RBF: An R package to compute a robust backfitting estimator for additive models

17 November 2020

Summary

Although highly flexible, non-parametric regression models typically require large sample sizes to be estimated reliably, particularly when they include many explanatory variables. 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. The `RBF` (Salibian-Barrera and Martínez 2020) package for R implements the robust estimator for additive models of Boente, Martínez, and Salibian-Barrera (2017), which can resist the damaging effect of outliers in the training set.

Statement of Need

The purpose of `RBF` is to provide a user-friendly implementation of a robust kernel-based estimation procedure for additive models that is resistant to the presence of potential outliers.

Implementation Goals

`RBF` implements a user interface similar to that of the R package `gam` (Hastie 2019), which computes the standard non-robust kernel-based fit for additive models using the backfitting algorithm. The `RBF` package also includes several modeling tools, including functions to produce diagnostic plots, obtain fitted values and compute predictions.

Background

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

If Y denots 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 location parameter $\mu \in \mathbb{R}$ and $g_j \, : \, \mathbb{R} \to \mathbb{R}$ are smooth functions. Note that if $g_j(X_j) = \beta_j \, X_j$ for some $\beta_j \in \mathbb{R}$ then Equation 1 reduces to a standard linear regression model.

The backfitting algorithm (Friedman and Stuetzle (1981)) fits model in Equation 1 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, Martínez, and Salibian-Barrera (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 and Fraiman (1989)). These robust estimators solve:

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

over $\mu \in \mathbb{R}$ and functions g_j with $E[g_j(X_j)] = 0$ and $E[g_j^2(X_j)] < \infty$. The loss function $\rho : \mathbb{R} \to \mathbb{R}$ is even, non-decreasing and non-negative, 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 1 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, Martínez, and Salibian-Barrera (2017)) which resulted in 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`. Optimal bandwidths were estimated using leave-one-out cross-validation as before.

```
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 2 contains partial residuals plots and both sets of estimated functions: blue solid lines indicate the robust fit and magenta dashed ones the classical one.

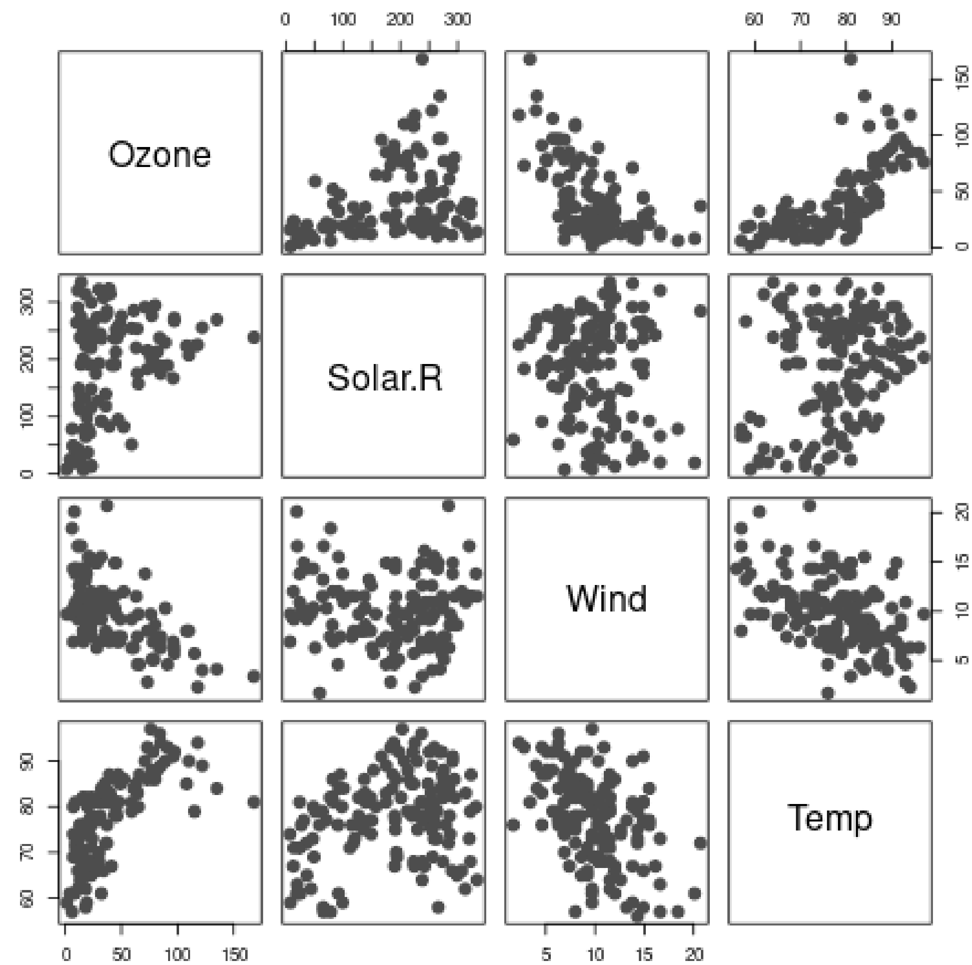

Figure 1: Scatter plot of variables of the Air Quality data set.

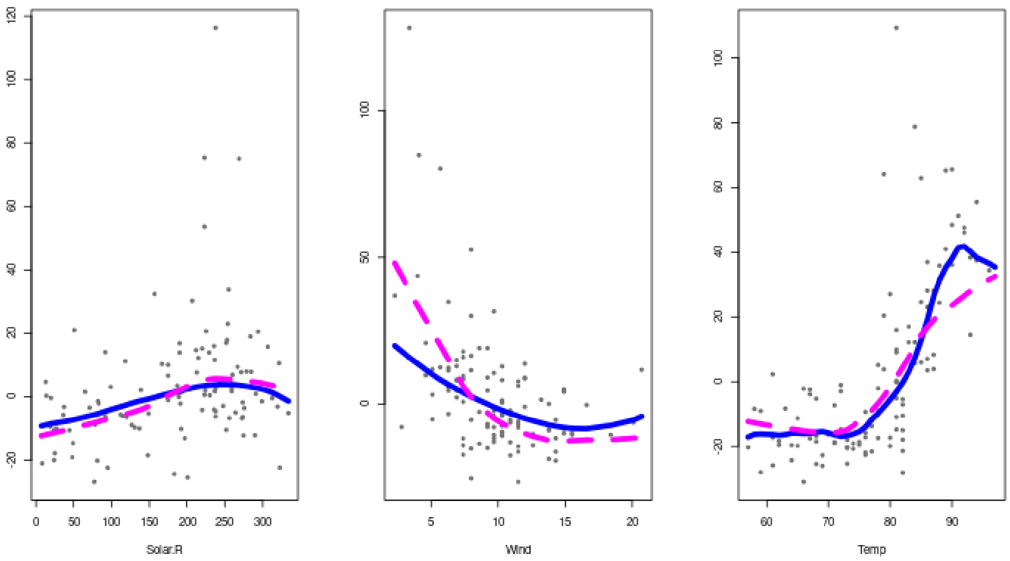

Figure 2: Plots of partial residuals with the robust backfitting fit, the estimated curves with the classical (in magenta) and robust (in blue) procedures.

The two fits differ mainly on the estimated effects of wind speed and temperature. The classical estimate for g_1(Temp) is consistently lower than the robust counterpart for Temp ≥ 85. For wind speed, the non-robust estimate $\hat{g}_2$(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 to detect the presence of atypical observations in the training data. Figure 3 displays a boxplot of these residuals. We note 4 possible outlying points (indicated with red circles).

To investigate whether the differences between the robust and non-robust estimators are due to the outliers, we recomputed the classical fit after removing them. Figure 4 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.

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 Luján (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).

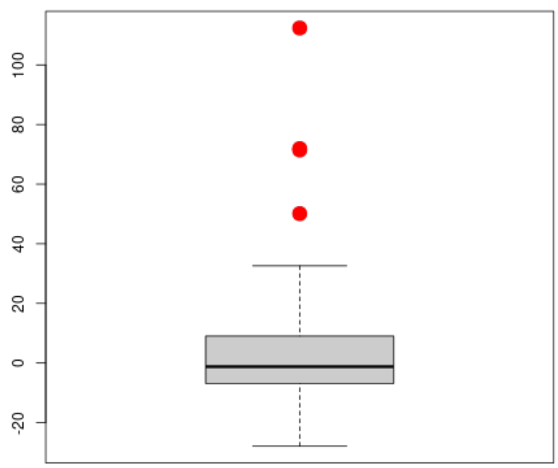

Figure 3: Boxplot of the residuals obtained using the robust fit.

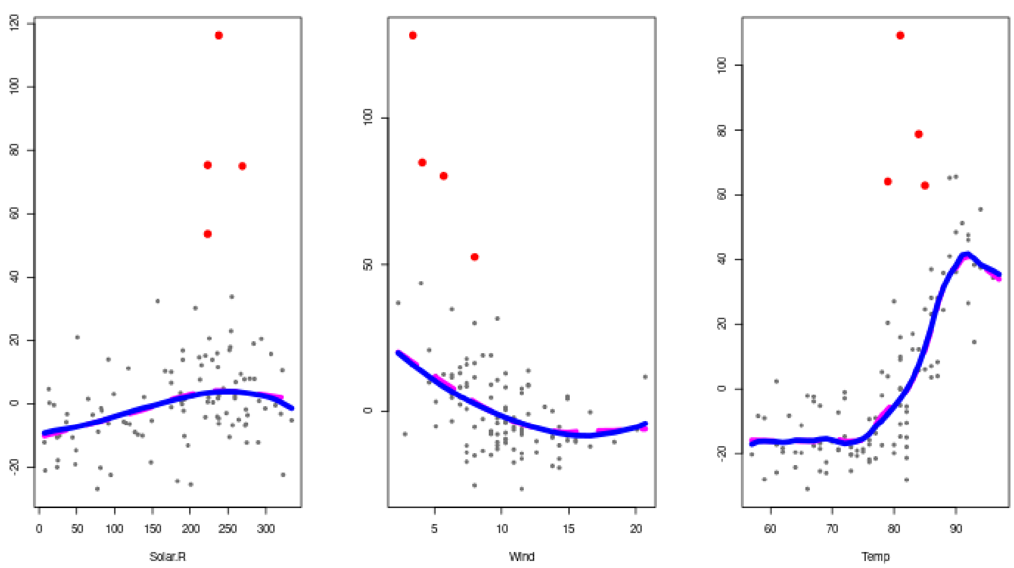

Figure 4: Plots of estimated curves and partial residuals with the robust backfitting fit. In magenta, the estimated curves with the classical backfitting procedure without potential outliers, and in blue the estimated curves with the robust approach. Red points correspond to the potential outliers.

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 Matías 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.