RBF: An R package for computing a robust backfitting estimation procedure for additive models

Alejandra M. Martínez[1] **and Matias Salibian-Barrera**[2]

1 Departamento de Ciencias Básicas, Universidad Nacional de Luján, Argentina **2** Department of Statistics, University of British Columbia, Canada

DOI: 10.21105/joss.0XXXX

Software

- Review
- Repository
- Archive

Editor: Editor Name

Submitted: 01 January XXXX
Published: 01 January XXXX

Summary

Although highly flexible, non-parametric regression models typically require large sample sizes to be fit reliably, particularly when many explanatory variables are present in the model. Additive models provide an alternative that is more flexible than linear models, not affected by the curse of dimensionality, and also allow the exploration of individual covariate effects. Standard algorithms to fit these models can be highly susceptible to the presence of a few atypical or outlying observations in the data.

`RBF` (Salibian-Barrera & Martínez, 2020) is an R package that implements a robust estimator for additive models based on the *backfitting* algorithm.

Statement of Need

The purpose of `RBF` is to provide a kernel-based estimation procedure for additive models that is resistant to the presence of potential outliers. The package also implements several modeling tools, including functions to produce diagnostic plots, obtain fitted values and compute predictions.

Implementation Goals

`RBF` provides a user interface similar to that of the R package `gam` (T. Hastie, 2019), which implements the standard non-robust kernel-based fit for additive models via the backfitting algorithm.

Background

Additve models offer a non-parametric generalization of linear models (T. J. Hastie & Tibshirani (1990)). They are flexible, interpretable and avoid the *curse of dimensionality* which is due to the fact that, as the number of explanatory variables increases, neighbourhoods become more sparse, and much fewer training observations are available to estimate the regression function at any one point.

Let Y be the response variable and $\mathbf{X} = (X_1, \ldots, X_d)^\top$ a vector of explanatory variables, then an additive regression model postulates that

$$Y \;=\; \mu + \sum_{j=1}^{d} g_j(X_j) \,+\, \epsilon\,, \tag{1}$$

where the error ϵ is independent of $\mathbf{X}$ and centered at zero. The objects to be estimated are the location parameter $\mu \in \mathbb{R}$ and the smooth functions $g_j : \mathbb{R} \to \mathbb{R}$. Note that when $g_j(X_j) = \beta_j X_j$ for some $\beta_j \in \mathbb{R}$, the above model reduces to a standard linear regression one.

The backfitting algorithm (Friedman & Stuetzle (1981)) fits the additive model above using kernel regression estimators for the smooth components g_j. It is based on the following observation: under Equation 1 the additive components satisfy $g_j(x) = E[Y - \mu - \sum_{\ell \neq j} g_\ell(X_\ell) | X_j = x]$. Each g_j is iteratively computed by smoothing the partial residuals as functions of X_j.

It is well known that these estimators can be seriously affected by a relatively small proportion of atypical observations in the training set. Boente et al. (2017) proposed a robust version of backfitting, which is implemented in the RBF package. Intuitively, the idea is to use the backfitting algorithm with robust smoothers, such as kernel-based estimators (Boente & Fraiman (1989)). These estimators solve the following optimization problem:

$$\min_{\mu, g_1, \ldots, g_d} E\left[\rho\left(\frac{Y - \mu - \sum_{j=1}^{d} g_j(X_j)}{\sigma}\right)\right]$$

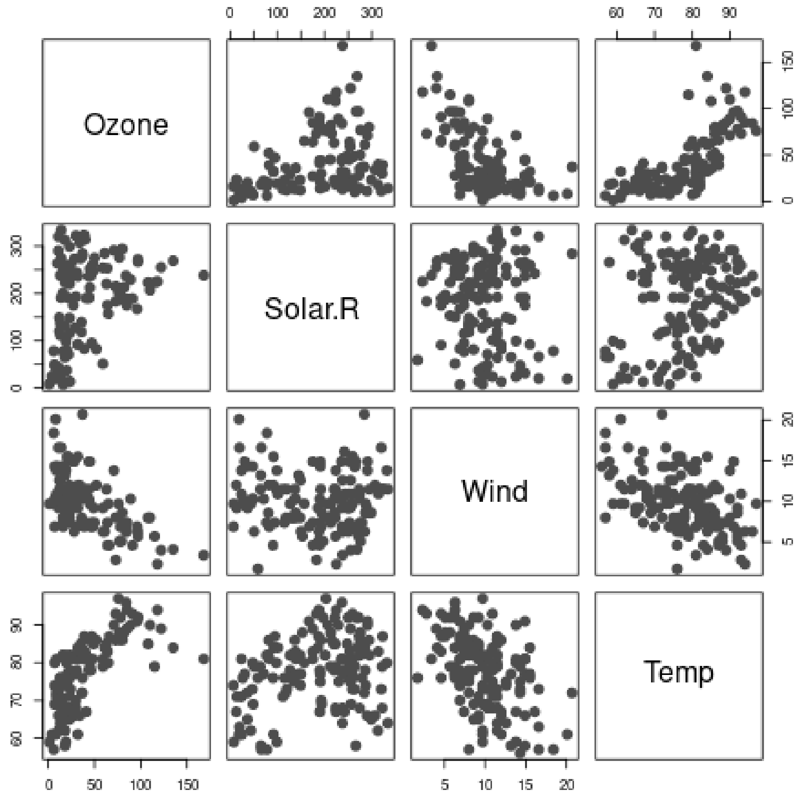
over $\mu \in \mathbb{R}$ and functions g_j with $E[g_j(X_j)] = 0$ and $E[g_j^2(X_j)] < \infty$, where $\rho : \mathbb{R} \to \mathbb{R}$ is an even, non-decreasing and non-negative loss function and σ is the residual scale. Different choices of the loss function ρ yield fits with varying robustness properties. Typical choices for ρ are Tukey's bisquare family and Huber's loss (Maronna et al. (2018)). Note that when $\rho(t) = t^2$, this approach reduces to the standard backfitting.

Illustration

The airquality data set contains 153 daily air quality measurements in the New York region between May and September, 1973 (Chambers et al. (1983)). The interest is in modeling the mean Ozone ("O_3") concentration as a function of 3 potential explanatory variables: solar radiance in the frequency band 4000-7700 ("Solar.R"), wind speed ("Wind") and temperature ("Temp"). We focus on the 111 complete entries in the data set.

Since the plot in Figure **??** suggests that the relationship between ozone and the other variables is not linear, we propose using an additive regression model of the form

$$\text{Ozone} = \mu + g_1(\text{Solar.R}) + g_2(\text{Wind}) + g_3(\text{Temp}) + \varepsilon \,. \tag{2}$$



To fit the model above we use robust local linear kernel estimates and Tukey's bisquare loss function. These choices can be specified using the arguments degree = 1 and type='Tukey' in the call to the function backf.rob. The model is specified with the standard formula notation in R.
The argument windows is a vector with the bandwidths to be used with each kernel smoother. To obtain optimal values we used a robust leave-one-out cross validation approach (Boente et al. (2017)) and obtained the following estimated optimal bandwidths:

```
R> bandw <- c(136.7285, 10.67314, 4.764985)
```

The code below computes the corresponding robust backfitting estimator for Equation 2:

```
R> data(airquality)
R> library(RBF)
R> ccs <- complete.cases(airquality)
R> fit.full <- backf.rob(Ozone ~ Solar.R + Wind + Temp, windows=bandw,
                     degree=1, type='Tukey', subset = ccs, data=airquality)
```

To compare the robust and classical estimates we use the R package gam to fit the model Equation 2 using the standard backfitting algorithm (optimal bandwidths were estimated using leave-one-out cross-validation):

```
R> library(gam)
R> aircomplete <- airquality[ccs, c('Ozone', 'Solar.R', 'Wind', 'Temp')]
R> fit.gam <- gam(Ozone ~ lo(Solar.R, span=.7) + lo(Wind, span=.7) +
                  lo(Temp, span=.5), data=aircomplete)
```

Figure Figure 1 contains partial residuals plots and both sets of estimated functions: blue solid lines for the robust fit and magenta dashed ones for the classical approach.

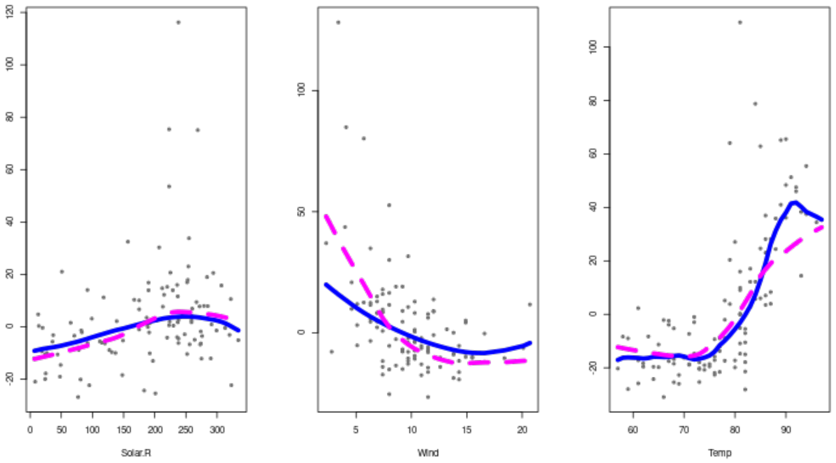


Figure 1: Plots of partial residuals with the robust backfitting fit, the estimated curves with the classical (in magenta) and robust (in blue) procedures.

The two fits differ mainly on the estimated effects of wind speed and temperature. The classical estimate for $g_1(\text{Temp})$ is consistently lower than the robust counterpart for Temp ≥ 85. For wind speed, the non-robust estimate $\hat{g}_2(\text{Wind})$ suggests a higher effect over Ozone concentrations for low wind speeds than the one given by the robust estimate, and the opposite difference for higher speeds.

Residuals from a robust fit can generally be used as a diagnostic tool to detect the presence of atypical observations in the training data. Figure Figure 2 displays a boxplot of these residuals. We note 4 possible outlying points (indicated with red circles).

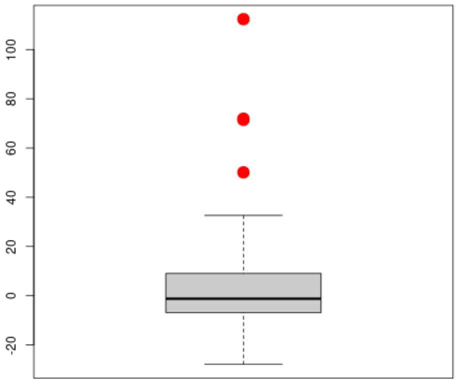


Figure 2: Boxplot of the residuals obtained using the robust fit.

To investigate whether the differences between the robust and non-robust estimators are due to the outliers, we repeated the classical analysis after removing them. Figure Figure 3 shows the estimated curves obtained with the classical estimator using the "clean" data together with the robust ones (computed on the whole data set). Outliers are highlighted in red. Note that both fits are now very close. An intuitive interpretation is that the robust fit has automatically down-weighted potential outliers and produced estimates very similar to the classical ones applied to the “clean’ ’ observations.

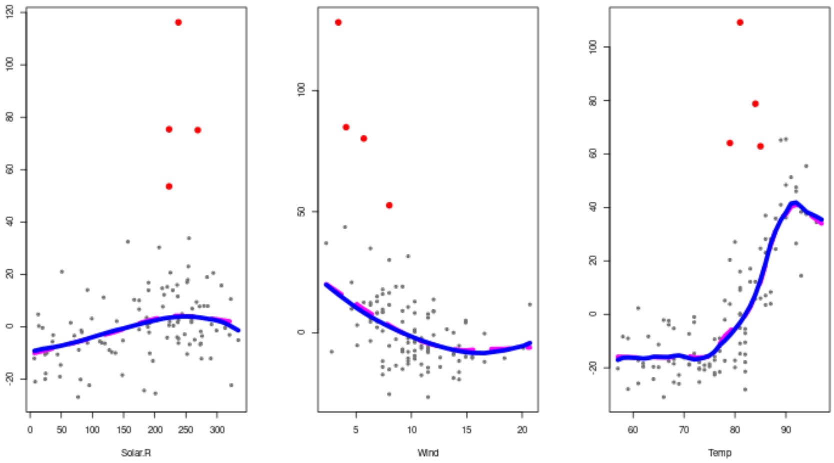


Figure 3: Plots of estimated curves and partial residuals with the robust backfitting fit. In magenta, the estimated curves with the classical backfitting procedure without potential outliers, and in blue the estimated curves with the robust approach. Red points correspond to the potential outliers.

Availability

The software is available at the Comprehensive R Archive Network CRAN and also at the GitHub repository. The GitHub repository also contains detailed scripts reproducing the data analysis above.

Acknowledgements

This research was partially supported by: 20020170100022BA from the Universidad de Buenos Aires; project PICT 2018-00740 from ANPCYT; Internal Projects CD-CBLUJ 301/19 and CD-CBLUJ 204/19 from the Department of Basic Science of the Universidad Nacional de Luj’an (UNLu); the Researchers in Training Project RESREC-LUJ 224/19 (UNLu); and by the Natural Sciences and Engineering Research Council of Canada (Discovery Grant RGPIN-2016-04288).

References

Boente, G., & Fraiman, R. (1989). Robust nonparametric regression estimation. *Journal of Multivariate Analysis*, *29*(2), 180–198.

Boente, G., Martínez, A., & Salibian-Barrera, M. (2017). Robust estimators for additive models using backfitting. *Journal of Nonparametric Statistics*, *29*(4), 744–767. https://doi.org/10.1080/10485252.2017.1369077

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis* (2nd ed.). Chapman & Hall.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, *76*(376), 817–823.

Hastie, T. (2019). *Gam: Generalized additive models*. https://CRAN.R-project.org/package=gam

Hastie, T. J., & Tibshirani, R. J. (Eds.). (1990). *Generalized additive models*. Chapman & Hall.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2018). *Robust statistics: Theory and methods (with r)* (2nd ed.). John Wiley & Sons.

Salibian-Barrera, M., & Martínez, A. (2020). *RBF: Robust backfitting*. https://CRAN.R-project.org/package=RBF