Boosting distributional soft regression trees

Nikolaus Umlauf[1], Johannes Seiler[1],
Mattias Wetscher[1], Nadja Klein[2]

[1] Universität Innsbruck, Austria
[2] Research Center Trustworthy Data Science and Security (UA Ruhr) and Department of Statistics (Technische Universität Dortmund), Germany

E-mail for correspondence: Nikolaus.Umlauf@uibk.ac.at

Abstract: Distributional soft trees offer a flexible and effective way to model full probabilistic regression models. On the one hand, unlike classical regression trees and forests, which use hard splits to partition data, soft trees provide smooth estimates through soft splits, leading to improved performance in many cases due to reduced approximation error. On the other hand, compared to structured additive distributional regression, distributional soft trees allow for more complex interactions of possibly high-dimensional feature vectors. In this article, we introduce a boosted version of a distributional adaptive soft regression tree that can be applied to very large datasets while performing variable selection on the fly. We demonstrate the strong predictive capabilities of this method through a complex regression problem involving the spatial mapping of recent child anaemia risk data in sub-Saharan Africa. Our results further highlight the potential of the proposed boosting method in large-scale complex regression problems.

Keywords: Boosting; GAMLSS; soft trees; variable selection.

1 Introduction

Distributional regression involves modeling the entire distribution of a response variable, rather than just its mean or median. This can provide a more comprehensive understanding of the relationship between covariates and the response, as well as enable more accurate probabilistic predictions beyond point forecasts. While there are various methods for obtaining a distributional model, this paper focuses on the class of structured additive distributional regression, also known as generalized additive models for location, scale, and shape (GAMLSS; Rigby and Stasinopoulos, 2005). GAMLSS can model every parameter of an arbitrary parametric target

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

distribution through input features, resulting in a probabilistic prediction model.

Distributional (adaptive) soft regression trees (DAdaSoRT; Umlauf and Klein, 2022) offer a flexible and effective way to model full probabilistic regression models, and have recently been shown to be a promising alternative to structured distributional methods. One key advantage of these DAdaSoRT, which embed classical soft trees into the distributional framework of GAMLSS, is the smoothness of their estimates on respective distributional parameters, which is achieved through the use of soft splits rather than hard splits. This smoothness can reduce approximation error and improve performance in many cases. In fact, DAdaSoRT have been shown to outperform both classical GAMLSS and full probabilistic distributional forests (DF; Schlosser et al., 2019) in certain situations. Figure 1 shows a simple

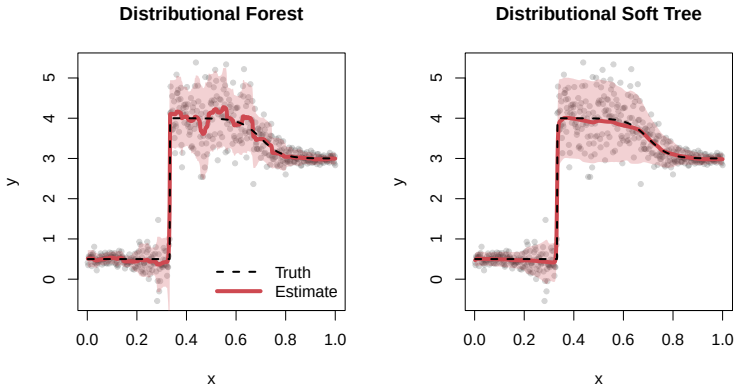

FIGURE 1. Simulated 2D data: Shown are the estimates for $E(Y|x)$ of a DF using 2000 trees (left) and a DAdaSoRT (right). The solid red lines represent the mean estimates, and the red shaded areas depict the 5% and 95% estimated quantiles of $E(Y|x)$. The dashed black lines show the true mean function.

2D regression example with a classical DF compared to a DAdaSoRT. Although DF is estimated with 2000 trees, the resulting estimate is quite wiggly and tends to overfit the data compared to the DAdaSoRT in the right panel of Figure 1. As mentioned before, the reason for this is mainly the hard splitting rule of classical trees and forests, which favors an approximation error that can even be amplified when modeling high-dimensional covariate interactions. The example also illustrates that DAdaSoRT can represent both smooth transitions and abrupt jumps of a function.

This article presents a new boosting algorithm designed to further improve the flexibility of distributional modeling using soft trees. Compared to the estimation method of Umlauf and Klein (2022), the boosting algorithm needs far less tuning, can be applied to very large data sets and selects the

most relevant features in the data on the fly . The latter capability is particularly useful as it helps to reduce the complexity of the modeling process, favours sparse and thus often better interpretable models and improves the accuracy of the results.

2 Model and Boosting Algorithm

DAdaSoRTs are introduced in Umlauf and Klein (2022), and we follow their notation for simplicity. Now, suppose there is data $\mathbf{y} = (y_1, \ldots, y_n)^\top$, such that for each output y_i, $i = 1, \ldots, n$ there is a q-dimensional feature vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})^\top$ available and $\mathbf{X} = (\mathbf{x}_1 | \ldots | \mathbf{x}_n)^\top$ is the $n \times q$ feature matrix. Assume $y \sim D_y\left(h_1(\theta_1) = \eta_1,\ h_2(\theta_2) = \eta_2, \ldots, h_K(\theta_K) = \eta_K\right)$, where D_y denotes a parametric distribution for the response variable y and $h_k(\cdot)$ are monotonic and twice differentiable link functions mapping to the distributional predictors η_k which are modeled by soft trees. Following Umlauf and Klein (2022), we use adaptive soft tree structures given by $\boldsymbol{\eta}_k \equiv f_k(\mathbf{X}) = \beta_{k,0} + \sum_{j=1}^{J_k} P_{k,j}(\mathbf{X}, \boldsymbol{\Omega}_{(k,j)})\beta_{k,j}$,where for $k = 1, \ldots, K$, $P_{k,j}(\cdot)$ represent the path probabilities from a soft splitting rule, $\boldsymbol{\Omega}_{(k,j)}$, $\beta_{k,0}$ and $\beta_{k,j}$ are weights that need to be estimated and J_k is the number of "basis functions" of $f_k(\cdot)$ obtained from left and right soft splitting.
In contrast to the multivariate soft splitting of Umlauf and Klein (2022), we use an univariate soft split for $P_{k,j}(\cdot)$, to automatically incorporate variable selection in the final DAdaSoRT by selecting only the best performing feature $\mathbf{x}_q$ according to the current log-likelihood contribution. Specifically, to set up a boosting type algorithm, we specify at each iteration $t = 0, \ldots, T$ and for each distributional predictor $\boldsymbol{\eta}_k$ the updating equation

$$\boldsymbol{\eta}_k^{[t+1]} = \boldsymbol{\eta}_k^{[t]} + \nu \cdot f_k^{[t]}(\mathbf{X}), \tag{1}$$

where ν is a step length parameter (e.g., $\nu = 0.1$). Therefore, predictors are improved slowly while each tree is estimated with maximum likelihood using offsets $\boldsymbol{\eta}_k^{[t]}$. In addition, the depth of the trees is kept small, which is a tuning parameter, so that a single $f_k^{[t]}(\cdot)$ only contributes a small amount to the overall model fit, similar to Bayesian additive regression trees (BART; Chipman et al. 2010). Moreover, instead of using all observations n for fitting a single tree in iteration t we only use a randomly selected subset $\boldsymbol{s}^{[t]} \subset \{1, \ldots, n\}$ of the data, i.e., each tree is build using (possibly) different data batches $\mathbf{X}_{\boldsymbol{s}^{[t]}}$. This leads to a regularization such that convergence of the algorithm is achieved when the log-likelihood evaluated on the batches becomes stationary around a certain level, i.e., in most applications, only enough boosting iterations T need to be provided without further tuning. In addition, it can be applied to very large data sets since the batchwise updating requires only a relatively small computational cost. We call this novel method batchwise boosting DAdaSoRT. An implementation is provided in the R package **softtrees** (Umlauf, 2023), see `help("BB-DAdaSoRT")`.

3 Child Anaemia Risk in Sub-Saharan Africa

Anaemia is a major health issue in low- and middle-income countries, particularly in sub-Saharan Africa, where over 50% of children under five are affected. We analyze haemoglobin (HgB) in a yet unexplored large-scale dataset with $> 340k$ observations from Demographic and Health Surveys. The data include climate, environmental and geospatial data. To perform model calibration checks, we split the data randomly into training and testing sets, with 80% of the data allocated to training and 20% to testing. We then benchmark the performance of a classical Bayesian additive model for location, scale, and shape (BAMLSS, Umlauf et al., 2018) with our proposed DAdaSoRT model. Notably, DAdaSoRT exhibited a considerably faster runtime, requiring approximately 6.5 hours to process 200 batches of 10000 data points, compared to approximately 65 hours for the Bayesian GAMLSS with 8000 MCMC iterations. Ultimately, we found that a model with skew exponential power type 3 distribution as implemented in the **gamlss.dist** package (Stasinopoulos and Rigby, 2022), achieved the best performance based on the out-of-sample continuous rank probability score (CRPS). Without further tuning, the skill score of this model, compared to a simple Gaussian intercept-only model, demonstrated an 11.13% improvement, while the skill score of the best-fitting BAMLSS model yielded an 11.05% improvement, indicating a marginal enhancement. However, this outcome is particularly promising as the identification of interactions is automated in our proposed DAdaSoRT model, unlike in BAMLSS. Additionally, compared to conventional distributional trees and forests, estimation with our approach is significantly more efficient and currently not feasible with available implementations of distributional trees or forests.

Figure 2 displays the out-of-sample quantile residuals of the final model. The histogram indicates approximately normally distributed residuals, while the worm plot reveals slight, yet statistically significant deviations from a zero mean for higher estimated quantiles. Overall, the model appears to be well calibrated, even on the test data.

The left panel of Figure 3 presents the log-likelihood contributions for each of the selected variables, indicating that land type and the age of the child in months are the two most influential factors. In the right panel, we depict the estimated spatial risk for $Pr(\text{HgB} < 110\,\text{g/L})$, illustrating substantial variations across the continent. Figure 4 showcases the marginal effects on $Pr(\text{HgB} < 110\,\text{g/L})$. All the figures related to $Pr(\text{HgB} < 110\,\text{g/L})$ demonstrate the exceptional ability of the proposed DAdaSoRT model to accurately approximate both sharp and smooth transitions. For instance, we observe sharp regional changes in the map of Figure 3 in contrast to the very smooth estimated effects in Figure 4.

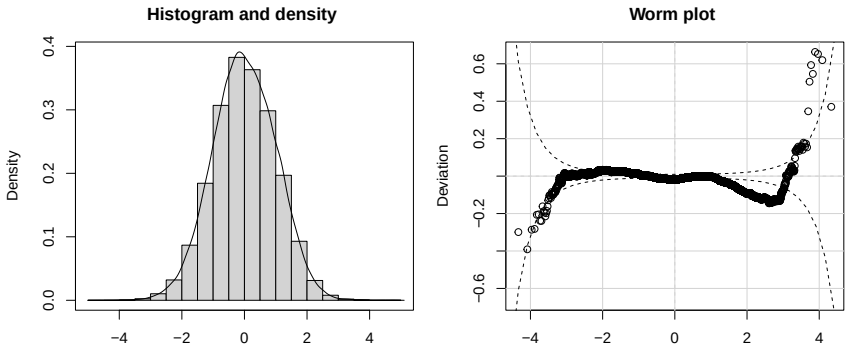

FIGURE 2. Model calibration plots using out-of-sample quantile residuals.

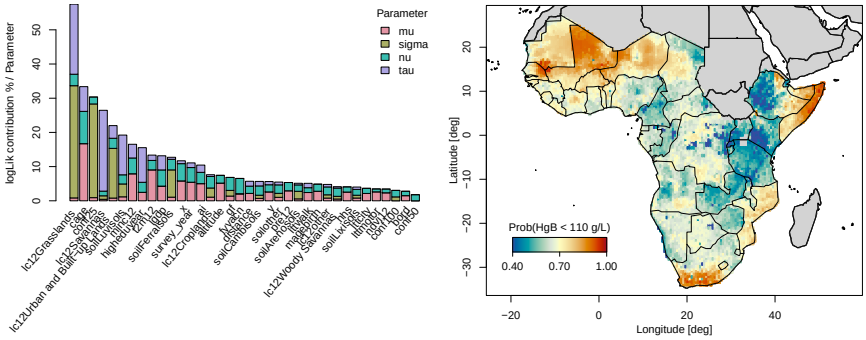

FIGURE 3. Variable log-likelihood contributions (left) and estimated anaemia risk for female infants at 30 months of age in 2020 (right).

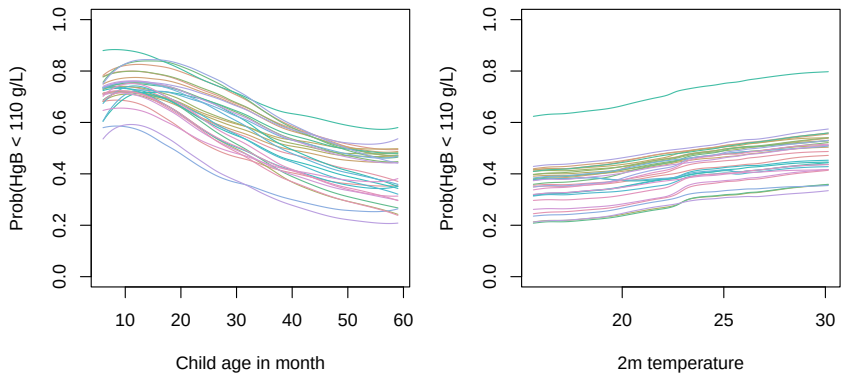

FIGURE 4. Estimated marginal effects on $Pr(\text{HgB} < 110\,\text{g/L})$ for each country in the data set.

Acknowledgments: This project was funded by the FWF grant #33941 and the DFG through the Emmy Noether grant KL 3037/1-1.

References

Rigby, R.A. and Stasinopoulos, D.M. (2005) Generalized Additive Models for Location, Scale and Shape. *Appl. Stat.*, **54**(3), 507 – 554.

Chipman, H.A., George, E.I, and McCulloch, R.E. (2010). BART: Bayesian Additive Regression Trees. *Ann. Appl. Stat.*, **4**(1), 266 – 298.

Umlauf, N., Klein, N, and Zeileis, A. (2018) BAMLSS: Bayesian additive models for location, scale and shape (and beyond). *J. Comput. Graph. Stat.*, **27**(3), 612 – 627.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A. (2019) Distributional Regression Forests for Probabilistic Precipitation Forecasting in Complex Terrain. *Ann. Appl. Stat.*, **13**(3), 1564 – 1589.

Umlauf, N. and Klein, N. (2022). Distributional Adaptive Soft Regression Trees. *arXiv*, URL: `https://arxiv.org/abs/2210.10389`

Stasinopoulos, M. and Rigby, R. (2022) **gamlss.dist**: Distributions for Generalized Additive Models for Location, Scale and Shape. R package version 6.0-5, URL: `https://CRAN.R-project.org/package=gamlss.dist`

Umlauf, N (2023). **softtrees**: Soft Distributional Regression Trees and Forests. R package version 1.1, URL: `https://github.com/freezenik/softtrees`