Generalized Additive Models & Functional Gradient Boosting with Geometrically Designed (GeD) Splines: Application to Insurance Data

Dimitrina S. Dimitrova[1], Emilio S. Guillén (presenter)[1], Vladimir K. Kaishev[1]

Insurance Data Science Conference, 17 - 18 June 2024.

[1] Faculty of Actuarial Science and Insurance, Bayes Business School.
Email: emilio.saenz-guillen@bayes.city.ac.uk

Insurance
Data
Science

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Motivation

- **Geometrically Designed Splines (GeDS)** (Kaishev et al., 2016, Dimitrova et al., 2023), — accurate and efficient tool for regression problems with one or two covariates and large datasets.

Motivation

- **Geometrically Designed Splines (GeDS)** (Kaishev et al., 2016, Dimitrova et al., 2023), — accurate and efficient tool for regression problems with one or two covariates and large datasets.
- GeD spline methodology is extended further by:
  1. **GAM-GeDS**: encompassing **Generalized Additive Models (GAM)**, thereby making GeDS highly multivariate.
  2. **FGB-GeDS**: incorporating **Functional Gradient Boosting (FGB)**, improving the construction of the underlying spline regression model.

Motivation

- **Geometrically Designed Splines (GeDS)** (Kaishev et al., 2016, Dimitrova et al., 2023), — accurate and efficient tool for regression problems with one or two covariates and large datasets.
- GeD spline methodology is extended further by:
  1. **GAM-GeDS**: encompassing **Generalized Additive Models (GAM)**, thereby making GeDS highly multivariate.
  2. **FGB-GeDS**: incorporating **Functional Gradient Boosting (FGB)**, improving the construction of the underlying spline regression model.

Implemented in the R package **GeDS**, available from CRAN:
https://cran.r-project.org/package=GeDS

GeDS estimation method

Free-knot spline regression technique based on a ***residual-driven (locally-adaptive) knot insertion scheme*** that produces a piecewise linear spline fit, over which ***smoother higher order spline fits*** are subsequently built.

GeDS estimation method

Free-knot spline regression technique based on a ***residual-driven (locally-adaptive) knot insertion scheme*** that produces a piecewise linear spline fit, over which ***smoother higher order spline fits*** are subsequently built.

GeDS method unfolds into two phases:

- **STAGE A** constructs a least squares linear spline fit to the data.
  - Starting with a straight-line, LS fit, which is then sequentially “broken” by iteratively introducing knots at those points ‘where the fit deviates most from the underlying functional shape determined by the data’.

GeDS estimation method

Free-knot spline regression technique based on a ***residual-driven (locally-adaptive) knot insertion scheme*** that produces a piecewise linear spline fit, over which ***smoother higher order spline fits*** are subsequently built.

GeDS method unfolds into two phases:

- **STAGE A** constructs a least squares linear spline fit to the data.
  - Starting with a straight-line, LS fit, which is then sequentially “broken” by iteratively introducing knots at those points ‘where the fit deviates most from the underlying functional shape determined by the data’.
- **STAGE B**
  - Builds smoother higher order spline fits using Schoenberg’s variation diminishing spline (VDS) approximation to the linear fit from Stage A.
  - For each higher spline order (quadratic, cubic...), compute the *averaging knot location* and re-estimate the spline coefficients by LS.

GAM-GeDS

The **Generalized Additive Model (GAM)** assumes the response variable, $Y \sim E.F.$, and relates its conditional expectation, $\mu = E\left[Y|X\right]$, to the predictor variables, $X_1, ..., X_P$, via a link function $g(\cdot)$:

$$g(\mu) = \alpha + \sum_{j=1}^{P} f_j(X_j), \text{ with } \mathbb{E}\left[f_j(X_j)\right] = 0, \quad j = 1, ..., P \tag{1}$$

Hastie and Tibshirani, 1990 — *local-scoring* and *backfitting* algorithms in conjunction with scatterplot smoothers, to fit GAMs.

GAM-GeDS

The **Generalized Additive Model (GAM)** assumes the response variable, $Y \sim E.F.$, and relates its conditional expectation, $\mu = E\left[Y|X\right]$, to the predictor variables, $X_1, ..., X_P$, via a link function $g(\cdot)$:

$$g(\mu) = \alpha + \sum_{j=1}^{P} f_j(X_j), \text{ with } \mathbb{E}\left[f_j(X_j)\right] = 0, \quad j = 1, ..., P \tag{1}$$

Hastie and Tibshirani, 1990 — *local-scoring* and *backfitting* algorithms in conjunction with scatterplot smoothers, to fit GAMs.

GAM with GeD Splines: Local-scoring algorithm using GeD splines as the function smoothers, f_j, within the backfitting algorithm.

FGB-GeDS

- **Functional Gradient Boosting** (Friedman, 2001): ensemble machine learning technique that iteratively combines multiple simple models ('weak-learners'), each striving to enhance the performance of the previous accumulative model.

  $\longrightarrow$ ***Component-wise Gradient Boosting*** (Bühlmann and Yu, 2003; Schmid and Hothorn, 2008): boosting algorithm for fitting additive models, inherently performing variable selection; implemented in **mboost** package (boosting with P-splines).

FGB-GeDS

- **Functional Gradient Boosting** (Friedman, 2001): ensemble machine learning technique that iteratively combines multiple simple models ('weak-learners'), each striving to enhance the performance of the previous accumulative model.

  $\longrightarrow$ ***Component-wise Gradient Boosting*** (Bühlmann and Yu, 2003; Schmid and Hothorn, 2008): boosting algorithm for fitting additive models, inherently performing variable selection; implemented in **mboost** package (boosting with P-splines).

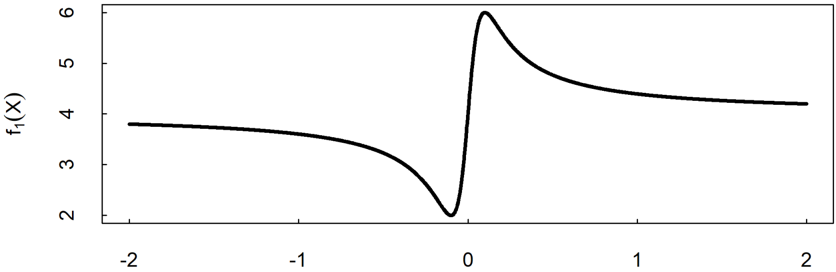
FGB with GEDS base-learners

- Flexible control of the strength of the base-learners:
  1. Weak GeDS initial learner + few boosting iterations with strong GeDS learners.
  2. Boosting iterations with weak GeDS learners based on single knot addition with memory.
- Optimal number of boosting iterations determined by a **stopping rule** based on a ratio of consecutive deviances.
- Final boosted fit expressed as a **single spline model**.

Simulated Data Application

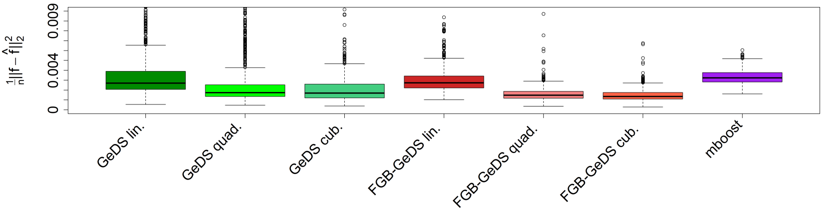
Consider the function:

$$f_1(x) = 40\frac{x}{1+100x^2} + 4 \quad , \quad x \in c(-2, 2)$$

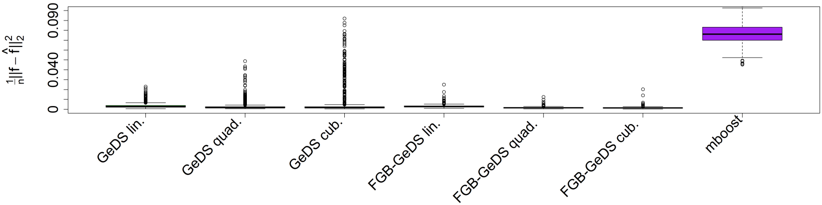


- Generate $1,000$ random samples, $\{X_i, Y_i\}_{i=1}^{N}$ with $Y_i \sim N(\mu_i, \sigma)$ with $\sigma = 0.2$, $\mu_i = \eta_i = f_1(X_i)$ and $X_i \sim U[-2, 2]$, $i = 1, ..., N$, where $N = 500$.

$$\begin{cases} \textbf{GeDS}: 10 \text{ int. knots} \\ \textbf{FGB-GeDS}: \text{initial learner with } 2 \text{ int. knots + } 1 \text{ boosting iter. with } 8 \text{ int. knots} \\ \textbf{mboost}: 10,000 \text{ boosting iter. with } 36 \text{ knots p/iter.} \end{cases}$$

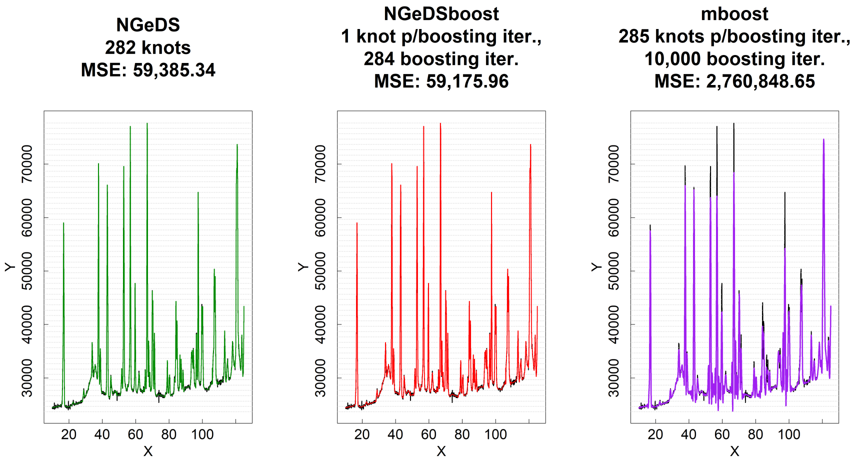


And setting **mboost** to have 10 int. knots p/boosting iter. instead:



Real Data Application

- High pressure neutron barium-iron arsenide ($BaFe_2As_2$) powder diffraction data (Kimber et al., 2009), with number of observations $N = 1151$.



Insurance Data Application

Motorcycle insurance data `swmotorcycle` available through the `R` package `CASdatasets` (Dutang and Charpentier, 2020).

$\longrightarrow$ We follow Delong et al., 2021 and model **gamma claim sizes**:

1. Gamma GLM regression + Gamma Neural Network regression.
2. `mboost`: FGB with P-splines.
3. GAM-GeDS.
4. FGB-GeDS.

- *Response*: `ClaimAmount/ClaimNb`, i.e., the average claim size.
- *Covariates*: `OwnerAge`; `Gender`; `Area`, `RiskClass`; `VehAge`.
- *Train/Test split*: **80%/20%**.
  - Simulate $\underline{100}$ different splits of data.

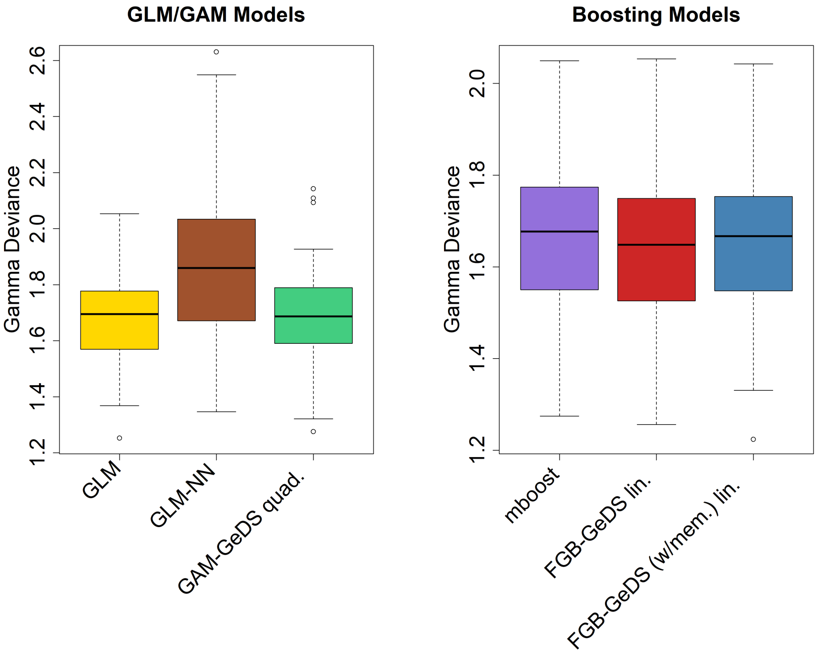

GLM/GAM Models
Gamma Deviance
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
GLM
GLM-NN
GAM-GeDS quad.
Boosting Models
Gamma Deviance
2.0
1.8
1.6
1.4
1.2
mboost
FGB-GeDS lin.
FGB-GeDS (w/mem.) lin.

Table 1: GLM/GAM Models

	Gamma Deviance			Internal knots
	Train Data	Test Data	Time (sec.)	(OwnerAge+VehAge)
GLM	1.585727	1.694797	0.008708	-
GLM NN	1.719903	**1.859394**	167.224576	-
GAM-GeDS quadratic	1.557612	**1.686492**	0.671260	5

Table 2: Boosting Models

	Gamma Deviance			Internal knots	Boosting
	Train Data	Test Data	Time (sec.)	p/boosting iter. (OwnerAge+VehAge)	iterations
mboost	1.610290	**1.676810**	0.156095	4	100
FGB-GeDS linear (2 starting knots)	1.575972	**1.648345**	0.130963	2	1
FGB-GeDS w/mem. linear (1 starting knot)	1.575536	1.667158	0.129040	1	3

Bühlmann, P., & Yu, B. (2003). Boosting with the l2 loss. *Journal of the American Statistical Association*, *98*(462), 324–339. https://doi.org/10.1198/016214503000125

Delong, Ł., Lindholm, M., & Wüthrich, M. V. (2021). Making tweedie's compound poisson model more accessible. *European Actuarial Journal*, *11*(1), 185–226. https://doi.org/10.1007/s13385-021-00264-3

Dimitrova, D. S., Kaishev, V. K., Lattuada, A., & Verrall, R. J. (2023). Geometrically designed variable knot splines in generalized (non-)linear models. *Applied Mathematics and Computation*, *436*, 127493. https://doi.org/https://doi.org/10.1016/j.amc.2022.127493

Dutang, C., & Charpentier, A. (2020). *Casdatasets: Insurance datasets* [R package version 1.0-11].

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine.. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Hastie, T., & Tibshirani, R. (1990). Generalized additive models. *Monographs on statistics and applied probability. Chapman & Hall*, *43*, 335.

Kaishev, V. K., Dimitrova, D. S., Haberman, S., & Verrall, R. J. (2016). Geometrically designed, variable knot regression splines. *Computational Statistics*, *31*(3), 1079–1105. https://doi.org/10.1007/s00180-015-0621-7

Kimber, S. A. J., Kreyssig, A., Zhang, Y.-Z., Jeschke, H. O., Valentí, R., Yokaichiya, F., Colombier, E., Yan, J., Hansen, T. C., Chatterji, T., McQueeney, R. J., Canfield, P. C., Goldman, A. I., & Argyriou, D. N. (2009). Similarities between structural distortions under pressure and chemical doping in superconducting BaFe2As2. *Nature Materials*, *8*(6), 471–475.

Schmid, M., & Hothorn, T. (2008). Boosting additive models using component-wise p-splines. *Computational Statistics & Data Analysis*, *53*(2), 298–311. https://doi.org/https://doi.org/10.1016/j.csda.2008.09.009