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]

RSS International Conference 2024

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

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Geometrically Designed Splines (GeDS)

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 (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.

Geometrically Designed Splines (GeDS)

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 (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

4. Functional Gradient Boosting with GeDS (FGB-GeDS)

- **Functional Gradient Boosting** (Friedman, 2001).

❊ **FGB-GeDS deals with major limitations of mainstream boosting algorithms:**

- **"Prone to overfitting"**

➠ Optimal number of boosting iterations determined by a **stopping rule** based on a ratio of consecutive deviances.

- **"Large number of parameters and unstable performance"**

➠ Strength of the base learners is **automatically regulated by the GeDS** technique itself, and flexibly controlled through the GeDS parameters.

- **"Black-box models"**

➠ Final FGB-GeDS boosted model expressed as a **single spline model**, which simplifies its evaluation and enhances interpretability.

Task: Fourier Transform Computation of Materials Science Data

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

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

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

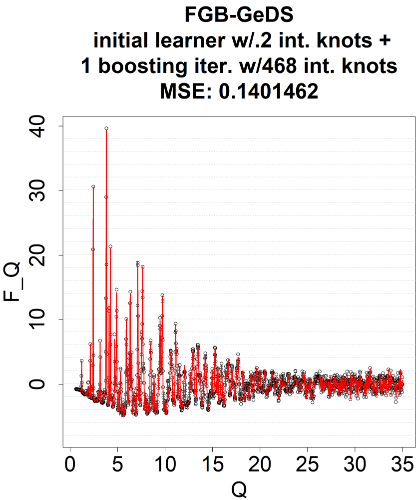
Step 1. Estimate $F(Q)$ through a GeDS fit $\equiv S(Q)$ to the sample $\mathcal{L}$.

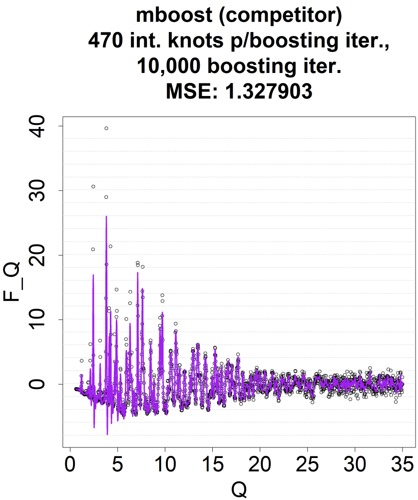
Step 2. Compute $G(r)$ using the fitted GeDS model, $S(Q)$.

For the time being, let us assume $Q_{\text{max}} \equiv \widetilde{Q}_{\text{max}}$, though in general $Q_{\text{max}} < \widetilde{Q}_{\text{max}}$:

- Signal in the data prevails up to a certain point; beyond this, only noise remains.
- Sequential (and costly) data collection: cut off at the appropriate Q_{max} for an optimal experimental design.

Step 1. Fit $F(Q)$, e.g, with an FGB-GeDS model





Step 2. Compute the Fourier transform of gold

Proposition

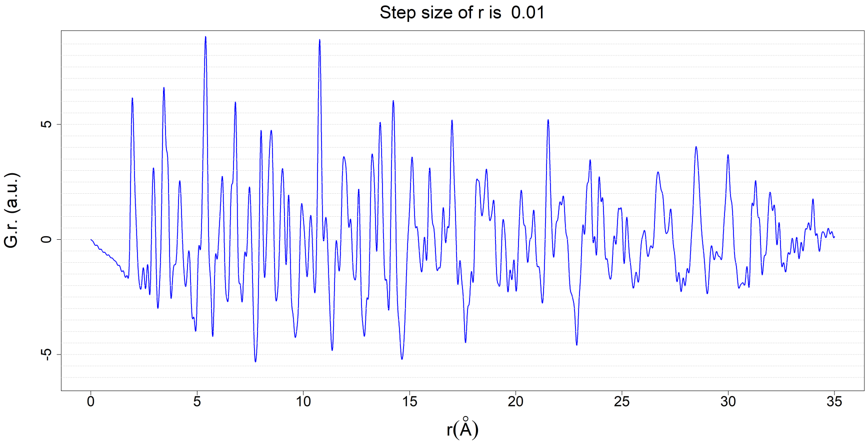
For the $\sin()$ transform,

$$G(r) = \frac{2}{\pi} \int_0^{Q_{\text{max}}} F(Q) \sin(Qr) dQ$$

of the function $F(Q)$, approximated by $S(Q)$ of order $n = 2s$, $s = 1, 2, 3, \ldots$ we have

$$G(r) \approx \frac{(-1)^s 2(n-1)!}{\pi r^n} \sum_{i=1}^{p} \hat{\theta}_i \left(t_{i+n} - t_i\right) \sum_{j=i}^{i+n} \frac{\sin(t_j r)}{\prod_{\substack{l=i \\ l \neq j}}^{i+n} (t_j - t_l)},$$

where $r \in \mathbb{R}^+$, $p = k + n$; $\hat{\theta}_i$, $i = 1, \ldots, p$ are the GeDS regression coefficients.


Step size of r is 0.01
G.r. (a.u.)
r(Å)
0
5
10
15
20
25
30
35
-5
0
5

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

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

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