$; displayed in the form of mean (sd).

### 1.5 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.