Augmented Spline Regression for Advanced Data Analysis: Generalized Additive Models & Functional Gradient Boosting with Geometrically Designed (GeD) Splines

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

[1]Faculty of Actuarial Science and Insurance, Bayes Business School.

RSS International Conference 2024
THE UK'S ANNUAL SHOWCASE FOR STATISTICS & DATA SCIENCE
BRIGHTON 2-5 September 2024

1. Geometrically Designed Splines (GeDS)

Free-knot spline regression technique based on a ***residual-driven (locally-adaptive) knot insertion scheme*** that produces an initial piecewise linear spline fit, over which ***smoother higher order spline fits*** are subsequently built (Kaishev et al., 2016, Dimitrova et al., 2023).

❉ 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.

- Applications in highly multivariate contexts: AI (e.g., image recognition/processing); robotics (e.g. motion planning for humanoid robots).
- Implemented in the R package **GeDS**, available from `CRAN`: https://cran.r-project.org/package=GeDS.

GeD Spline Regression

GeDS method unfolds into two stages:

- **STAGE A** constructs a least squares (LS) 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", based on a measure defined by residuals.
- **STAGE B** builds smoother higher order spline fits using Schoenberg's variation diminishing spline (VDS) approximation, based on the linear fit from Stage A.
  - For each higher spline order (quadratic, cubic, ...), compute the corresponding *averaging knot location* and re-estimate the spline coefficients by LS.

Properties of GeDS estimated knots and regression coefficients:

- Knots possess *Schoenberg variation diminishing optimality*.
- *Asymptotic normality* of estimators in the case of normal noise, allowing for the construction of *pointwise asymptotic confidence intervals*.
- Asymptotic conditions on the rate of growth of the knots for *negligible bias/variance ratio* of the GeDS estimators.

2. Generalized Additive Models with GeD Splines

GAM with GeD Splines: *Local-scoring* algorithm using GeD splines as the function smoothers, f_j, at each *backfitting* iteration.

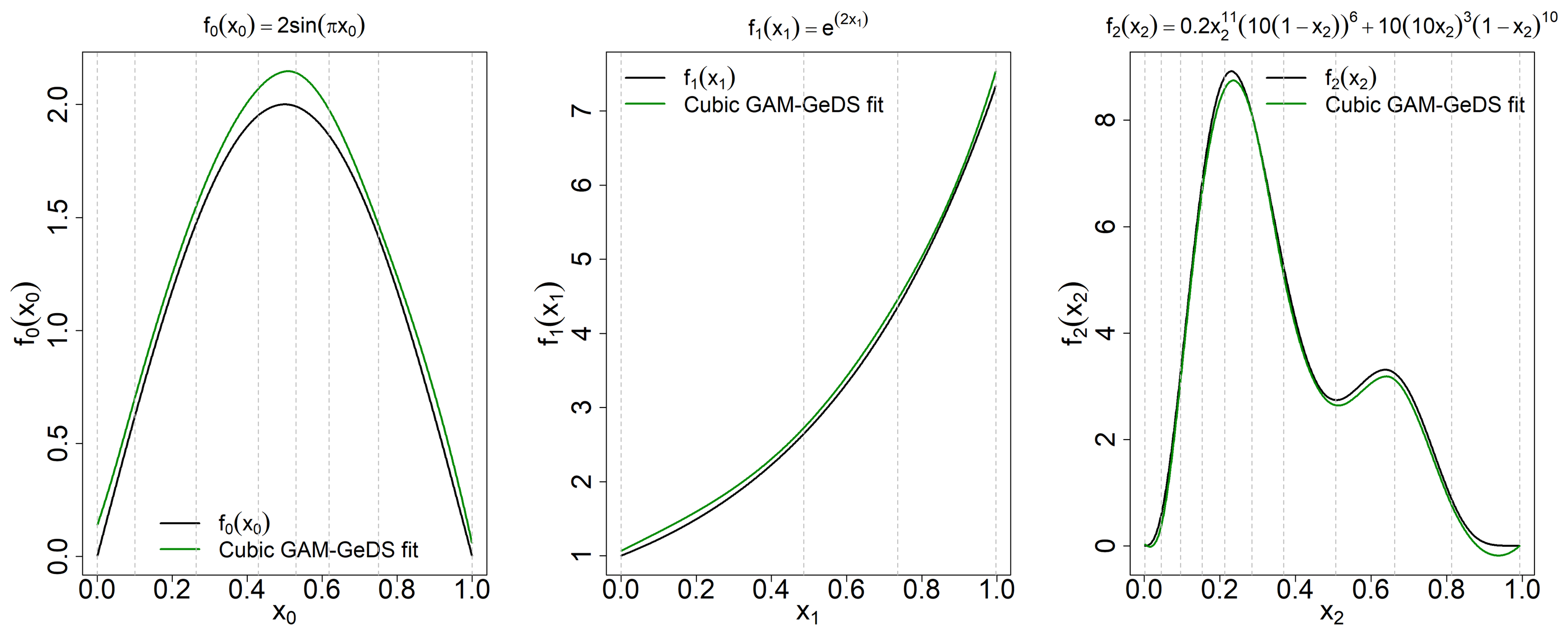
Example (Gu and Wahba, 1991):

$$f(\mathbf{x}) = \underbrace{2 \times \sin(\pi \times x_0)}_{f_0(x_0)} + \underbrace{\exp(2x_1)}_{f_1(x_1)} + \underbrace{0.2x_2^{11}(10(1-x_2))^6 + 10(10x_2)^3(1-x_2)^{10}}_{f_2(x_2)}$$

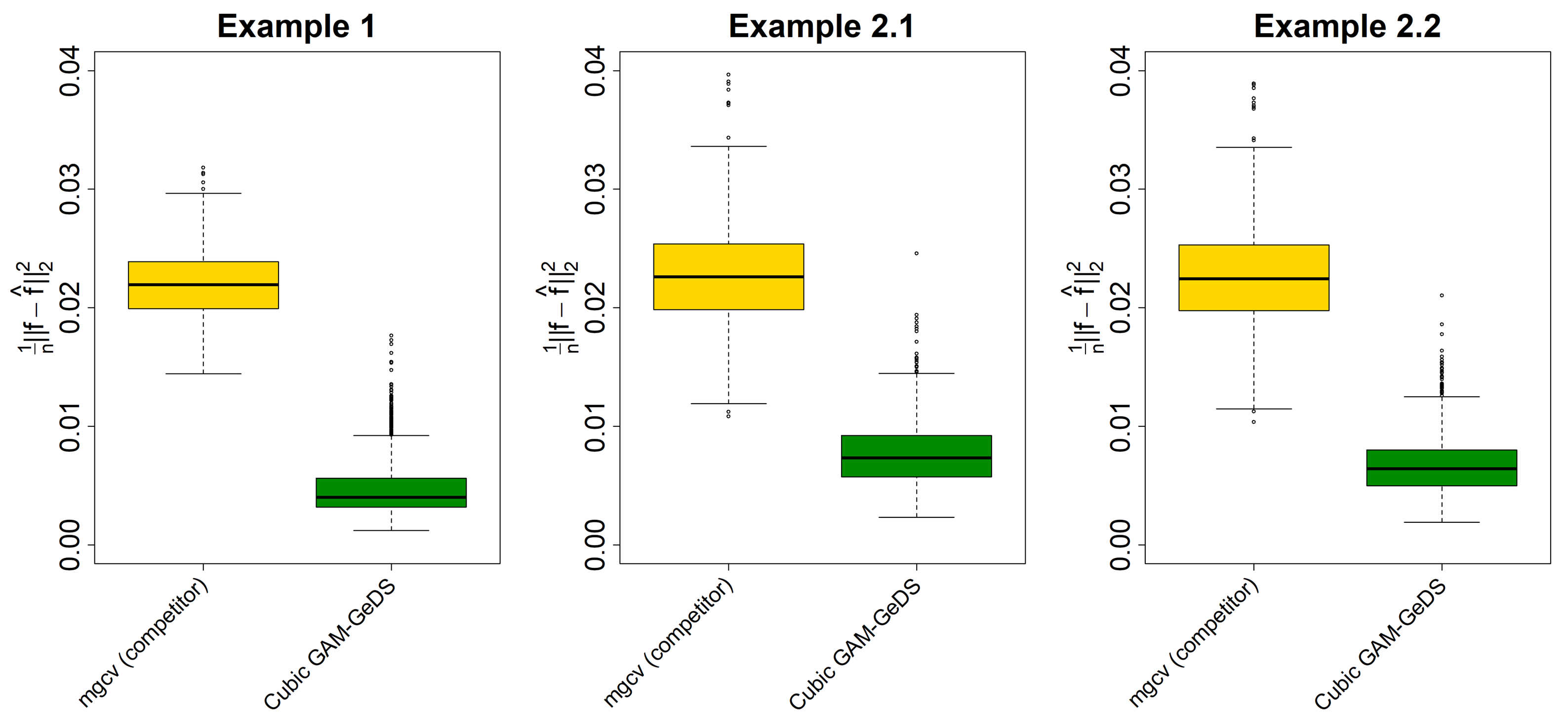
In **Example 1**, we fit $y = f(\mathbf{x}) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, including a noise predictor, x_3. In **Example 2** we replace $f(x_0)$ by a factor variable x_0 with 4 levels: **2.1** includes the noise predictor x_3, **2.2** deletes it. For all the examples, $x_0, x_1, x_2, x_3 \sim$ Uniform$(0,1)$.

GAM-GeDS (partial) fits + MSE boxplots

Cubic GAM-GeDS partial fits for example 1:



MSE boxplots w.r.t. $f(\mathbf{x})$, examples 1, 2.1 & 2.2:



3. Functional Gradient Boosting with GeD Splines

Deals with major limitations of mainstream Gradient Boosting algorithms:

- "Prone to overfitting"
  - ➠ FGB-GeDS determines the optimal number of boosting iterations through a **stopping rule** based on a ratio of consecutive deviances.
- "Many parameters and unstable performance"
  - ➠ Strength of the base learners is **automatically regulated by the GeDS technique** at each boosting iteration, and flexibly controlled through the GeDS parameters.
- "Black-box models"
  - ➠ Final FGB-GeDS boosted model is expressed as a **single spline model**, which simplifies its evaluation and enhances interpretability.

Application: Compute the Fourier Transform of Gold (Au)

Given a sample, $\mathcal{L} = \{F(Q_i), Q_i\}_{i=1}^N$, $0 < Q_1 < ... < Q_N < \widetilde{Q}_{\max}$, we are interested in estimating the **Fourier transform** (imaginary part):

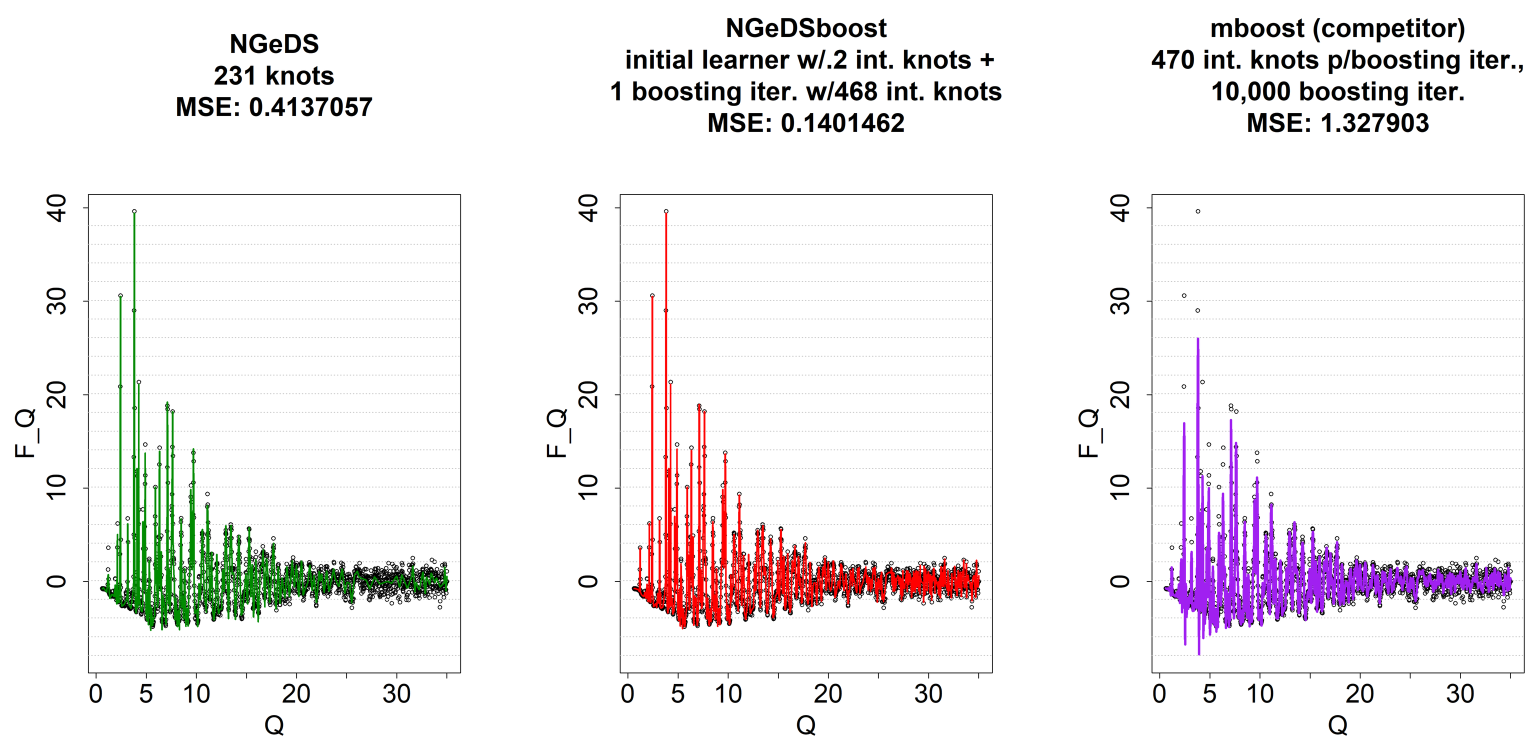
$$G(r) = \frac{2}{\pi}\int_0^{Q_{\max}} F(Q)\sin QrdQ.$$

Assuming $Q_{\max}$ is known, this involves two steps:

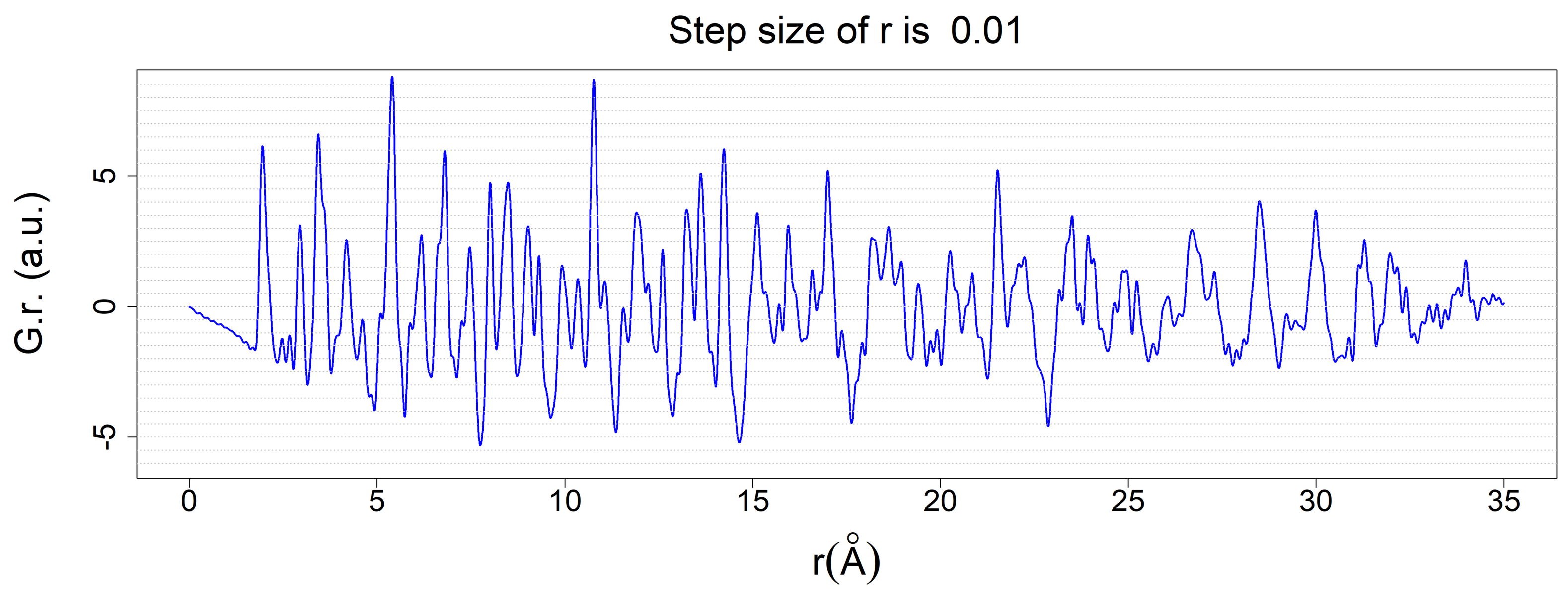
Step 1. Estimate $F(Q)$ through a GeDS fit $\equiv S(Q)$ to the sample $\mathcal{L}$.
Step 2. Compute $G(r)$ using the fitted GeDS model, $S(Q)$.

For the time being, let $Q_{\max} \equiv \widetilde{Q}_{\max}$, though in general $Q_{\max} < \widetilde{Q}_{\max}$ (signal in data prevails up to a certain point), and needs to be optimally estimated.

Step 1: Estimate $F(Q)$



Step 2: Compute the Fourier transform $G(r)$



References

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

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