An outlier-robust fit for Generalized Additive Models with applications to disease outbreak detection

Azadeh Alimadad [*]   Matias Salibian-Barrera [†]

February 14, 2011

Keywords: Generalized Additive Models, robustness, outliers, robust quasi-likelihood

Abstract

We are interested in a class of unsupervised methods to detect possible disease outbreaks, i.e. rapid increases in the number of cases of a particular disease that deviate from the pattern observed in the past. The motivating application for this paper deals with detecting outbreaks using Generalized Additive Models to model weekly counts of certain infectious diseases. We can use the distance between the predicted and observed counts for a specific week to determine whether an important departure has occurred. Unfortunately, this approach may not

[*] Azadeh Alimadad is a PhD candidate at the Faculty of Health Sciences, Simon Fraser University, and the Modeling of Complex Social Systems Program (MoCSSy), Simon Fraser University, Burnaby, BC, Canada. Azadeh Alimadad gratefully acknowledges technical support from the IRMACS Centre, SFU, and partial financial support from the SFU CTEF MoCSSy Program.

[†] Matias Salibian-Barrera is an Associate Professor at the Department of Statistics, University of British Columbia, Vancouver, BC, Canada. This work was partially supported by a Discovery Grant of the Natural Sciences and Engineering Research Council of Canada.

work as desired because GAMs can be very sensitive to the presence of a small proportion of observations that deviate from the assumed model. Thus, the outbreak may affect the predicted values causing these to be close to the atypical counts, and thus mask the outliers by having them appear not to be too extreme or atypical. We illustrate this phenomenon with influenza-like-illness doctor visits data from the US for the 2006-2008 flu seasons. One way to avoid this masking problem is to derive an algorithm to fit GAM models that can resist the effect of a small number of atypical observations. In this paper we discuss such an outlier-robust fit for Generalized Additive Models based on the backfitting algorithm. The basic idea is to replace the maximum likelihood based weights used in the Generalized Local Scoring Algorithm with those derived from robust quasi-likelihood equations (Cantoni and Ronchetti, 2001b). These robust estimators for generalized linear models work well for the Poisson family of distributions, and also for Binomial distributions with relatively large numbers of trials. We show that the resulting estimated mean function is resistant to the presence of outliers in the response variable and that it also remains close to the usual GAM estimator when the data do not contain atypical observations. We illustrate the use of this approach on the detection of the recent outbreak of H1N1 flu by looking at the weekly counts of influenza-like-illness (ILI) doctor visits, as reported through the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet), and also apply our method to the numbers of requested isolates in Canada. Weeks with a sudden increase in ILI visits or requested isolates are much more clearly identified as atypical by the robust fit because the observed counts are far from the ones predicted by the fitted GAM model.

1 Introduction

Generalized Additive Models (GAM) (Hastie and Tibshirani, 1986; Wood, 2006) are flexible extensions to Additive Models (Friedman and Stuetzle, 1981; Huber, 1985) to the case of non-normally distributed response variables. They extend Additive Models in the same spirit as Generalized Linear Models (Nelder and Wedderburn, 1972) extend linear models. In particular, they allow the modeling of a properly transformed mean response as a sum of smooth functions of individual covariates. More specifically, let Y be a random variable with distribution function in an exponential family, let $\mathbf{X} = (X_1, \ldots, X_p)' \in \mathbb{R}^p$ be a vector of covariates and assume that for an appropriate link function g we have

$$g\left(E\left[Y|\,\mathbf{X}\right]\right) = f_0 + \sum_{j=1}^{p} f_j(X_j)\,, \tag{1}$$

where $f_0 \in \mathbb{R}$, and $f_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \ldots, p$ are "smooth" functions. Given a sample $(Y_1, \mathbf{X}_1)$, $\ldots$, $(Y_n, \mathbf{X}_n)$ following model (1), we are interested in estimating $E\left[Y|\mathbf{X}\right]$. Provided estimates $\hat{f}_j$, $j = 0$, $\ldots$, p are available, a natural estimator for $E[Y|\mathbf{X}]$ is

$$g^{-1}\left(\hat{f}_0 + \sum_{j=1}^{p} \hat{f}_j(X_j)\right) .$$

Hastie and Tibshirani (1986) proposed the Generalized Local Scoring Algorithm (GLSA) to calculate estimated smooth functions $\hat{f}_j$, $j = 0$, $\ldots$, p. This algorithm extends the backfitting algorithm used to fit Additive Models to the case of non-Gaussian responses in the same spirit as the iterative weighted least squares algorithm used to fit generalized linear models extends least squares using likelihood-based weights. Wood (2006) considers a penalized likelihood approach where the smooth components are modeled using splines or other appropriate function basis.

This paper is motivated by a problem arising from the application of these models to weekly counts of infectious diseases. More specifically, we are interested in detecting outbreaks, i.e. sudden increases in the number of reported cases of a particular disease, or other departures from the pattern of past observed counts. Although in many applications, a careful exploratory analysis based on scatter plots can provide with an adequate answer as to whether there has been a change in the behaviour of the counts of interest, surveillance systems following a large number of diseases and / or health districts may require an automatic method to flag observations that potentially depart from historical patterns.

One way to identify atypical observations is to fit a reasonable model to the data, and use the distance between the predicted and observed responses to determine whether an important departure has occurred. Weekly counts of relatively prevalent diseases or viral infections (e.g. influenza, HIV, Hepatitis C) typically exhibit strong (and non-linear) temporal and seasonal patterns. Generalized Additive Models are a natural tool to model these data. Unfortunately, this approach may not work as desired because it is easy to see that GAMs can be very sensitive to the presence of a small proportion of observations that deviate from the assumed model. In other words: a few atypical observations could seriously affect the non-parametric estimates of the smooth regression function (see, for example, Figure 1). Thus, the outbreak may affect the predicted values causing them to be closer to the atypical counts, and thus masking them. Additionally, the effect of the outbreak may cause other ("good") observations to (falsely) appear as deviating significantly from the model.

We illustrate these problems with data on patient visits reported through the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). These data are available on-line at `http://www.cdc.gov/flu/weekly/fluactivity.htm`. We use the weekly counts of ILI visits for the 2006, 2007 and 2008 seasons. A season consists of weeks 40 to week 20 of the following year, so that, for example, the 2008 season includes up to week 20 of 2009, where the H1N1 flu epidemic had already caused a noticeable increase in ILI visits. Figure 2 shows the data, the GAM fits obtained with the GLSA and the penalized splines approach of Wood (2006) as implemented in libraries `gam` and `mgcv` in `R` (R Development Core Team, 2009), respectively, and the corresponding standardized residuals (right plot on panel (b)). The observations for the 2008-2009 season are shown with solid circles. The bandwidth for the backfitting estimate was chosen using leave-one-out cross validation. Note how the standardized residuals for the last weeks of the 2008-2009 outbreak are only slightly higher than those observed in past seasons, indicating a potentially bad season, but not much worse than the others in this data set. Also, note how the standardized residuals are reasonably scattered around zero before week 16 but clearly shift downwards later to compensate for the high number of visits on Weeks 16 to 20 of 2009.

Other potential problems caused by a small proportion of observations deviating from the assumed model are illustrated with three synthetic examples displayed in Figure 1.

[Figure 1 about here.]

We used both the backfitting algorithm of Hastie and Tibshirani (1986) and the penalized splines approach of Wood (2006), as implemented in the `R` packages `gam` and `mgcv`, respectively. The bandwidth for the backfitting estimate was chosen using leave-one-out cross-validation. In plot (a) we have a few outliers at the end of the curve. This is a particularly bad configuration and we see that both fitted mean functions completely accomodate them. A similar situation is depicted in plot (b) where the outliers are now well within the range of the covariate, but their effect is no less dramatic. Plot (c) illustrates the dangers that a few relatively scattered extreme outliers can pose, in this case affecting the cross-validation criteria used to select either the bandwidth of the smoother (for the `gam` fit) or the penalty term (for the `mgcv` fit).

One way to avoid these problems is to derive an algorithm to fit GAM models that is not seriously affected by a small number of atypical observations. We will call such a fit a "robust fit". A commonly used measure for the degree of outlier protection provided by an estimator is its breakdown point (Donoho and Huber, 1983). For parametric models the breakdown point is the smallest proportion of contaminated data that can take the estimate beyond any finite bound or make it otherwise non-informative (see Davies and Gather, 2005). In this paper we will not address the delicate problem of formally defining breakdown for non parametric regression, but rather, present a new fitting algorithm that yields estimators that are consistently closer to the true mean function than the standard gam fit under different model-violations, while still performing very well when there are no outliers in the data.

An outlier-resistant fit for generalized additive models can be obtained by robustifying the penalized splines approach (see, for example, Wood (2006)). Recently Croux, Gijbels and Prosdocimi (2010) proposed a robust fit based on this approach. In this paper we will focus on the Generalized Local Scoring Algorithm (GLSA) (Hastie and Tibshirani, 1986). This approach consists of applying the back-fitting algorithm with weights derived from the iterative weighted least squares (IWLS) algorithm used to solve the maximum likelihood equations of the corresponding generalized linear model. The basic idea of our robust fit is to replace these maximum likelihood based weights with others derived from robust quasi-likelihood equations. Hence, one of the building blocks of our proposal are the robust estimates for Quasi-likelihood (QL) models (Cantoni and Ronchetti, 2001b). These work well for the Poisson family of distributions, and also for Binomial distributions with a moderately large number of trials. Furthermore, note that outliers can also affect data-based methods used to determine the tuning (or penalty) constants involved in the smoothing steps of the GLSA algorithm (see, for example, Cantoni and Ronchetti, 2001a). Intuitively, one does not want to penalize fits that do not predict well those observations that are potential outliers. Hence, we also propose a robustified leave-one-out cross-validation criterion that downweights outlying observations. Numerical studies indicate that this algorithm works well in practice, in particular: it is resistant to extreme observations (even at the boundary of the data range) and it behaves similarly to the GLSA when the data do not contain outliers.

The rest of this paper is organized as follows. Section 2 introduces the model and briefly reviews the generalized local scoring algorithm. Section 3 dicusses our robust version of this algorithm used to compute the robust fit and a robust cross-validation criteria for bandwidth selection. Two examples are discussed in Section 4 while Section 5 reports the results of a simulation study and concluding remarks are found in Section 6. Some technical details have been relegated to the Appendix.

2 The Generalized Local Scoring Algorithm

Regression models provide a framework to study the mean of a response variable Y as a function f (generally unknown to some degree) of one or more covariates $\mathbf{X} \in \mathbb{R}^p$ via the relationship $E(Y|\mathbf{X}) = f(\mathbf{X})$. Generalized linear models assume that there exist a function g (the link function) such that $g(E(Y|\mathbf{X})) = \boldsymbol{\beta}'\mathbf{X}$ for a certain vector of regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$. Let $Y_1, \ldots, Y_n$ be n independent random variables following this model with associated covariates $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Let $\mu_i = E(Y_i|\mathbf{X}_i) = g^{-1}(\eta_i)$, with $\eta_i = \boldsymbol{\beta}'\mathbf{X}_i$ and $v_i = V(Y_i|\mathbf{X}_i) = v(\mu_i)$, where v is a known function. Let $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ be the observed realizations of the above random variables. When the distribution of $Y_i|\mathbf{X}_i$, $i = 1, \ldots, n$, belong to an exponential family, the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\beta}}_n$ satisfies the following equations

$$\sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{v_i} \right) \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_n} = \mathbf{0} \,.$$

It is not difficult to see that these equations can also be written as

$$\sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} W_i \, \mathbf{x}_i = \mathbf{0} \,, \tag{2}$$

(see McCullagh and Nelder, 1999), where the weights are given by

$$W_i = v_i^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \,. \tag{3}$$

These equations can also be justified only assuming that the model for the first and second moments of the response variable are appropriate. This approach is known as quasi-likelihood (see Wedderburn, 1974; and McCullagh, 1983). The Fisher's scoring algorithm is an iterative procedure that can be used to solve equation (2) above. It is based on the corresponding

Newton-Raphson iterations but it replaces the entries in the matrix of partial derivatives by their expected values. Each iteration of Fisher's scoring can be expressed as the solution of the following iterative weighted least squares scheme (see Nelder and Wedderburn, 1972):

$$\left(\sum_{i=1}^{n} W_i(\boldsymbol{\beta}^{(j)}) \, \mathbf{x}_i \, \mathbf{x}_i^t \right) \, \boldsymbol{\beta}^{(j+1)} = \sum_{i=1}^{n} W_i(\boldsymbol{\beta}^{(j)}) \, \mathbf{x}_i \, z_i(\boldsymbol{\beta}^{(j)}) \, , \qquad j = 0, 1, \ldots , \tag{4}$$

where

$$z_i(\boldsymbol{\beta}^{(j)}) = \eta_i(\boldsymbol{\beta}^{(j)}) \, + \, \left(y_i - \mu_i(\boldsymbol{\beta}^{(j)}) \right) \, \frac{\partial \eta_i}{\partial \mu_i}(\boldsymbol{\beta}^{(j)}) \, ,$$

and we have made explicit the dependence of η_i, μ_i and W_i on the regression parameters $\boldsymbol{\beta}$.

In many applications, however, it is of interest to relax this model to include potentially non-linear covariate effects even after applying the link function transformation g. Generalized additive models extend these models to other exponential distributions in the same spirit as generalized linear models extend linear models. More specificially, we will assume that for the link function g we have

$$g\left(E\left(Y | \, X_1, X_2, \ldots, X_p \right) \right) \, = \, f_0 + \sum_{j=1}^{p} f_j \left(X_j \right) \, ,$$

where $f_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \ldots, p$, denote unspecified but smooth univariate functions with $E(f_j(X)) = 0$ and f_0 is a constant. Estimation of these functions can be done by replacing the weighted least squares representation in (4) by an appropriate algorithm to fit a weighted additive model. Specifically, the Generalized Local Scoring Algorithm (Hastie and Tibshirani, 1986) can be described in the following steps:

(a) Let $m = 0$, $f_0^m = g(\bar{y})$, $f_k^m = 0$ for $k = 1, \ldots, p$, where $\bar{y} = \sum_{i=1}^{n} y_i / n$.

(b) Let $z_i^m = \eta_i^m + (y_i - \mu_i^m)(\frac{\partial \eta_i}{\partial \mu_i})$ for $i = 1, \ldots, n$, with $\eta_i^m = f_0^m + \sum_{j=1}^{p} f_j^m(X_{ij})$, $\mu_i^m = g^{-1}(\eta_i^m)$ and weights $W_i^m = (\frac{\partial \mu_i}{\partial \eta_i})^2 (V_i^m)^{-1}$.

(c) Fit a weighted additive model to the z_i^m's and obtain estimated functions f_j^{m+1}, $j = 1, \ldots, p$, additive predictors η_i^{m+1}, and fitted values μ_i^{m+1}, $i = 1, \ldots, n$.

(d) Compute a convergence criterion, e.g.

$$\Delta(\eta^{m+1}, \eta^m) = \frac{\sum_{j=1}^{p} \|f_j^{m+1} - f_j^m\|}{\sum_{j=1}^{p} \|f_j^m\|} \,,$$

where $\|f_j^m\|$ is the Euclidean norm (in $\mathbb{R}^n$) of the vector of n evaluations of f_j^m.

(e) Let $m = m + 1$ and repeat steps (ii) to (iv) until $\Delta(\eta^{m+1}, \eta^m)$ is below some predetermined small threshold.

Note that step (c) in the above algorithm involves fitting a weighted additive model. Following Hastie and Tibshirani (1986), we use the back-fitting algorithm with a weighted smoother. This scheme can be applied with any univariate scatterplot smoother. For ease of computation we use the locally weighted scatterplot smoothing (LOESS) of Cleveland (1979), as implemented in the function `gam` of the package `gam` for `R`. This smoother performs local polynomial regression (we used local linear fits in our applications), and the weights determine the importance of each observation in the local neighbourhoods, central points receiving larger weights.

Hastie and Tibshirani (1987) suggest a different convergence criterion for step (d), namely one based on the deviances rather than the additive predictors. However, in our experience, both criteria yield very similar results, which can be understood given the continuity of both the link and deviance functions. Moreover, in most applications we have seen the qualitative conclusions derived from a generalized additive fit (robust or not) remained the same when either convergence criteria was used. Nonetheless, one can imagine situations with a very flat likelihood surface where the GLSA algorithm might converge to rather different solutions depending on the convergence criterion used. In such cases we prefer to use the more stringent criterion we described above.

3 An outlier-robust variant of the General Local Scoring Algorithm

Most likelihood based methods are highly sensitive to slight departures of the assumptions utilized to derive them. A small proportion of the data not following the assumed model may severely affect both the estimates and the final conclusion of the analysis. Moreover, note that many diagnostic methods may fail to identify violations of the assumed model because the estimates on which they are based might have been seriously distorted by the atypical observations they try to find. This phenomenom is sometimes called "masking" in the literature (e.g. Rousseeuw and van Zomeren, 1990). Interestingly, in some applications these atypical observations are the ones researchers are more interested in, and finding them is the main objective of the analysis. Some examples include intruder detection methods, image analysis, and disease outbreak detection. Our main motivating example described in Section 1 falls in the last category: the goal is to develop an algorithm for automatic (and early) detection of sudden and atypical increases in the number of reported cases (or some other related indicators) of a particular disease.

Noting that the Generalized Local Scoring Algorithm relies on the iterative weighted least squares representation in equation (4) of Fisher's scoring iterations for GLM models, we propose a similar algorithm based on iterative weighted least squares equations for robust quasi-likelihood equations as in Cantoni and Ronchetti (2001b). These estimates work well with log-linear models and with logistic models for binomial experiments with relatively large numbers of trials. In the rest of this section we derive our robust version of the Generalized Local Scoring Algorithm.

Following Cantoni and Ronchetti (2001b), consider M-estimators for linear quasi-likelihood models defined as the solution of estimating equations of the form

$$\sum_{i=1}^{n} \left(\phi(y_i, \mu_i) \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \, - \, a_n(\boldsymbol{\beta}) \Bigg|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_n} \; = \, \mathbf{0} \, ,$$

where $a_n(\boldsymbol{\beta}) \; = \; \sum_{i=1}^{n} E\{ \, \phi(y_i, \mu_i)[\partial \mu_i / \partial \boldsymbol{\beta}] \, \}$ is a correction factor to ensure that the above estimating equations are unbiased,

$$\phi(y_i, \mu_i) = \psi_c \left(\frac{y_i - \mu_i}{\sqrt{v_i}} \right) \; \frac{1}{\sqrt{v_i}} \, ,$$

and ψ_c is a member of Huber's family of psi-functions

$$\psi_c(r) = \begin{cases} r & |r| \leq c \\ \text{c sign(r)} & |r| > c \end{cases} \tag{5}$$

Thus, the estimating equations are

$$\sum_{i=1}^{n} \psi_c \left(\frac{y_i - \mu_i}{\sqrt{v_i}} \right) \frac{1}{\sqrt{v_i}} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} - a_n(\boldsymbol{\beta}) \Bigg|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_n} \; = \, \mathbf{0} \, . \tag{6}$$

Note that our choice of estimating equations reflects the assumption that there are no outliers in the covariates (see Cantoni and Ronchetti (2001b); Künsch *et al.* (1989); and Stefanski *et al.* (1986) for a more detailed discussion). The solutions $\hat{\boldsymbol{\beta}}_n$ of the equation above are asymptotically normally distributed.

In the following we will adapt the Generalized Local Scoring Algorithm (GLSA) described above to fit Robust Generalized Linear Models. The first step is to find a representation similar to (4) for equation (6) that can be used to derive robust weights and the corresponding transformed responses. It is not difficult to verify (details can be found in the Appendix) that the Fisher-scoring algorithm to solve (6) can be written as

$$\left(\sum_{i=1}^{n} \omega_i(\boldsymbol{\beta}^{(j)}) \mathbf{x}_i \mathbf{x}_i' \right) \boldsymbol{\beta}^{(j+1)} = \sum_{i=1}^{n} \omega_i(\boldsymbol{\beta}^{(j)}) \mathbf{x}_i \, z_i(\boldsymbol{\beta}^{(j)}) \, , \tag{7}$$

where

$$z_i(\boldsymbol{\beta}^{(j)}) = \eta_i(\boldsymbol{\beta}^{(j)}) + \frac{h_i(\boldsymbol{\beta}^{(j)})}{\ell_i(\boldsymbol{\beta}^{(j)})} \, , \quad \omega_i(\boldsymbol{\beta}^{(j)}) = \ell_i(\boldsymbol{\beta}^{(j)}) \frac{1}{\sqrt{v_i(\boldsymbol{\beta}^{(j)})}} \frac{\partial \mu_i}{\partial \eta_i}(\boldsymbol{\beta}^{(j)}) \, ,$$

$$l_i = \left[\frac{E(\psi_c'(r_i))}{\sqrt{v_i}} \frac{\partial \mu_i}{\partial \eta_i} + \frac{1}{2} E(\psi_c'(r_i) r_i) \frac{1}{v_i} \frac{\partial v_i}{\partial \eta_i} + E(\frac{\partial}{\partial \eta_i} E(\psi_c(r_i))) \right] ,$$

$$h_i = \psi_c(r_i) - E(\psi_c(r_i)) , \quad \text{and} \quad r_i = (y_i - \hat{\mu}_i)/\sqrt{v_i} .$$

The choice of the tuning constant c: When estimating regression coefficients for linear or generalized linear models, the score equations of robust estimators are generally chosen based on considerations of asymptotic efficiency or robustness properties. For example, in linear regression models, one considers maximal outlier resistance and a pre-determined asymptotic efficiency when the errors are normally distributed (Maronna *et al.*, 2006). For generalized linear models Cantoni and Ronchetti (2001b) select the tuning constant c for the function ψ_c considering the stability of tests of hypotheses. In the case of generalized additive models, however, statistical inference is still an active area of research (see, for example, Dominici *et al.*, 2002; Ramsay *et al.*, 2003; and Figueiras *et al.*, 2005). Consequently, in this paper we will consider c simply as a downweighting threshold. Note that observations with a Pearson residual of absolute value larger than c will be downweighted in the estimating equations; the larger the residual, the lower the weight. In all our examples and simulations we set $c = 1.5$, which produced good results in a variety of situations. In our experience, values of c betweeen 1 and 4 produced similar qualitative results. As the value of c gets larger, the robust fit more closely resembles the classical one.

3.1 The Robust Generalized Local Scoring Algorithm

Our Robust Generalized Local Scoring Algorithm builds on Hastie and Tibshirani's GLSA, replacing the MLE-based weights (3) with those derived from the robust quasi-likelihood score equations (6). The algorithm can be described in the following steps:

(a) Let $m = 0$, $f_0^m = g(\bar{y})$, $f_k^m = 0$ for $k = 1, \ldots, p$, where $\bar{y} = \sum_{i=1}^n y_i/n$. Alternatively, set $f_0^m = g(\tilde{y})$, where $\tilde{y} = \text{median}_{1 \leq i \leq n}\, y_i$.

(b) Let $z_i^m = \eta_i^m + h_i^m/\ell_i^m$ for $i = 1, \ldots, n$, with $\eta_i^m = f_0^m + \sum_{j=1}^p f_j^m(X_{ij})$, $\mu_i^m = g^{-1}(\eta_i^m)$ and weights

$$\omega_i^m = \ell_i^m \frac{1}{\sqrt{v_i^m}} \frac{\partial \mu_i}{\partial \eta_i}\,,$$

with

$$l_i^m = \left[\frac{E(\psi'(r_i^m))}{\sqrt{v_i^m}} \frac{\partial \mu_i}{\partial \eta_i} + 1/2 E(\psi'(r_i^m) r_i^m) \frac{1}{v_i^m} \frac{\partial v_i}{\partial \eta_i} + E(\frac{\partial}{\partial \eta_i} E(\psi(r_i^m))) \right]\,,$$

$$h_i^m = \psi_c(r_i^m) - E(\psi_c(r_i^m))\,,$$

and

$$r_i^m = (y_i - \mu_i^m)/\sqrt{v_i^m}\,.$$

(c) Fit a weighted additive model to the z_i^m's and obtain estimated functions f_j^{m+1}, $j = 1, \ldots, p$, additive predictors η_i^{m+1}, and fitted values μ_i^{m+1}, $i = 1, \ldots, n$.

(d) Compute a convergence criterion, e.g.

$$\Delta(\eta^{m+1}, \eta^m) = \frac{\sum_{j=1}^p \|f_j^{m+1} - f_j^m\|}{\sum_{j=1}^p \|f_j^m\|}\,.$$

where $\|f_j^m\|$ is the Euclidean norm (in $\mathbb{R}^n$) of the vector of n evaluations of f_j^m.

(e) Let $m = m + 1$ and repeat steps (ii) to (iv) until $\Delta(\eta^{m+1}, \eta^m)$ is below some predetermined small threshold.

In the Appendix we show how to calculate ℓ_i^m for score functions in Huber's family of functions (5) and responses with Poisson or Binomial distribution. Code for `R` (R Development Core Team, 2009) implementing the above algorithm is available on-line at http:// www.stat.ubc.ca / ˜ matias / soft.html.

3.2 Asymptotic properties

The literature on asymptotic properties of backfitting estimators for either additive or generalized additive models is not yet extensive (see, for example, Opsomer (2000), Opsomer and Kauermann (2000, 2002), Kauermann and Opsomer (2003)). The main difficulty in obtaining general and comprehensive results for these estimators derives from the lack of closed-form expressions for them. For generalized additive models, following Opsomer and Kauermann (2000, 2002) and Berhane and Tibshirani (1998), one can extend the properties of the backffiting estimators for additive models by inspecting the behaviour of the weights used in the GLSA. More specifically, consider the estimator obtained from the Robust GLSA where a local polynomial regression estimator is used as the univariate smoother. Recall from (7) that, at each iteration, the Robust GLSA fits a weighted additive model with responses $z_1, \ldots, z_n$ where $E(z_i) = \eta_i = f_0 + \sum_{j=1}^{p} f_j(X_j)$, $i = 1, \ldots, n$, and weights $w_i = l_i / [\sqrt{v_i}\, g'(\mu_i)] = r(\mathbf{X}_i)$, say. During the iterations, μ_i and v_i are replaced by the corresponding current estimators. The following conditions are slightly adapted from Opsomer and Kauermann (2000). Although they apply to models with two covariates, the results can be extended to the general case following Opsomer (2000). Assume that local polynomials of degree p_1 and p_2 were used, and that:

A.1 the kernel in the local polynomial regression is bounded, continuous, has compact support, $\int x^{p_1+1} K(x) dx \neq 0$, and $\int x^{p_2+1} K(x) dx \neq 0$;

A.2 the covariates have a joint distribution with density function $h(x_1, x_2) > 0$ for all (x_1, x_2)

in its support set, h and the marginal densities h_1 and h_2 are bounded, continous, differentiable, have compact support and the first derivatives of h_1 and h_2 have a finite number of sign changes over their support;

A.3 the weight function is bounded, continous, differentiable and positive over the support of h, the partial derivatives of the weight function and the derivatives of the conditional univariate weigths $r_j(x_j) = E(r(X)|X_j = x_j)$, $j = 1, 2$, all have a finite number of sign changes over the support of h;

A.4 the additive functions f_1 and f_2 are continous and differentiable up to order $p_1 + 1$ and $p_2 + 1$, respectively;

A.5 the variance function satisfies $v(x_1, x_2) > 0$ for all (x_1, x_2) in the support of h; and

A.6 as $n \to \infty$ the bandwidths τ_1, τ_2 satisfy: $\tau_j \to 0$ and $n\,\tau_j/\log(n) \to \infty$, $j = 1, 2$.

It is tedious but not hard to see that when the score function ψ_c is one of Huber's in (5) the weights in the Robust Generalized Local Scoring algorithm satisfy the needed regularity conditions in A.3 above, and thus we can apply Theorem 2.2 of Opsomer and Kauermann (2000) and conclude that if the starting values of μ_i are close to the true ones, then the robust estimators proposed here are asymptotically unbiased.

3.3 Robust smoothing parameter selection

Smoothers and other non-parametric regression estimators depend on a smoothing or penalty parameter that typically needs to be selected by the user. There are a number of possible ways to do this in the context of generalized additive models. In this section we focus on cross-validation methods, but refer the interested reader to Hastie and Tibshirani (1990), for a more detailed discussion.

We are particularly concerned with alleviating the potential damaging effect of outliers on automatic cross-validation procedures. Since our robust generalized local scoring algorithm was derived from quasi-likelihood score equations (rather than maximum likelihood equations), we will measure the loss incurred by the fitted value $\hat{\mu}_i^{(-i)}$ with $(y_i - \hat{\mu}_i^{(-i)})/v(\hat{\mu}_i^{(-i)})$, where $\hat{\mu}_i^{(-i)}$ is the predicted value obtained without using the i-th observation in the data. Hence the cross-validation criterion is

$$\mathrm{CV} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_{(-i)}}{v(\hat{\mu}_i^{(-i)})} \right)^2 . \tag{8}$$

Alternatively, as suggested in Hastie and Tibshirani (1990) we can look at the sum of the deviances

$$\mathrm{CV}_d = \sum_{i=1}^{n} D(y_i, \hat{y}_{(-i)}) = \sum_{i=1}^{n} d_i \, , \tag{9}$$

say, where $D(y_i, \lambda) = 2\,(l(y_i, \hat{\lambda}) - l(y_i, \lambda))$, $l(y, \lambda)$ is the log-likelihood function for a single observation y and $l(y, \hat{\lambda})$ is the maximum value of $l(y, \lambda)$.

As noted in Ronchetti and Staudte (1994) and Cantoni and Ronchetti (2001a), outliers in the data may affect automatic smoothing parameter selection methods regardless of the robustness of the fitting algorithm. In other words, even if we use a robust estimator, outliers in the data may affect negatively the selection of the smoothing parameter. It is easy to intuitively understand why this is case. If every observed response y_i in our goodness-of-fit criterion has the same importance, a value of the smoothing parameter that produces a fit adjusting most observations well and leaving a few potential outliers far will have an unduly large value of the cross-validation criterion. Hence, even if the fit is robust, the classical criteria may favour bandwidths that yield fits that accomodate the outliers. To avoid this problem we downweight outlying observations so that the "cost" of not fitting them is reduced. We consider weights of the form $w_i = \psi_c(\tilde{r}_i)/\tilde{r}_i$, and define a Robust Cross-validation criteria as:

$$\mathrm{RCV}(\alpha) = \sum_{i=1}^{n} w_i^2 \, \tilde{r}_i^2 = \sum_{i=1}^{n} \psi_c(\tilde{r}_i)^2 \, . \tag{10}$$

If one uses the deviance-based cross-validation criterion (9), similar arguments as those in the previous paragraph suggest a robustified criterion of the form

$$\mathrm{RCV}_d(\alpha) = \sum_{i=1}^{n} \rho\left(d_i\right) \ , \tag{11}$$

for a bounded function $\rho : \mathbb{R} \to \mathbb{R}_+$ (see Oh *et al.*, 2004; Oh *et al.*, 2007; Leung, 2005). Although both approaches are intutively appealing, we use the former one because it is more directly connected with the robust estimating equations we considered for this work. We also performed some numerical experiments which suggest that in practice both methods tend to give qualitatively similar fits.

4 Examples

4.1 Influenza-like Illness visits in the US

Weekly counts of Influenza-like-Illness outpatient visits in the US are reported by the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet), and available on-line (http://www.cdc.gov/ flu/ weekly/ fluactivity.htm). A season consists of weekly counts from week 40 through week 20 of the following calendar year. Historically these counts exhibit a non-linear pattern peaking around week 7. In the spring of 2009 the H1N1 flu started spreading around the world, including the US, where a large number of cases were detected and treated. We see this phenomenon reflected in the atypically high counts for weeks 17 to 20 of 2009 season (the 2008-9 season is indicated with solid circles in the left panel of Figure 2).

We consider data for the 2006-2007, 2007-2008 and 2008-2009 seasons and fit a GAM model with a logarithmic link function. The single covariate is time (week number). The GAM fit obtained with the `gam` package for `R` is quite good until week 13, approximately, and is affected noticeably by the large counts in weeks 17 to 20 of 2009. The bandwidth was obtained with leave-one-out cross validation. The penalized splines fit of Wood (2006) obtained with the `R` package mgcv produces a fit that is very close to this one, and very hard to distinguish on the plot. Moreover, the residuals plot in panel (b) of Figure 2, shows that standardized residuals for the last weeks of the 2008-2009 outbreak are only slightly higher than those observed in past seasons, indicating a potentially bad season, but not severely worse than the others in this data set. Also note how for the latter part of the 2008/9 season, the residuals are not scattered around zero as it is the case for previous weeks, indicating a potential change in the model.

Next we applied our proposed outlier-robust GAM fit to the same data, with the bandwidth chosen with robust leave-one-out cross validation as in (10). We used a score function in Huber's family with tuning constant $c = 1.5$ in both the estimation and cross-validation steps. Other choices of c in the range 1 to 3 produced results that were qualitatively the same. All calculations were carried out in R (see Section 5).

[Figure 2 about here.]

The resulting fit is displayed in Figure 2, along with the "classical" GAM fits. We see that the robust fit is not affected by the atypical large counts of the last weeks of the 2008-2009 season, and as a result, the "non-outlier" residuals remain nicely scattered around zero. More importantly, the residuals for the 4 atypically high weekly counts now appear noticeably larger than all the others one, clearly separating them from the other observations, and making them much easier to identify as potentially deviating from the pattern that fits previous years' data.

4.2 Virus isolates in Canada

Consider now weekly counts of virus isolates obtained in Canada between weeks 35 of 2006 and week 24 of 2009. These data are available on-line from the World Health Organization Global Influenza Programme FluNet at http://www.who.int/flunet. We applied both the regular and robust GAM fits to these data, using leave-one-out cross-validation to select the smoothing parameter and $c = 1.5$ in (5) for both our estimation and cross-validation steps. Figure 3 contains the plot of the data, the fits, and the associated residual plots.

[Figure 3 about here.]

The counts for the 2008 - 2009 season are indicated with solid dots (note the sharp increase in counts after week 20 of 2009). Similarly to what occurs with the US data on ILI visits discussed in the previous section, we see that the non-robust fit is affected by the sudden increase in counts in weeks 21 to 24 of 2009. The residuals plots in the right panel of Figure 3 also show clearly how these weeks appear extreme using the robust fit, but do not seem nearly as atypical according to the non-robust fit.

4.3 Synthetic example revisited

We added the proposed robust GAM fit to the three artificial data sets in Figure 1. Figure 4 contains the fits obtained with the GLSA, the penalized splines approach of Wood (2006) and the robust GLSA discussed here.

[Figure 4 about here.]

The bandwidth was chosen using leave-one-out cross-validation for the classical GAM fit, and the robust leave-one-out cross-validation criterion for the robust GAM fit. In particular, panel (c) shows the effect that outliers can have on data-based selection methods for the smoothing parameter, where the selected fit results in oversmoothing. Also note that in panels (a) and (b) the robust and classical GAM fits are very close to each other in the regions where there are no outliers present. Overall, we find the robust fit to be much closer to the mean function of the model that generated the majority of the observations than either of the standard gam fits.

4.4 Bivariate example

To illustrate the performance of our method on an example with more than one covariate, consider the model $Y|(X_1, X_2) \sim \mathcal{P}\left(\exp\left(f_1(X_1) + f_2(X_2)\right)\right)$ where

$$f_1\left(X_1\right) = 3\,\sin\left(X_1 \pi\, 5/4\right)\ , \qquad f_2\left(X_2\right) = 3\,\cos\left(X_2\, \pi/2\right)\ ,$$

and $X_j \sim \mathcal{U}(0,1),\ j = 1, 2$, are independent random variables. To explore the sensitivity of the fit to the presence of outliers we randomly generated $W_1, \ldots, W_{400}$ iid random variables from the model

$$W \;=\; (1 - B)\; Y \;+\; B\tilde{Y}\,,$$

where $Y|(X_1, X_2)$ is as above, $\tilde{Y}|(X_1, X_2) \sim \mathcal{P}\left(\exp\left(f_1(X_1) + f_2(X_2) + 500\right)\right)$, and $P\left(B = 1\right) = 1 - P\left(B = 0\right) = 0.15$, independent from the X's and the Y's. Figure 5 shows the data, the true mean surface and the fits obtained with the GAM, MGCV and RGAM methods.

[Figure 5 about here.]

We can see that only the robust fit is able to estimate the true surface mean reasonably well. Both the back-fitting and the penalized splines estimates leave most of the observations below the fitted surface. This can also be seen in the residuals plots in Figure 6. Note that the residuals obtained with the robust fit show the majority of the points scattered around the fitted surface without any discernible trend, with the remaining points lying relatively far away. In contrast, the mean surface estimated with both non-robust fits lies between the outliers and the non-outlying points. Moreover, the residuals of the “good points” mistakenly show a curved structure that may lead to (wrongly) questioning the validity of the model for these observations.

[Figure 6 about here.]

5 Simulation Study

We performed a simulation study to explore the properties of the proposed fit in a variety of situations. More specifically we generated data sets with and without outliers, with different mean functions and proportions and location of the outliers. For each simulation scenario described below we generated 500 samples and calculated the mean squared error incurred in each of them:

$$\mathrm{MSE}_j \;=\; \frac{1}{n}\sum_{i=1}^{n}\left(\mu_i - \hat{\mu}_i\right)^2 \;, \quad j = 1, \ldots, 500 \,.$$

To compare the robust and non-robust fits we report the median and MAD of these 500 MSE's for each method considered here.

We considered Poisson and Binomial responses, with outliers either at the beginning or at the end of the range of the covariate used in the experiment in the following manner. Let $(\tilde{y}_1, x_1)$, ..., $(\tilde{y}_n, x_n)$ be the data, and consider the observations $\tilde{y}_j$ with $x_{(k_1)} \leq x_j \leq x_{(k_2)}$, for fixed numbers k_1 and k_2, where $x_{(m)}$ is the m-th order statistic. Then

$$y_j \;=\; (1 - z_j)\; \tilde{y}_j + z_j\, w_j \;,$$

where $z_j \sim \mathcal{B}(1, \delta)$, and $w_j = 10$ or $w_j \sim \mathcal{P}(30)$ depending on whether y_j has a Binomial or Poisson distribution. Hence, the number of outliers (and their position in the Poisson case) in each sample is random. The covariates were kept fixed throughout each of the simulation settings. All calculations where performed in `R` (R Development Core Team, 2009). The GAM fit was obtained using the function `gam` available in the package `gam`, the penalized splines approach of Wood (2006) was calculated with the the `R` package `mgcv` (MGCV), while the robust GAM fit was calculated using the authors' R code, which is available on-line at `http://www.stat.ubc.ca/~matias/soft.html`. We used a score function in Huber's family with tuning constant $c = 1.5$ in both the estimation and cross-validation steps. The bandwidths for GAM and RGAM were selected using leave-one-out CV and robust leave-one-out CV as in

(8) and (10), respectively.

Poisson responses For the Poisson model $Y|X \sim \mathcal{P}(\lambda(X))$ we used $\lambda(X) = f_j(X)$, $j = 1, 2$, where

$$f_1(X) = \exp(\sin(2X/120) + \cos(7X/60) + 1)\,, \tag{12}$$

and

$$f_2(X) = (65 - (\exp(3X) + X^2 10 + X^4 - 50\sin(\pi/2X^5)))/10\,. \tag{13}$$

[Figure 7 about here.]

Figure 7 contains typical data sets illustrating these mean functions, and the location of the outliers used in our numerical experiments. With f_1 we used $n = 80$ and $x_i = i$, $i = 1, \ldots, 80$. With f_2 we used $n = 100$ and $x_i \sim \mathcal{U}(-1, 1)$, $i = 1, \ldots, 100$. The covariates were kept fixed throughout the simulation.

[Table 1 about here.]

[Table 2 about here.]

Summaries of the mean squared errors are displayed in Tables 1 and 2.

These results show that the robust GAM fit is able to resist the damaging effect of outliers for both mean functions and location of outliers. The median MSE of the non-robust GAM fits across the 500 simulated data sets can be as much as 14 times higher than that of the robust GAM fit. Also note that the MAD of these MSE's is also consistently (and noticeably) larger than that corresponding to the robust GAM fits, indicating that the robust GAM fit is also more stable than its non-robust counterparts. The difference in median MSE's is very large in all cases where outliers are present, while, as expected, the robust fit does slightly worse than the non-robust one when there are no outliers in the data.

Binomial responses For the Binomial model we used

$$Y|\,X \sim \mathcal{B}\,(10,\, p\,(X))\ , \tag{14}$$

with $\operatorname{logit}(p\,(X)) = f_j\,(X)$, $j = 1, 2$, and

$$f_1\,(X) = -\sin(5X/120)/0.8 - 1\,, \tag{15}$$

and

$$f_2\,(X) = \cos(7X/120)/1.2 + 1/1.2\,. \tag{16}$$

[Figure 8 about here.]

Figure 8 contains two typical data sets generated with these models. We used $n = 100$ and $x_i = i$, $i = 1, \ldots, 100$. The covariates were kept fixed throughout the simulation.

Summaries of the mean squared errors are displayed in Table 3.

[Table 3 about here.]

As in the Poisson case, these results are very favorable for the robust GAM fit. Again in this case the median MSE of the non-robust GAM fits are consistently higher than that of the robust GAM fit, while the MAD of these MSE's is larger than that corresponding to the robust GAM fit, indicating that the robust GAM fit is more stable and closer to the true mean function than either of the non-robust GAM fits.

6 Conclusion

In this paper we proposed a robust GAM fit that is resistant to the presence of observations that deviate from the pattern of the majority of the data. Moreover this fit behaves similarly to the non-robust one when there are no atypical points in the data. We illustrated the use of this method for the automatic detection of potential disease outbreaks using two data sets that reflect the current H1N1 influenza pandemic. Moreover, our simulation studies indicate that this robust GAM fit performs well across a variety of mean functions and outlier locations. We also recommend using a robust leave-one-out cross-validation criterion to select the bandwidth (or penalty) parameter of the smoother used in the back-fitting algorithm. All the calculations were carried out in `R` using code that is publicly available at `http://www.stat.ubc.ca/~matias/soft.html`.

7 Appendix

We need to solve $f(\boldsymbol{\beta}) = \mathbf{0}$ where

$$f(\boldsymbol{\beta}) = \sum_{i=1}^{n} [(\psi_c(r_i) \frac{1}{\sqrt{v_i}} \frac{\partial \mu_i}{\partial \eta_i} \mathbf{x}_i - a(\boldsymbol{\beta})] \, ,$$

or

$$\sum_{i=1}^{n} (\psi_c(r_i) \frac{1}{\sqrt{v_i}} \frac{\partial \mu_i}{\partial \eta_i} \mathbf{x}_i) - n \, a(\boldsymbol{\beta}) = \mathbf{0} \, .$$

Note that

$$a(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} E(\psi_c(r_i)) \frac{1}{\sqrt{v_i}} \frac{\partial \mu_i}{\partial \eta_i} \mathbf{x}_i \, .$$

Hence the equation to be solved is

$$f(\boldsymbol{\beta}) = \sum_{i=1}^{n} [\psi_c(r_i) - E(\psi_c(r_i))] \frac{1}{\sqrt{v_i}} \frac{\partial \mu_i}{\partial \eta_i} \mathbf{x}_i = \mathbf{0} \, .$$

If we use Newton-Raphson iterations to solve it, at the j-th step of the iteration, $\boldsymbol{\beta}^{(j+1)}$ satisfies $(\nabla f(\boldsymbol{\beta}^{(j)})) \, (\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}) \; = \; -f(\boldsymbol{\beta}^{(j)})$, where ∇f denotes the gradient of f. Fisher's scoring method replaces the observed Hessian ∇f by its expected value so that our iterations become

$$E(\nabla f(\boldsymbol{\beta}^{(j)})) \, \boldsymbol{\beta}^{(j+1)} \; = \; E(\nabla f(\boldsymbol{\beta}^{(j)})) \, \boldsymbol{\beta}^{(j)} - f(\boldsymbol{\beta}^{(j)}) \, .$$

Note that

$$\nabla f(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \frac{\partial h_i}{\partial \boldsymbol{\beta}} \frac{1}{\sqrt{v_i}} + h_i \frac{\partial}{\partial \boldsymbol{\beta}} \left(\frac{1}{\sqrt{v_i}} \right) \right\} \frac{\partial \mu_i}{\partial \eta_i} \mathbf{x}_i \, ,$$

where $h_i = \psi_c(r_i) - E(\psi_c(r_i))$. Since $E(h_i) = 0$ we have

$$E(\nabla f(\boldsymbol{\beta})) = \sum_{i=1}^{n} E(\frac{\partial h_i}{\partial \beta})\frac{1}{\sqrt{v_i}}\frac{\partial \mu_i}{\partial \eta_i}\mathbf{x}_i\,.$$

Let $\gamma_i = E\left[\partial h_i/\partial \boldsymbol{\beta}\right]$. Then

$$f(\boldsymbol{\beta}) = \sum_{i=1}^{n} h_i \frac{1}{\sqrt{v_i}}\frac{\partial \mu_i}{\partial \eta_i}\mathbf{x}_i\,,$$

and

$$E(\nabla f(\boldsymbol{\beta}))\,\boldsymbol{\beta} \;=\; \sum_{i=1}^{n} \gamma_i \frac{1}{\sqrt{v_i}}\frac{\partial \mu_i}{\partial \eta_i}\,\mathbf{x}_i'\boldsymbol{\beta}\,.$$

It is not difficult to see that

$$\gamma_i \;=\; \left[-\frac{E(\psi'(r_i))}{\sqrt{v_i}}\,\frac{\partial \mu_i}{\partial \eta_i} - \frac{1}{2}\,E(\psi'(r_i)r_i)\frac{1}{v_i}\,\frac{\partial v_i}{\partial \eta_i}\; - E(\frac{\partial}{\partial \eta_i}E(\psi(r_i)))\right]\mathbf{x}_i = \ell_i\,\mathbf{x}_i\,,$$

say, where $\ell_i \in \mathbb{R}$. Therefore

$$E\left(\nabla f(\boldsymbol{\beta}^{(j)})\right)\boldsymbol{\beta}^{(j+1)} = \sum_{i=1}^{n} \ell_i\,\frac{1}{\sqrt{v_i}}\frac{\partial \mu_i}{\partial \eta_i}\mathbf{x}_i\,\mathbf{x}_i'\,\boldsymbol{\beta}^{(j+1)}\,,$$

and

$$\begin{aligned} E(\nabla f(\boldsymbol{\beta}^{(j)}))\,\boldsymbol{\beta}^{(j)} - f(\boldsymbol{\beta}^{(j)}) \;&=\; \sum_{i=1}^{n} \frac{\ell_i}{\sqrt{v_i}}\frac{\partial \mu_i}{\partial \eta_i}\,\mathbf{x}_i\eta_i - \sum_{i=1}^{n} h_i\frac{1}{\sqrt{v_i}}\frac{\mu_i}{\eta_i}\mathbf{x}_i\,, \\ &=\; \sum_{i=1}^{n}(\ell_i\eta_i - h_i)\frac{1}{\sqrt{v_i}}\frac{\partial \mu_i}{\partial \eta_i}\mathbf{x}_i \\ &=\; \sum_{i=1}^{n}\ell_i\frac{1}{\sqrt{v_i}}\frac{\partial \mu_i}{\partial \eta_i}(\eta_i - \frac{h_i}{\ell_i})\mathbf{x}_i\,. \end{aligned}$$

Finally, our iterations $E\left(\nabla f(\boldsymbol{\beta}^{(j)})\right)\boldsymbol{\beta}^{(j+1)} = E\left(\nabla f(\boldsymbol{\beta}^{(j)})\right)\boldsymbol{\beta}^{(j)} - f(\boldsymbol{\beta}^{(j)})$ become

$$\left(\sum_{i=1}^{n}\ell_i\,\frac{1}{\sqrt{v_i}}\frac{\partial \mu_i}{\partial \eta_i}\mathbf{x}_i\,\mathbf{x}_i'\right)\,\boldsymbol{\beta}^{(j+1)} = \sum_{i=1}^{n}\ell_i\frac{1}{\sqrt{v_i}}\frac{\partial \mu_i}{\partial \eta_i}(\eta_i - \frac{h_i}{\ell_i})\mathbf{x}_i\,,$$

which has the form

$$\left(\sum_{i=1}^{n}\omega_i\mathbf{x}_i\mathbf{x}_i'\right)\,\boldsymbol{\beta}^{(j+1)} = \sum_{i=1}^{n}\omega_i\mathbf{x}_i\,z_i\,,$$

where $z_i = \eta_i - h_i/\ell_i$ and $\omega_i = \left(\ell_i/\sqrt{v_i}\right)\,\frac{\partial \mu_i}{\partial \eta_i}$.

References

[1] Berhane, K. and Tibshirani, R. (1998), "Generalized additive models for longitudinal data," *The Canadian Journal of Statistics*, **26**, 517-535.

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

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

[4] Cleveland, W.S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, **74**, 829–836.

[5] Croux, C., Gijbels, I. and Prosdocimi, I. (2010), "Robust estimation of mean and dispersion functions in extended generalized additive models," Unpublished manuscript.

[6] Davies, P.L. and Gather, U. (2005), "Breakdown and groups," *The Annals of Statistics*, **33**, 977-988.

[7] Dominici, F., McDermott, A., Zeger, S.L., and Samet, J.M. (2002), "On the use of generalized additive models in time-series studies of air pollution and health," *American Journal of Epidemiology*, **156**(3), 193-203.

[8] Donoho, D.L. and Huber, P.J. (1983), "The notion of breakdown point," in *A Festschrift for Erich Lehmann*, (P. J. Bickel, K. A. Doksum and J. L. Hodges, eds.) 157-184. Wadsworth, Belmont, CA.

[9] Figueiras, A., Roca-Pardiñas, J., Cadarso-Suárez, C. (2005), "A bootstrap method to avoid the effect of concurvity in generalised additive models in time series studies of air pollution," *Journal of Epidemiology and Community Health*, **59**, 881-884.

[10] Friedman, J.H. and Stuetzle, W. (1981), "Projection pursuit regression," *Journal of the American Statistical Association*, **76**, 817-823.

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

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

[13] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986), *Robust statistics: the approach based on influence functions*, New York: John Wiley and Sons.

[14] Huber, P.J. (1985), "Projection pursuit (with discussion)," *The Annals of Statistics*, **13**, 435-475.

[15] Kauermann, G. and Opsomer, J.D. (2003), "Local likelihood estimation in generalized additive models," *Scandinavian Journal of Statistics*, **30**, 317-337.

[16] 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.

[17] Nelder, J.A. and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, Series B*, **135**, 370-384.

[18] Maronna, R.A., Martin, D.R., and Yohai, V.J. (2006), *Robust statistics: theory and methods*, New York: John Wiley and Sons.

[19] McCullagh, P. and Nelder, J.A. (1999), *Generalized linear models*, New York: Chapman & Hall / CRC.

[20] Oh, H-S., Nychka, D. W. and Lee, T. C. M. (2007), "The role of pseudo-data for robust smoothing with application to wavelet regression," *Biometrika*, **94**, 893-904.

[21] Oh, H-S., Nychka, D. W., Brown, T. and Charbonneau, P. (2004), "Period analysis of variable stars by robust smoothing," *Applied Statistics*, **53**, 15-30.

[22] Opsomer, J.D. (2000), "Asymptotic properties of backfitting estimators," *Journal of Multivariate Analysis*, **73**, 166-179.

[23] Opsomer, J.D. & Kauermann, G. (2000), "Weighted local polynomial regression, weighted additive models and local scoring," Preprint 00-7, Department of Statistics, Iowa State University

[24] Opsomer, J.D. & Kauermann, G. (2002), "A Note on Local Scoring and Weighted Local Polynomial Regression in Generalized Additive Models," Unpublished manuscript available on-line at `http://www.stat.colostate.edu/~jopsomer/papers/Local_scoring.pdf`

[25] R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[26] Ramsay, T., Burnett, R.T., and Krewski, D. (2003), "The effect of concurvity in generalized additive models linking mortality to ambient particulate matter," *Epidemiology*, **14**(1), 18-23.

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

[28] Rousseeuw, P.J. and van Zomeren, B.C. (1990), "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical Association*, **85**, 633-639.

[29] Ruppert, D., Wand, M.P. and Carroll, R.J. (2003), *Semiparametric regression.* Cambridge series in Statistical and probabilistic mathematics, New York: Cambridge University Press.

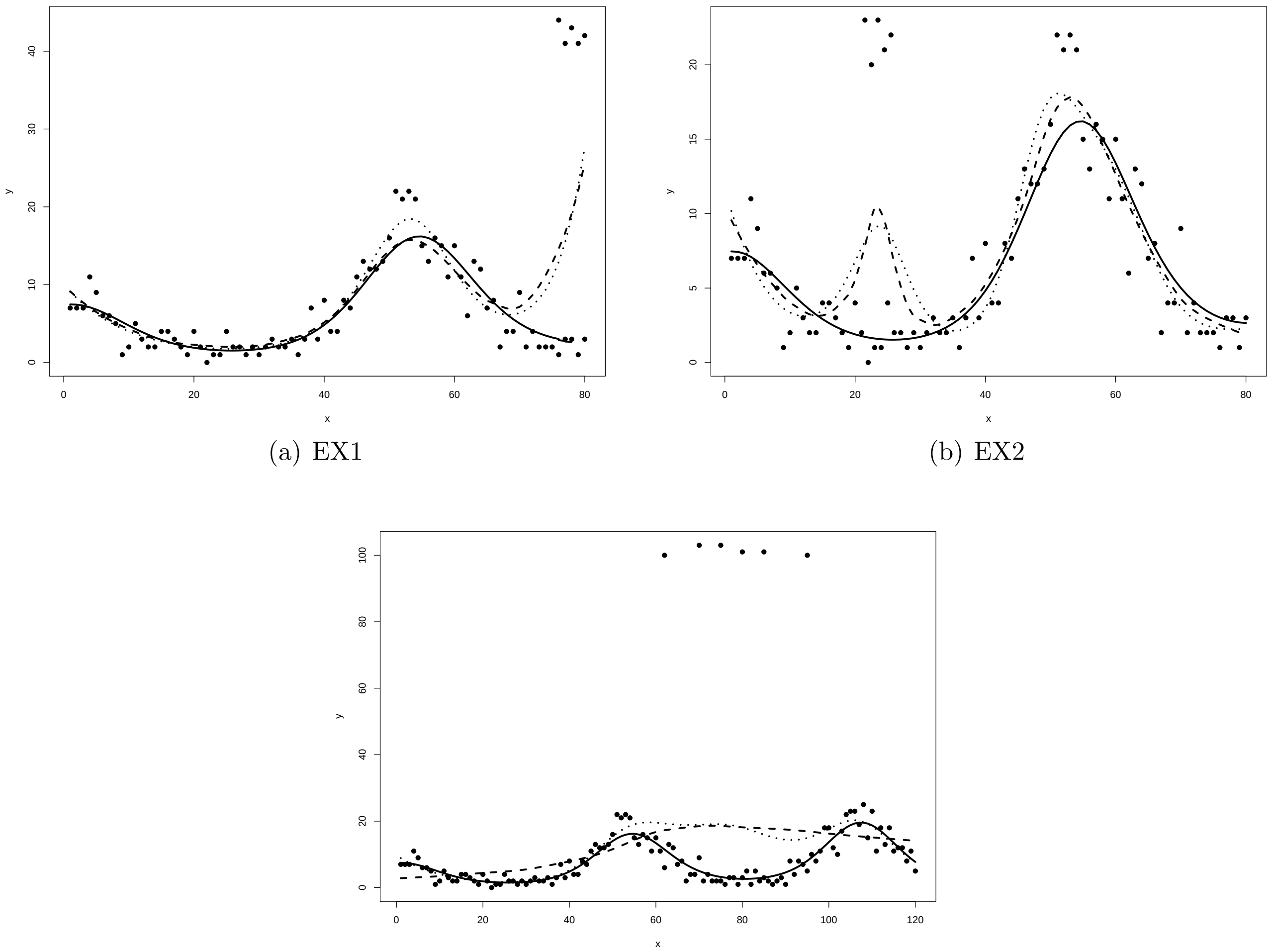
[30] Silverman, B.W. (1985), "Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion)," *Journal of the Royal Statistical Society,*

Series B, **47**, 1-52.

[31] 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.

[32] Wedderburn, R. W. M. (1974), "Quasi-likelihood functions, generalized linear mdoels and the Gauss-Newton method," *Biometrika*, **61**, 439-447.

[33] Wood, S. (2006). *Generalized Additive Models. An introduction with R*, Boca Raton, Florida: Chapman & Hall / CRC.

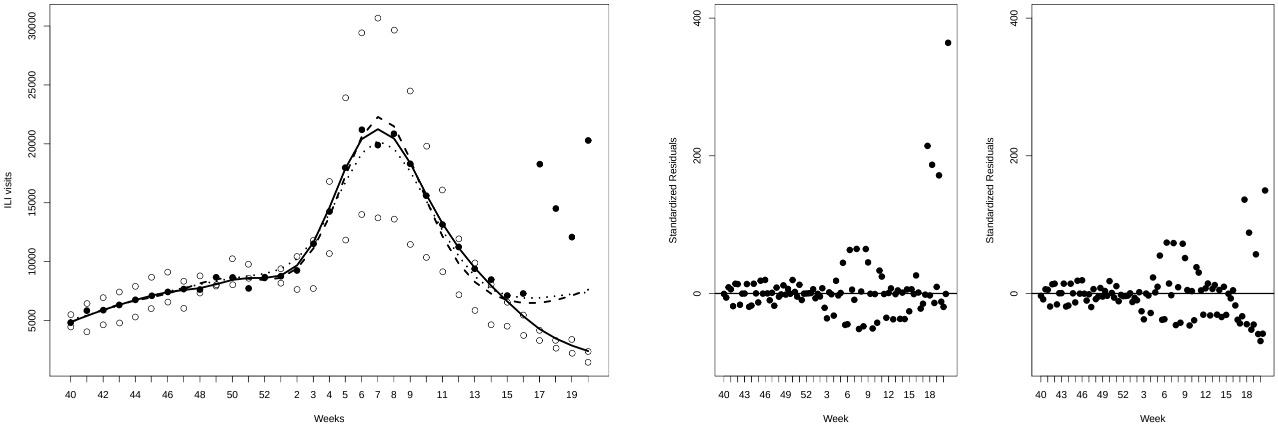


(a) EX1

(b) EX2

(c) EX3

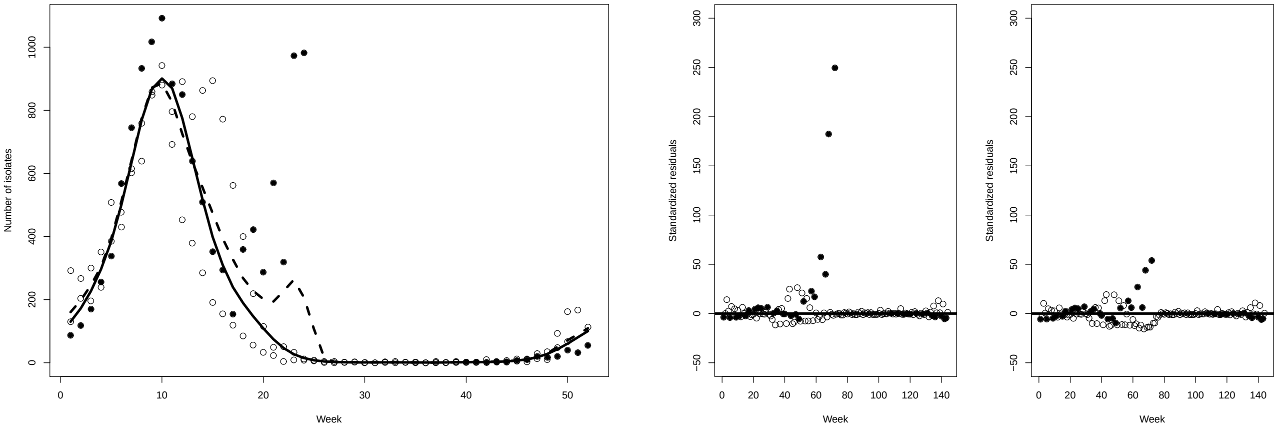
Figure 1: Three synthetic examples of the effect of a small proportion of outliers on the GAM fit. The solid line indicates the true mean function, the dashed line shows the GAM fit, obtained with the library gam in R, and the dotted line corresponds to the GAM fit obtained with the library mgcv in R.



(a) Robust and standard GAM fits

(b) Residuals

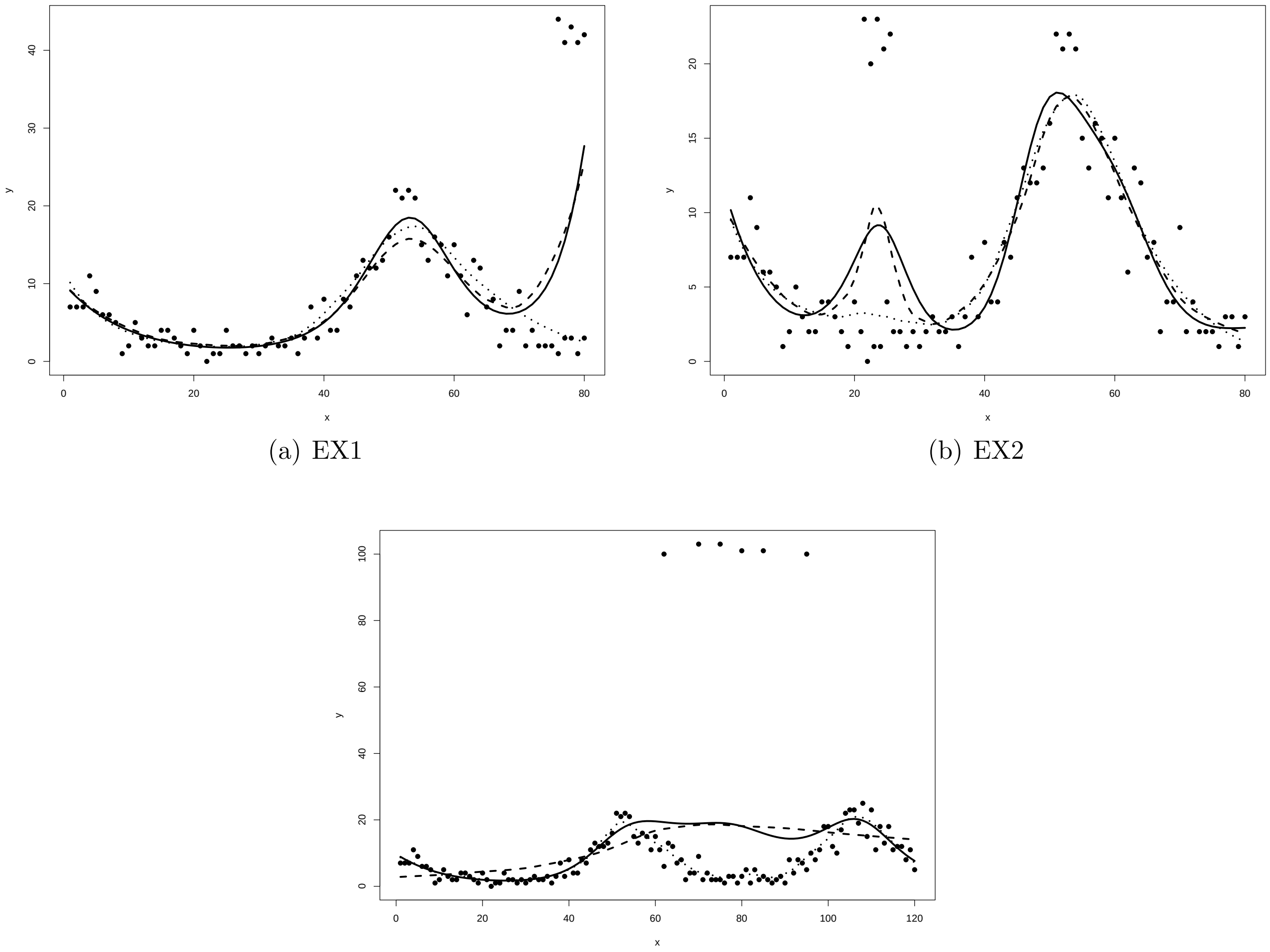
Figure 2: GAM, mgcv, and robust GAM fits (panel (a)) to the weekly number of Influenza-Like-Illness visits in the US for the 2006-2008 flu seasons, and the corresponding standardized residuals (panel (b)). In panel (a) the dotted line denotes the GAM fit, the dashed line corresponds to the mgcv fit, while the robust GAM fit is indicated with a solid line. Panel (b) contains the standardized residuals associated with the robust and standard GLSA algorithms (left and right panels, respectively). The residuals of the mgcv fit are very similar to those of the non-robust GLSA fit.



(a) Robust and standard GAM fits

(b) Residuals from both fits

Figure 3: Weekly counts of virus isolates in Canada between 2006 and 2009. In panel (a) the dashed line denotes the GAM fit while the robust GAM fit is indicated with a solid line. Panel (b) contains the standardized residuals associated with the robust (left plot) and standard (right plot) fits.

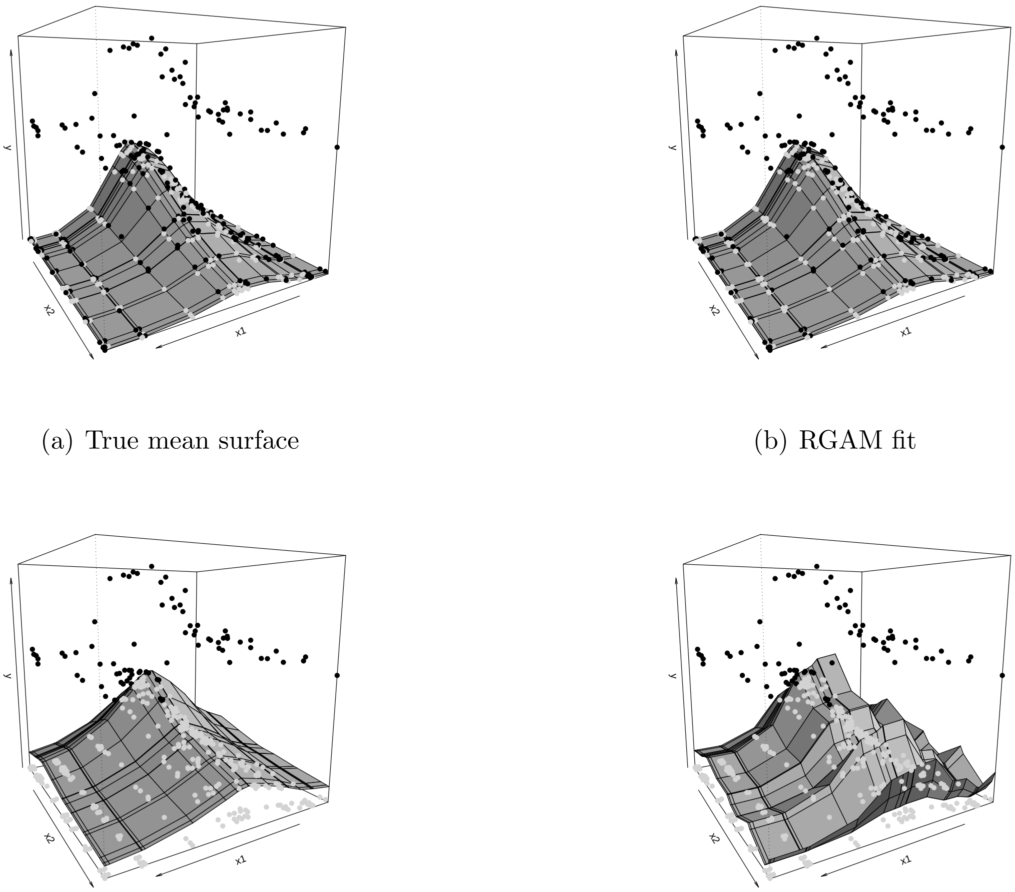


(a) EX1

(b) EX2

(c) EX3

Figure 4: Synthetic datasets illustrating the effect of three different outlier settings. The dashed line represents the GAM fit obtained with the library gam in R, the solid line corresponds to the GAM fit obtained with the library mgcv in R, while the robust GAM fit is indicated with a dotted line.



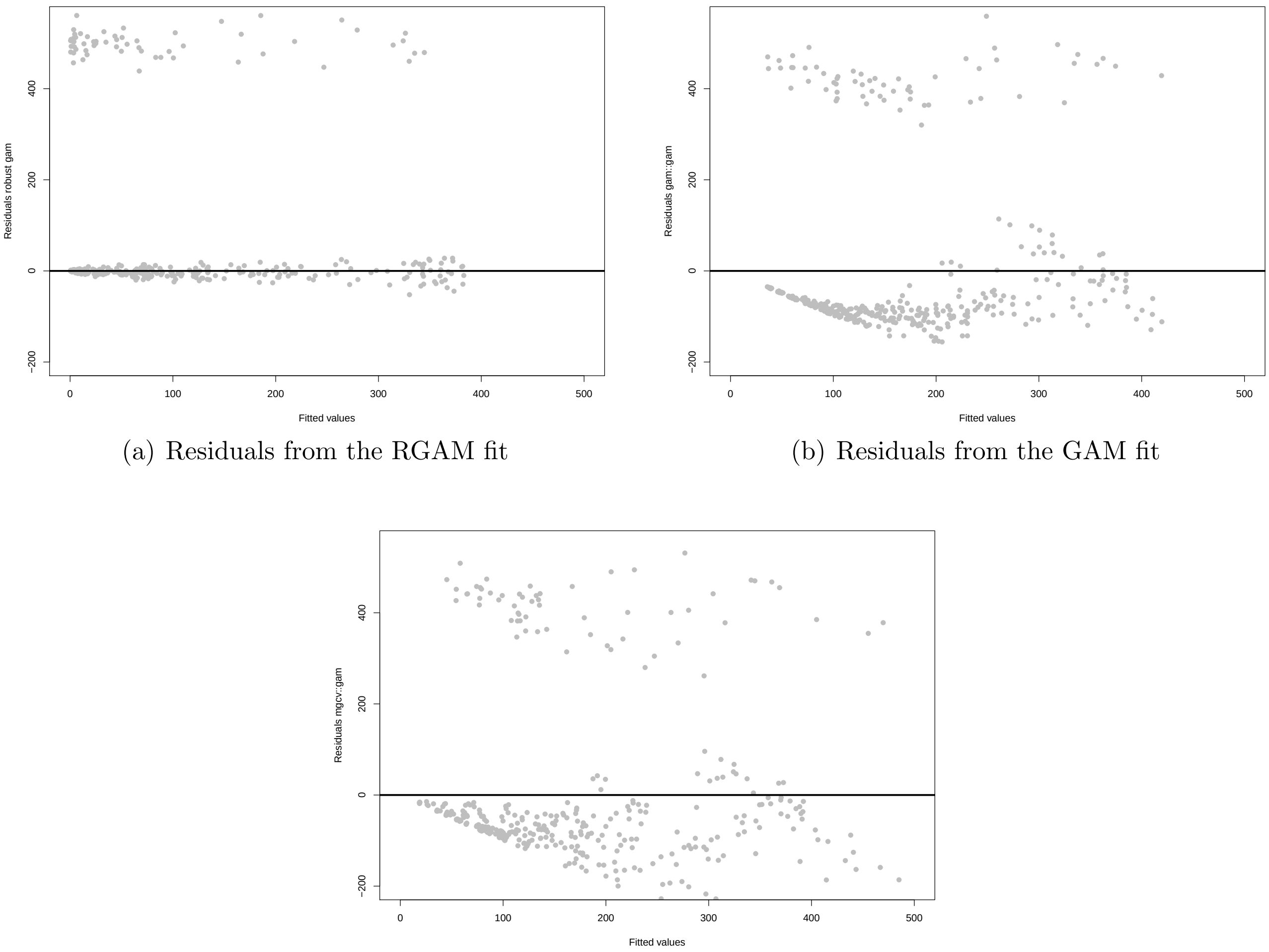
(a) True mean surface

(b) RGAM fit

(c) GAM fit

(d) MGCV fit

Figure 5: A bivariate synthetic example. Solid black points indicate observations above the surface, while the light gray ones are below the surface.

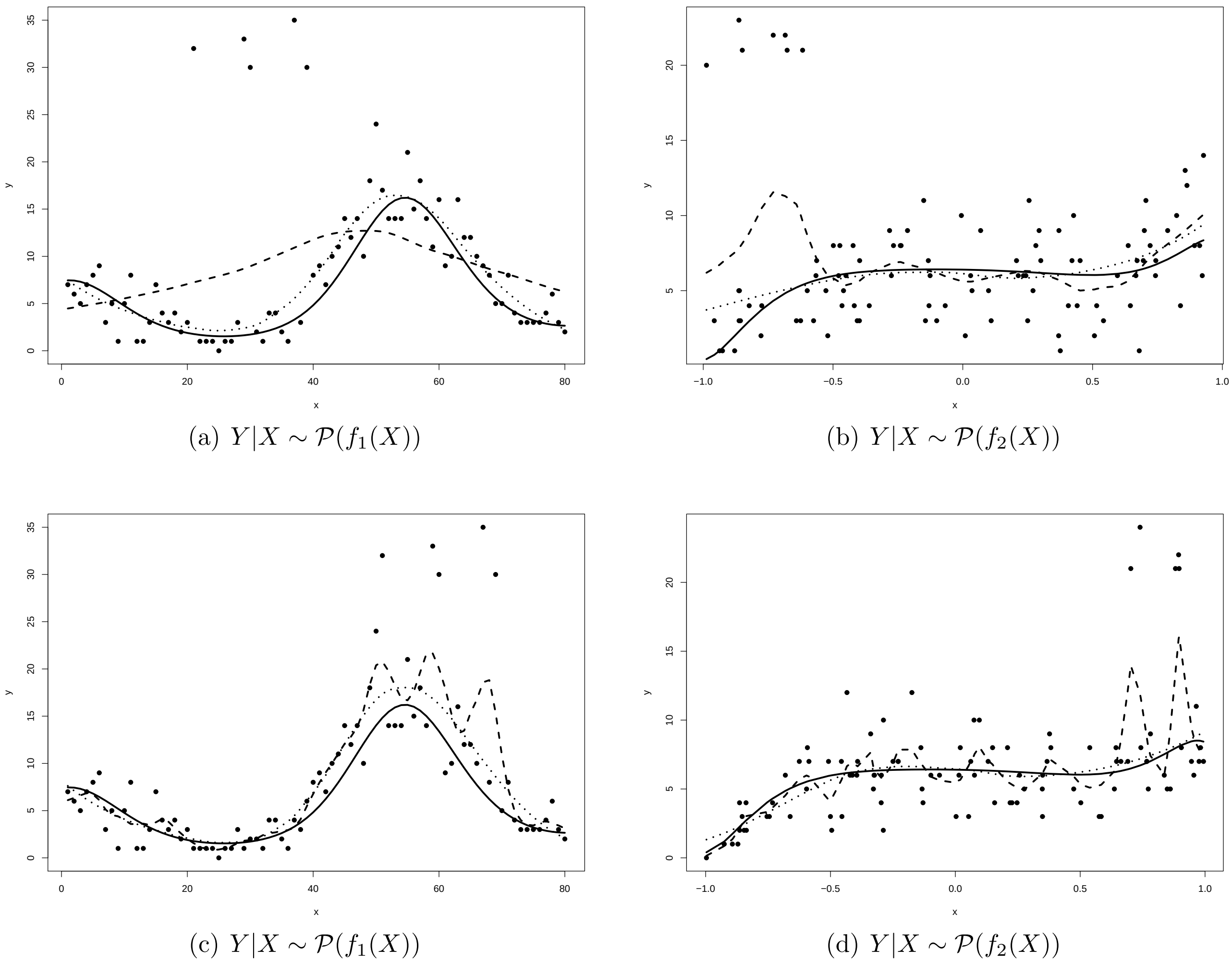


(a) Residuals from the RGAM fit

(b) Residuals from the GAM fit

(c) Residuals from the MGCV fit

Figure 6: Residual plots for the bivariate example.



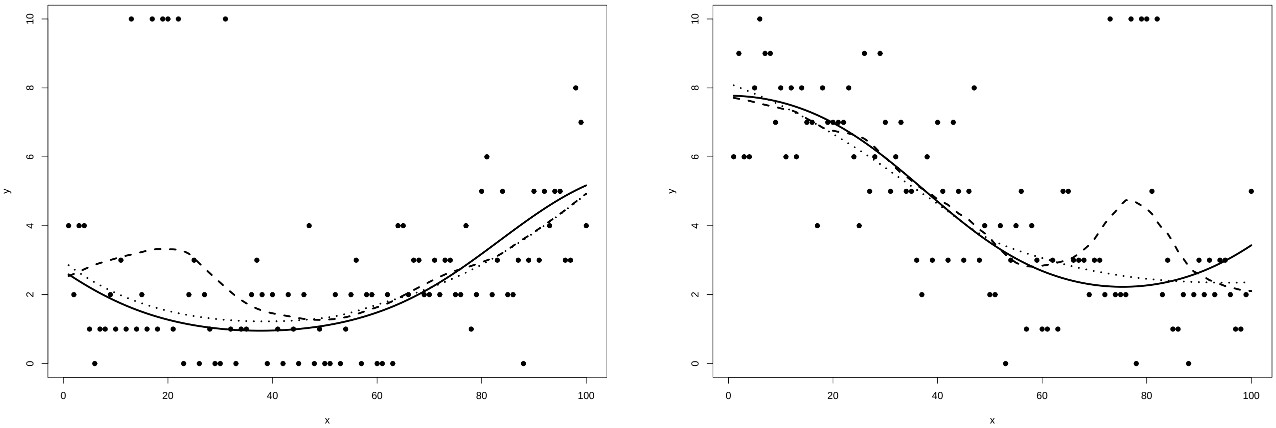
(a) $Y|X \sim \mathcal{P}(f_1(X))$

(b) $Y|X \sim \mathcal{P}(f_2(X))$

(c) $Y|X \sim \mathcal{P}(f_1(X))$

(d) $Y|X \sim \mathcal{P}(f_2(X))$

Figure 7: Illustrative examples of simulation data sets with $\delta = 0.20$ and a Poisson response. The solid line denotes the true mean function. The dashed line represents the GAM fit obtained with library gam in R; while the robust GAM fit is indicated with a dotted line.



(a) $\operatorname{logit}(p(X)) = f_1(X)$

(b) $\operatorname{logit}(p(X)) = f_2(X)$

Figure 8: Illustrative examples of simulation data sets with $\delta = 0.20$ and a Binomial response. The solid line denotes the true mean function. The dashed line represents the GAM fit obtained with library gam in R; while the robust GAM fit is indicated with a dotted line.

	Beginning			End		
δ	R-GAM	GAM	MGCV	RGAM	GAM	MGCV
0.00	0.85 (0.48)	0.70 (0.37)	0.59 (0.28)			
0.10	0.93 (0.49)	10.4 (6.83)	3.51 (3.08)	1.12 (0.63)	1.50 (1.21)	1.62 (1.37)
0.20	1.11 (0.54)	16.0 (6.10)	10.4 (8.33)	1.56 (1.00)	3.69 (2.78)	4.53 (3.50)
0.30	1.52 (0.91)	21.2 (8.44)	21.1 (13.9)	2.56 (1.88)	7.95 (5.92)	9.72 (6.51)

Table 1: Median and MAD (within parentheses) mean squared error over 500 samples for each contamination scenario and estimation method. The response variable Y satisfies $Y|X \sim \mathcal{P}(f_1(X))$ with f_1 as in (12)

	Beginning			End		
δ	R-GAM	GAM	MGCV	RGAM	GAM	MGCV
0.00	0.46 (0.26)	0.44 (0.22)				
0.10	0.46 (0.28)	1.22 (0.96)	1.38 (1.16)	0.55 (0.33)	0.85 (0.65)	1.05 (0.81)
0.20	0.53 (0.37)	2.77 (1.98)	3.72 (2.54)	0.81 (0.66)	1.86 (1.41)	2.29 (1.75)
0.30	0.95 (0.75)	6.39 (3.96)	7.66 (4.83)	1.59 (1.55)	4.14 (2.77)	4.82 (3.14)

Table 2: Median and MAD (within parentheses) mean squared error over 500 samples for each contamination scenario and estimation method. The response variable Y satisfies $Y|X \sim \mathcal{P}(f_2(X))$ with f_2 as in (13)

	$\text{logit}\,(p(X)) = f_1(X)$			$\text{logit}\,(p(X)) = f_2(X)$		
δ	R-GAM	GAM	MGCV	RGAM	GAM	MGCV
0.00	0.85 (0.64)	0.73 (0.44)	0.66 (0.47)	1.09 (0.79)	1.11 (0.68)	0.98 (0.61)
0.10	1.07 (0.78)	4.09 (2.52)	4.79 (3.79)	1.21 (0.90)	2.23 (1.48)	2.17 (1.65)
0.20	1.73 (1.17)	11.9 (6.88)	15.2 (8.73)	1.46 (1.10)	4.68 (3.37)	5.24 (4.31)
0.30	3.11 (1.98)	25.3 (11.6)	30.3 (13.2)	1.99 (1.46)	8.81 (5.67)	10.7 (7.35)

Table 3: Median and MAD (within parentheses) mean squared error over 500 samples. Response variable follows model (14) for each contamination scenario and estimation method. The regression functions f_1 and f_2 are defined in (15) and (16), respectively.