A robust fit for Generalized Additive Models

Matías Salibián-Barrera [1]

[1] Department of Statistics, University of British Columbia, 333 - 6356 Agricultural Road, Vancouver, BC, V6T 1Z2, Canada

Keywords: Robustness, Generalized Additive Models

1 Introduction

Generalized Additive Models (GAM) (Hastie and Tibshirani, 1986, 1990) are a powerful exploratory tool that is widely used in practice. Unfortunately, popular fitting algorithms for these models (e.g. the General Local Scoring Algorithm (GLSA), Hastie and Tibshirani, 1990) can be highly sensitive to a small proportion of observations that depart from the model. In particular, a few atypical observations could seriously affect the non-parametric estimates of the smooth regression functions.

Figure 1 illustrates a possible scenario. The data consist of 80 observations y_i, $1 \leq i \leq 80$, from Poisson distributions with $y_i \sim \mathcal{P}(\lambda_i)$, $\log(\lambda_i) = \sin(2\,x_i/120) + \cos(7\,x_i/60) + 1$, and $x_i = i$ for $1 \leq i \leq 80$. Five outliers were placed around x_{23}. Note that the classical GLSA fit is heavily affected by the outliers.

Robust proposals for non-parametric regression models have been studied recently by Cantoni and Ronchetti (2001a) and Bianco and Boente (2002). In this paper we propose a new robust fit for GAM models. The building blocks of this proposal are robust estimates for Quasi-Likelihood (QL) models (Cantoni and Ronchetti, 2001b; see also Stefanski, Carroll and Ruppert, 1986; and Künsch, Stefanski and Carroll, 1989) and the GLSA algorithm (Hastie and Tibshirani, 1986, 1990). Specifically, we adapt the GLSA algorithm using robust estimating equations to determine appropriate weights that transform the robust QL score equations into re-weighted least squares equations. We then iteratively fit weighted additive models, in the same spirit as GLSA. Bandwidth selection can be done automatically using a robust cross-validation criteria (Ronchetti and Staudte, 1994).

Figure 2 compares the classical GLSA fit with the one obtained with the robust approach. Note that the robust fit is able to stay close to the true mean function. The bandwidths were chosen using a robust cross-validation criteria. Simulation results suggest that the fit obtained with this algorithm is able to resist the effect of outliers in a number of different situations and that it also performs well when the data follow the model.

References

Bianco, A. and Boente, G. (2002). Robust nonparametric generalized regression estimation. Unpublished manuscript.

Cantoni, E. and Ronchetti, E. (2001a). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing*, 11, 141-146.

Cantoni, E. and Ronchetti, E. (2001b). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96, 1022-1030.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1, 297-318.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*, Chapman & Hall : New York.

Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, 84, 460-466.

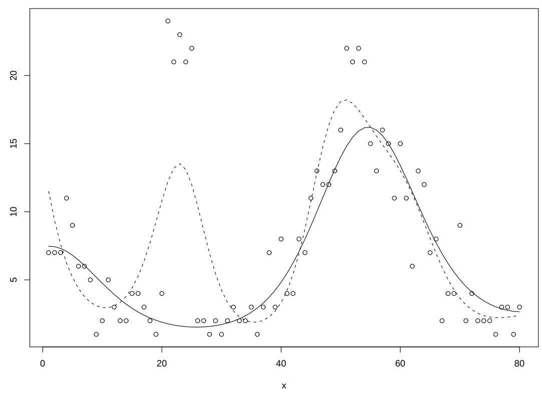

FIGURE 1. Example – Solid line is the true mean function – Dashed line is the fit obtained with the General Local Scoring Algorithm. Note the dramatic effect of the outliers in the resulting fit.

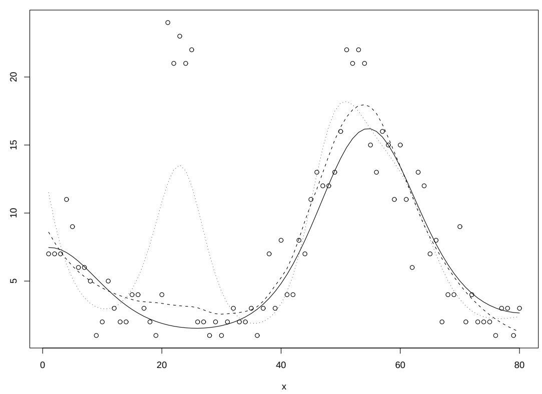

FIGURE 2. Example – Solid line is the true mean function – Dotted line is the fit obtained with the General Local Scoring Algorithm – Dashed line is the robust fit with automatic bandwidth selection.

Ronchetti, E. and Staudte, R. (1994). A robust version of Mallow's C_p. *Journal of the American Statistical Association*, 89, 550, 559.

Stefanski, L. A., Carroll, R. J., and Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika*, 73, 413-425.