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

17 November 2020

# Summary

Additive models provide an alternative to fully non-parametric regression models. They are flexible and are not affected by the curse of dimensionality. They also allow to explore the individual effect of each covariate on the overall mean function, and thus provide similar interpretations to those obtained with linear models. Standard algorithms to fit additive models are known to be highly susceptible to the presence of a few atypical or outlying observations in the data.

RBF (Salibian-Barrera and Martínez 2020) is an R package that implements a robust regression estimator for additive models using an algorithm called *backfitting*.

# Statement of Need

The purpose of RBF is to:

* Provide a kernel-based estimation procedure for additive models.
* Allow an estimation using a robust procedure that avoids problems that come from atypical observations in the data.

# Implementation Goals

RBF has an interface for the user similar to the most widely used R packages to fit additive models: gam (Hastie 2019), mgcv (Wood 2019), gamlss (Stasinopoulos and Rigby 2019) and VGAM (Yee 2020), among others.

# Background

Hastie and Tibshirani (1990) introduce additive models as a non-parametric generalization of linear models. These models are flexible and interpretable and avoid the curse of dimensionality which is related to the fact that, as dimension increases, neiborhoods of a point of covariates become more sparse. Assuming that we have , , independent and identically distributed random vectors with the same distribution as , additive models postulate that

where the error is independent of and centered at zero. The objects to be estimated are the location parameter and the smooth functions . Note that when for some , the above model reduces to the usual linear regression one. Similarly, the functions can be interpreted as the marginal effect of the -th covariate on the expected value of the response variable when all other explanatory variables remain fixed.

One of the most popular estimation procedures for additive models is the backfitting algorithm proposed by Friedman and Stuetzle (1981). Noting that under model the additive components satisfy , the backfitting procedure iteratively computes estimates of each by smoothing the partial residuals as functions of the observed values of .

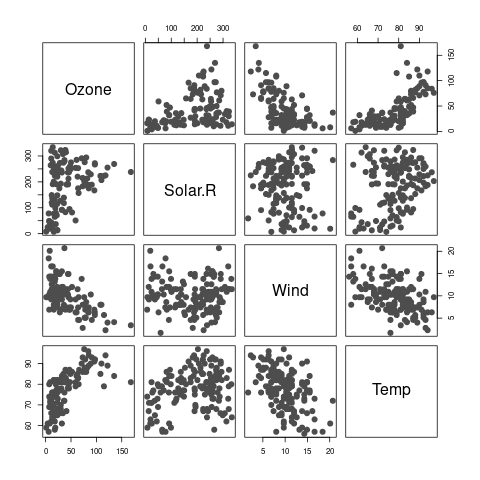
It is well known that these estimators can be seriously affected by a relatively small proportion of atypical observations. Recently, Boente, Martínez, and Salibian-Barrera (2017) proposed a robust version of backfitting, implemented in the RBF package. The intuitive idea consists of using the backfitting algorithm with robust univariate smoothers, such as the kernel-based estimators in Boente and Fraiman (1989). This approach corresponds to finding the solution to the following optimization problem:

over and functions with and , where is an even, non-decreasing and non-negative loss function and is the residual scale. Typical choices for the loss function are Tukey’s bisquare family and Huber’s loss Maronna et al. (2018). Note that when , this approach reduces to the classical backfitting.

# Illustration

We consider the airquality data set available in R which contains 153 daily air quality measurements in the New York region between May and September, 1973 (see Chambers et al. (1983)). The interest is in explaining the mean Ozone concentration ( , measured in ppb) as a function of 3 potential explanatory variables: solar radiance measured in the frequency band 4000-7700 (Solar.R, in Langleys), wind speed (Wind, in mph) and temperature (Temp, in degrees Fahrenheit). In our analysis we focus on the 111 complete cases in the data set.

Figure shows the scatter plot of the data which indicates that the relationship between ozone and other variables does not appear to be linear.



Scatter plot of variables of the Air Quality data set.

We propose to fit an additive model of the form We will use robust smoothers with local linear kernel estimates and Tukey’s bisquare loss function. These choices can be set in the call to the function backf.rob using the arguments degree = 1 and type='Tukey'. The code below computes the robust backfitting estimator for the airquality data, restricting the analysis to cases that do not contain missing entries using the argument subset:

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)

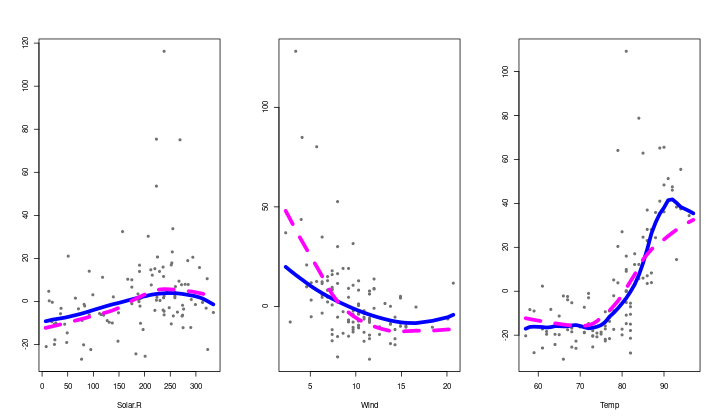
The argument windows is a vector of bandwidths (in our case ) to be used with the kernel smoothers of each explanatory variable. We used leave-one-out combined with the robust cross-validation for selecting the bandwidths as it is described in Boente, Martínez, and Salibian-Barrera (2017). We obtained the following triplet:

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

As it is often informative to compare the robust and classical fits, we use the R package gam that implements the classical backfitting algorithm with local regression smoothers. We ran a similar leave-one-out cross-validation experiment to select its smoothing paramaters.

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)

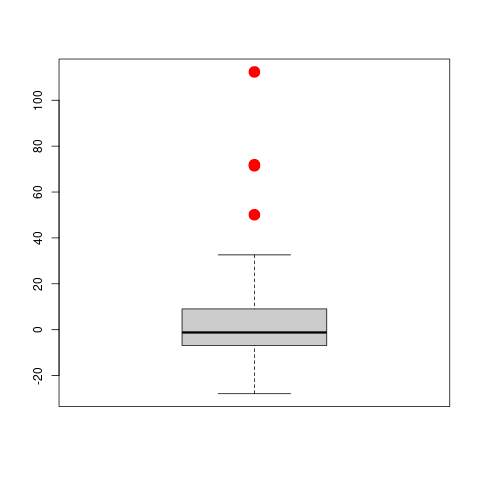
Figure contains the partial residuals plots and both sets of estimated functions: robust in blue and solid lines and non-robust in magenta and dashed lines.



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

The main differences between the two fits are in the estimated effects of wind speed and temperature. In particular, the classical estimate for yields a consistently lower effect on mean Ozone than the robust counterpart for moderate-to-high temperatures (85 degrees and higher). In the case of wind speed, the non-robust estimate indicates a higher effect of wind speed over Ozone concentrations for low speeds than the one given by the robust estimate, and the opposite difference for higher speeds.

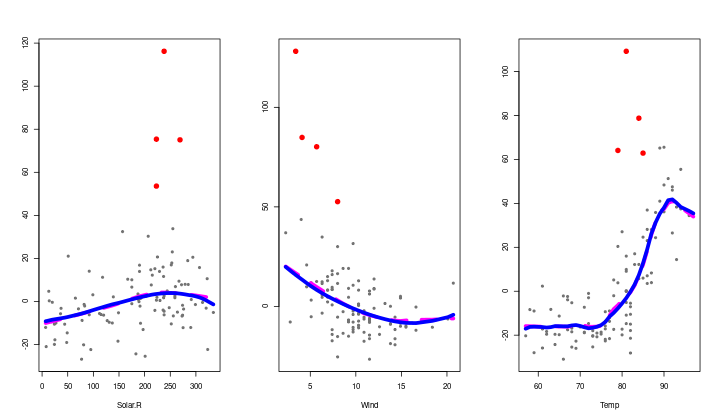
To detect potentially atypical observations in the data, we construct a boxplot of the residuals obtained by the robust fit, and highlight in red those residuals that are unusually large.



Boxplot of the residuals obtained using the robust fit.

The boxplot in Figure shows 4 observations detected as outliers highlighted in red, corresponding to observations 23, 34, 53 and 77.

To verify that the main differences between the robust and non-robust backfitting estimators are due to the possible outliers, we repeated the classical analysis without them. Figure shows the estimated curves obtained with the classical estimator using the clean, data together with the robust ones computed on the original data set and the partial residuals of the potential outliers highlighted in red. Note that both fits are now very close. An intuitive interpretation is that the robust fit automatically down-weighted potential outliers and produced estimates very similar to those obtained with the classical backfitting algorithm applied on the rest of the data.



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](https://CRAN.R-project.org/) and also at the [GitHub repository](https://github.com/msalibian/RBF).

# Acknowledgements

This research was partially supported by 20020170100022BA from the Universidad de Buenos Aires, PICT 2018-00740 from ANPCYT, Internal Projects from the Department of Basic Science CD-CBLUJ 301/19 and CD-CBLUJ 204/19 and Researchers in Training Project RESREC-LUJ 224/19, Universidad Nacional de Luján, Argentina.

# References

Boente, Graciela, and Ricardo Fraiman. 1989. “Robust Nonparametric Regression Estimation.” *Journal of Multivariate Analysis* 29 (2): 180–98.

Boente, Graciela, Alejandra Martínez, and Matias Salibian-Barrera. 2017. “Robust Estimators for Additive Models Using Backfitting.” *Journal of Nonparametric Statistics* 29 (4): 744–67. <https://doi.org/10.1080/10485252.2017.1369077>.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. *Graphical Methods for Data Analysis*. 2nd ed. London: Chapman & Hall.

Friedman, J. H., and W. Stuetzle. 1981. “Projection Pursuit Regression.” *Journal of the American Statistical Association* 76 (376): 817–23.

Hastie, T. J., and R. J. Tibshirani, eds. 1990. *Generalized Additive Models*. London: Chapman & Hall.

Hastie, Trevor. 2019. *Gam: Generalized Additive Models*. <https://CRAN.R-project.org/package=gam>.

Maronna, Ricardo A., R. Douglas Martin, Victor J. Yohai, and Matı́as Salibián-Barrera. 2018. *Robust Statistics: Theory and Methods (with R)*. 2nd ed. John Wiley & Sons.

Salibian-Barrera, Matias, and Alejandra Martínez. 2020. *RBF: Robust Backfitting*. <https://CRAN.R-project.org/package=RBF>.

Stasinopoulos, Mikis, and Bob Rigby. 2019. *Gamlss: Generalised Additive Models for Location Scale and Shape*. <https://CRAN.R-project.org/package=gamlss>.

Wood, Simon. 2019. *Mgcv: Mixed Gam Computation Vehicle with Automatic Smoothness Estimation*. <https://CRAN.R-project.org/package=mgcv>.

Yee, Thomas W. 2020. *VGAM: Vector Generalized Linear and Additive Models*. <https://CRAN.R-project.org/package=VGAM>.