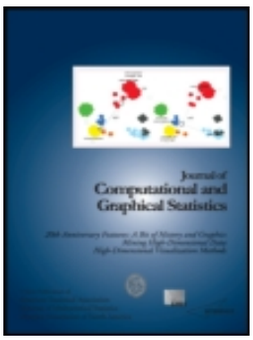

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

Robust Estimation for Generalized Additive Models

Raymond K. W. Wong, Fang Yao & Thomas C. M. Lee

To cite this article: Raymond K. W. Wong, Fang Yao & Thomas C. M. Lee (2014) Robust Estimation for Generalized Additive Models, Journal of Computational and Graphical Statistics, 23:1, 270-289, DOI: 10.1080/10618600.2012.756816

To link to this article: http://dx.doi.org/10.1080/10618600.2012.756816

Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

Robust Estimation for Generalized Additive Models

Raymond K. W. WONG, Fang YAO, and Thomas C. M. LEE

This article studies M-type estimators for fitting robust generalized additive models in the presence of anomalous data. A new theoretical construct is developed to connect the costly M-type estimation with least-squares type calculations. Its asymptotic properties are studied and used to motivate a computational algorithm. The main idea is to decompose the overall M-type estimation problem into a sequence of well-studied conventional additive model fittings. The resulting algorithm is fast and stable, can be paired with different nonparametric smoothers, and can also be applied to cases with multiple covariates. As another contribution of this article, automatic methods for smoothing parameter selection are proposed. These methods are designed to be resistant to outliers. The empirical performance of the proposed methodology is illustrated via both simulation experiments and real data analysis. Supplementary materials are available online.

Key Words: Bounded score function; Generalized information criterion; Generalized linear model; Robust estimating equation; Robust quasi-likelihood; Smoothing parameter selection.

1. INTRODUCTION

Generalized additive models (GAMs) (e.g., Hastie and Tibshirani 1990) are extensions of additive models (AMs). They can be applied to handle a wider class of data such as binary and count data. Their parametric counterparts are the well-known generalized linear models (GLMs) (e.g., McCullagh and Nelder 1989). Both GLMs and GAMs assume the response variable follows an exponential family distribution. They also share the same goal of modeling the relationship between the predictors and the mean of the response. While GLMs achieve this goal by using parametric methods, GAMs allow nonparametric fitting and hence are more flexible.

Robust estimation for GLMs has been widely studied. For example, robust logistic regression has been considered by Copas (1988) and Carroll and Pederson (1993). For more

Raymond K. W. Wong is associated with the Department of Statistics, University of California at Davis, One Shields Avenue, Davis, CA 95616 (E-mail: *rkwwong@ucdavis.edu*). Fang Yao is associated with the Department of Statistics, University of Toronto, 100 St. George Street, Toronto, Ontario M5S 3G3, Canada (E-mail: *fyao@utstat.toronto.edu*). Thomas C. M. Lee is associated with the Department of Statistics, University of California at Davis, One Shields Avenue, Davis, CA 95616 (E-mail: *tcmlee@ucdavis.edu*).

Journal of Computational and Graphical Statistics, Volume 23, Number 1, Pages 270–289
DOI: 10.1080/10618600.2012.756816
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jcgs*.

general settings, Stefanski, Carroll, and Ruppert (1986) and Künch, Stefanski, and Carroll (1989) proposed using bounded score functions to define robust estimates; Morgenthaler (1992) used L_1 norm for likelihood calculations; and Preisser and Qaqish (1999) and Cantoni and Ronchetti (2001) constructed robust estimating equations for conducting, respectively, robust estimation and robust inference procedures. For the robust estimation of GAMs, two recent articles are devoted to the subject: Alimadad and Salibian-Barrera (2011) and Croux, Gijbels, and Prosdocimi (2011). The estimation procedures developed in these two articles produce promising empirical results. However, they also have some minor shortcomings: the procedure by Alimadad and Salibian-Barrera (2011) uses brute force cross-validation for smoothing parameter selection and hence it is computationally expensive, while no theoretical support is provided for the method by Croux, Gijbels, and Prosdocimi (2011).

Following the idea by Stefanski, Carroll, and Ruppert (1986) and Preisser and Qaqish (1999), we use robust estimating equations to define robust estimates for GAMs. Computing the corresponding robust estimates is not always trivial as it requires the solving of a system of nonlinear equations. To circumvent this issue, we study the theoretical properties of a new transformation that is capable of converting this nonlinear problem into a least-squares type calculation. This transformation contains unknown quantities so it cannot be performed in practice. However, it motivates an efficient algorithm for computing the robust estimates. The main idea is to decompose the original nonlinear equation-solving problem into a sequence of relatively fast and well-studied AM fittings. It can also be paired with different nonparametric smoothers, and applied to problems with multiple covariates. In this work, we also develop automatic and reliable methods for choosing the amount of smoothing. These methods are based on the work by Konishi and Kitagawa (1996), and they accommodate the presence of outliers and worked well in simulations.

The rest of this article is organized as follows. Background material is provided in Section 2. The proposed robust estimators and the aforementioned computational algorithm are presented in Section 3, while some theoretical development is given in Section 4. The issue of smoothing parameter selection is then addressed in Section 5, and Section 6 discusses the case of multiple covariates. Empirical performances of the proposed methodology are evaluated via simulations and real data example in Sections 7 and 8, respectively. Concluding remarks are offered in Section 9 while technical details are deferred to the online Appendix.

2. BACKGROUND

2.1 NOTATION AND DEFINITIONS

A standard setting for GAM fitting is as follows. The responses $\{y_i\}_{i=1}^n$ are assumed to be independent and follow the exponential family distribution with unknown expectation μ_i and known variance function $V(\mu_i)$. The expectation μ_i is related to the linear predictor η_i via a monotonic link function g: $\eta_i = g(\mu_i)$. Suppose there are m covariates $x_{1i}, \ldots, x_{mi}$. In GAMs η_i is modeled as a sum of smooth functions $f_1, \ldots, f_m$ of these covariates:

$$\eta_i \equiv \sum_{j=1}^{m} f_j(x_{ji}). \tag{1}$$

For clarity, we will first focus on the case when $m = 1$ and delay our discussion for $m > 1$ to Section 6. To simplify notation, when $m = 1$, we write $f_1 = f$ and $x_{1i} = x_i$ for all i. That is, Equation (1) reduces to $\eta_i = f(x_i)$.

One common nonparametric approach to estimating f is penalized basis expansion fitting. With a set of prespecified basis functions $\{b_1(\cdot), \ldots, b_p(\cdot)\}$, the smooth function f, now written as $f(x; \boldsymbol{\beta})$, is assumed to have the following representation:

$$f(x; \boldsymbol{\beta}) = \sum_{j=1}^{p} b_j(x)\beta_j, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a vector of basis coefficients. To estimate $\boldsymbol{\beta}$, regularization methods such as penalized likelihood are often used. Let $\mathbf{D}$ be a prespecified penalty matrix and $\lambda > 0$ be a smoothing parameter. Then $\boldsymbol{\beta}$ can be estimated by maximizing

$$\sum_{i=1}^{n} l(y_i, \mu_i) - \lambda \boldsymbol{\beta}^{\mathrm{T}} \mathbf{D} \boldsymbol{\beta},$$

where l is the log-likelihood function or a quasi log-likelihood function. Differentiating this functional with respect to $\boldsymbol{\beta}$ yields the following system of estimating equations

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial}{\partial \boldsymbol{\beta}} \mu_i - \mathbf{S}\boldsymbol{\beta} = \mathbf{0}, \quad \text{with } \mathbf{S} = 2\lambda \mathbf{D}. \tag{3}$$

The traditional estimator of $\boldsymbol{\beta}$, denoted as $\breve{\boldsymbol{\beta}}$, is the solution of Equation (3). Popular members of this class of nonparametric smoothers include smoothing splines (e.g., Green and Silverman 1994) and penalized regression splines (e.g., Ruppert, Wand, and Carroll 2003).

2.2 Influence Function of $\breve{\boldsymbol{\beta}}$

Influence function is a useful concept for studying the robustness properties of an estimator. Suppose the data $\{z_i\}_{i=1}^n$ are generated from a distribution $G(z, \theta)$ with an unknown parameter θ. Further suppose that the estimator $\hat{\theta}$ for θ can be expressed as $\hat{\theta} = H(\hat{G})$, where H is a functional and $\hat{G}$ is the empirical cumulative distribution function (cdf) $\hat{G}(z, \theta) = \sum_{i=1}^{n} \mathrm{I}_{\{z_i \le z\}}/n$. The influence function of $\hat{\theta}$ at z is defined as

$$\mathrm{IF}(z; H, G) = \lim_{\varepsilon \to 0} \frac{H\{(1-\varepsilon)G + \varepsilon\delta_z\} - H(G)}{\varepsilon},$$

where δ_z is the point mass 1 at z. This influence function measures the impact of an infinitesimal contamination at z on the estimator. If an estimator is robust, $\mathrm{IF}(z; H, G)$ should not be arbitrarily large for any value of z. In other words, $\mathrm{IF}(z; H, G)$ should be bounded for all values of z if the estimator is robust (for a more thorough discussion on influence functions, see e.g., Hampel et al. 1986).

Let $F(y, x)$ be the joint cdf of the response y and the covariate x. To derive the influence function for $\breve{\boldsymbol{\beta}}$, we first note that $\breve{\boldsymbol{\beta}}$ is an M-estimator defined by the score function

$$\breve{\boldsymbol{\psi}}(y_i, \boldsymbol{\beta}) = \frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial}{\partial \boldsymbol{\beta}} \mu_i - \frac{1}{n} \mathbf{S}\boldsymbol{\beta}, \tag{4}$$

and that it can be expressed as $\check{\boldsymbol{\beta}} = \check{T}(\hat{F})$, where $\hat{F}$ is the empirical joint cdf $\hat{F}(y, x) = \sum_{i=1}^{n} I(\{y_i \le y\} \cap \{x_i \le x\})/n$ and the functional $\check{T}$ is defined implicitly by $\int \check{\boldsymbol{\psi}}\{z, \check{T}(F)\} dF(z, x) = \mathbf{0}$. Here, $I(A)$ is the indicator function of the set A. From the literature by Hampel et al. (1986), its influence function is given by

$$\mathrm{IF}(y; \check{\boldsymbol{\psi}}, F) = -\left\{ \int \frac{\partial}{\partial \boldsymbol{\beta}} \check{\boldsymbol{\psi}}(z, \boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\check{T}(F)} dF(z, x) \right\}^{-1} \check{\boldsymbol{\psi}}\{y, \check{T}(F)\}.$$

Note that we use the notation $\mathrm{IF}(y; \check{\boldsymbol{\psi}}, F)$ instead of $\mathrm{IF}(y; \check{T}, F)$ to stress the dependence on the score function. Now as $\check{\boldsymbol{\psi}}$ is unbounded in y and the term inside the bigger pair of braces is a constant with respect to y, $\mathrm{IF}(y; \check{\boldsymbol{\psi}}, F)$ is also unbounded in y, suggesting that $\check{\boldsymbol{\beta}}$ is not a robust estimator.

3. METHODOLOGY

3.1 ROBUST ESTIMATING EQUATIONS

To achieve robust estimation for GAMs, one could modify the estimating Equation (3) so that the resulting influence function is bounded. Following this idea, we define our robust estimator, $\hat{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$ as the solution of

$$\sum_{i=1}^{n} \boldsymbol{\psi}(y_i, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \nu(y_i, \mu_i) \zeta(\mu_i) \frac{\partial}{\partial \boldsymbol{\beta}} \mu_i - a(\boldsymbol{\beta}) - \frac{1}{n} \mathbf{S}\boldsymbol{\beta} \right\} = \mathbf{0}, \tag{5}$$

where

$$a(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} E\left\{\nu(y_i, \mu_i)\right\} \zeta(\mu_i) \frac{\partial}{\partial \boldsymbol{\beta}} \mu_i$$

with the expectation taken with respect to the conditional distribution $y_i | x_1, \ldots, x_m$, ν is a weight function that down-weighs the effects of outliers, and ζ is a scaling function to be defined below. Note that if $\nu(y, \mu) = (y - \mu)/V(\mu)$ and $\zeta(\mu) = 1$, then $a(\boldsymbol{\beta}) = 0$, and $\boldsymbol{\psi}$ and $\hat{\boldsymbol{\beta}}$ reduces to $\check{\boldsymbol{\psi}}$ and $\check{\boldsymbol{\beta}}$, respectively. We further note that an additional weight function can be introduced to Equation (5) to alleviate the effects of high leverage points. To facilitate theoretical developments, we largely omit the use of this additional weight function, although an example is given in Section 8.

Similarly as before, we write $\hat{\boldsymbol{\beta}} = T(\hat{F})$, where now $T(\hat{F})$ is defined by $\int \boldsymbol{\psi}\{z, T(F)\}$ $dF(z, x) = \mathbf{0}$. Thus, the corresponding influence function is

$$\mathrm{IF}(y; \boldsymbol{\psi}, F) = -\left\{ \int \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\psi}(z, \boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=T(F)} dF(z, x) \right\}^{-1} \boldsymbol{\psi}\{y, T(F)\}.$$

To make $\boldsymbol{\psi}$ and hence $\mathrm{IF}(y; \boldsymbol{\psi}, F)$ bounded, one could select a bounded ν guaranteed by some function ϕ,

$$\nu(y, \mu) = \phi\left\{ \frac{y - \mu}{V^{\frac{1}{2}}(\mu)} \right\} \frac{1}{V^{\frac{1}{2}}(\mu)},$$

and a natural candidate is the following Huber-type function with cutoff c that does not depend on the sample size n and is related to the efficiency of the robust estimation:

$$\phi_c(r) = \begin{cases} r, & |r| \le c \\ c \times \text{sign}(r), & |r| > c \end{cases}. \tag{6}$$

We know that the choice ϕ_c is sufficient for most practical use, but theoretical derivations often require twice differentiability that can be achieved by imposing smoothness constraints in a small neighborhood of c. We define the scaling function $\zeta(\mu_i) = 1/E\{\phi'(r_i)\}$, where $r_i = (y_i - \mu_i)/V^{1/2}(\mu_i)$. For given μ_i, this can be separately obtained by numerical approximation or even explicit calculation (e.g., for binomial and Poisson with ϕ_c).

Notice that the estimator $\hat{\boldsymbol{\beta}}$ is an M-estimator, and that it can also be treated as a penalized likelihood estimator. This is because $\hat{\boldsymbol{\beta}}$ can also be obtained as the maximizer of

$$\sum_{i=1}^{n} q(y_i, \mu_i) - \lambda \boldsymbol{\beta}^{\mathrm{T}} \mathbf{D} \boldsymbol{\beta},$$

where the quasi-likelihood term q is given by

$$q(y_i, \mu_i) = \int_{y_i}^{\mu_i} \nu(y_i, t)\zeta(\mu_i)dt - \frac{1}{n}\sum_{j=1}^{n}\int_{y_j}^{\mu_j} E\{\nu(y_j, t)\zeta(\mu_j)\}dt \quad \text{for all } i. \tag{7}$$

This term q corresponds to a robustified likelihood of our estimation procedure and hence we shall call it robust quasi-likelihood.

3.2 A General Algorithm for Robust GAM Estimation

Due to the nonlinear nature of ν, obtaining the robust estimate $\hat{\boldsymbol{\beta}}$, the solution to Equation (5), is not a trivial calculation. Here we propose a practical algorithm for carrying out this task. The idea is to approximate the solution of Equation (5) by iteratively solving Equation (3), taking the advantage that many fast methods and softwares are available for the solving of Equation (3). We first provide an intuitive argument that motivates our algorithm.

Suppose for now good estimates $\hat{\mu}_i$'s for μ_i's are available. Define

$$\tilde{y}_i = [\nu(y_i, \hat{\mu}_i) - E\{\nu(y_i, \hat{\mu}_i)\}]\,\zeta(\hat{\mu}_i)V(\hat{\mu}_i) + \hat{\mu}_i. \tag{8}$$

Also define $\tilde{\boldsymbol{\beta}}$ as the solution to Equation (3) with the y_i's replaced by these $\tilde{y}_i$'s. That is, $\tilde{\boldsymbol{\beta}}$ solves

$$\sum_{i=1}^{n} \frac{\tilde{y}_i - \mu_i}{V(\mu_i)} \frac{\partial}{\partial \boldsymbol{\beta}} \mu_i - \mathbf{S}\boldsymbol{\beta} = \mathbf{0}. \tag{9}$$

Straightforward algebra shows that both $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ solve the same estimating equations. From this, two important questions arise: (i) are $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ the same? And if yes, (ii) what do we gain by this?

Under certain conditions, the next section establishes the asymptotic equivalence of $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$. This implies that, if the $\tilde{y}_i$'s were known, our gain would be that the robust estimator $\hat{\boldsymbol{\beta}}$ can be computed quickly as the solution to Equation (9).

Of course in practice $\tilde{y}_i$'s are unknown, but the above discussion suggests a fast iterative method for solving Equation (5). The idea is, given a current set of estimates of μ_i's, first

calculate the next estimates of $\tilde{y}_i$'s through Equation (8), then plug in these new $\tilde{y}_i$'s into Equation (9), and solve for the next set of estimates of μ_i's.

Many common GAM fitting methods, such as local scoring and iterative reweighted least-squares, for solving Equation (3) are iterative, with each iteration as effective as a weighted AM fitting. This means a direct application of the above idea for solving Equation (5) will involve iterations within iterations. The proposed algorithm eliminates this issue by further combining the calculation of $\tilde{y}_i$'s and the weighted AM fitting in one single step. Starting with initial estimates $\hat{\mu}_i^{(0)}$'s for μ_i's, this algorithm iterates until convergence the following two steps for $t = 0, 1, \ldots$:

1. Compute, for all i,

$$\tilde{z}_i^{(t+1)} = \big(\tilde{y}_i^{(t)} - \hat{\mu}_i^{(t)}\big) g'\big(\hat{\mu}_i^{(t)}\big) + \hat{\eta}_i^{(t)},$$

where

$$\tilde{y}_i^{(t)} = \big[v\big(y_i, \hat{\mu}_i^{(t)}\big) - E\big\{v\big(y_i, \hat{\mu}_i^{(t)}\big)\big\}\big]\zeta\big(\mu_i^{(t)}\big) V\big(\hat{\mu}_i^{(t)}\big) + \hat{\mu}_i^{(t)}$$

and

$$\hat{\eta}_i^{(t)} = g\big(\hat{\mu}_i^{(t)}\big).$$

2. Fit a weighted AM with $\tilde{z}_i^{(t+1)}$ as the response and use $[V(\hat{\mu}_i^{(t)})\{g'(\hat{\mu}_i^{(t)})\}^2]^{-1}$ as the weights. Take the fitted values as the next set of iterative estimates $\hat{\eta}_i^{(t+1)}$'s.

We have a few remarks about this algorithm. First, the initial estimates $\hat{\mu}_i^{(0)}$'s can be obtained as the solution of Equation (3)—that is, by nonrobust fitting. We used these initial estimates throughout all our numerical work, and they were remarkably reliable as initial guesses. Second, the above algorithm can be coupled with any types of nonparametric smoothers, as long as the weighted fitting described in Step 2 is feasible. Third, the algorithm can also be applied to cases with more than one covariates. A bivariate example is given in Section 8. Fourth, in practice, we do not update the value of $\zeta(\mu_i^{(t)})$ when the number of iterations t is bigger than a threshold, say 10. We discovered that this strategy speeds up the convergence of the algorithm without sacrificing the quality of the estimates. Finally, for problems with normal errors and identity link function, $\tilde{y}_i$ in Equation (8) recovers the pseudo data derived by Oh, Nychka, and Lee (2007), and the above algorithm reduces to their ES-algorithm for computing robust nonparametric regression estimates.

4. ASYMPTOTIC EQUIVALENCE

Recall $\tilde{\boldsymbol{\beta}}$ is the solution to Equation (9) while $\hat{\boldsymbol{\beta}}$ is the solution to Equation (5). Denote the corresponding estimates for f derived from $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ through Equation (2) as $\tilde{f}$ and $\hat{f}$, respectively. This section establishes the asymptotic equivalence between $\tilde{f}$ and $\hat{f}$. We note that the analysis below is applicable for a special but wide class of estimators, namely, those with their penalty $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta}$ derived from the norm of a reproducing kernel Hilbert space (RKHS). Briefly, $\mathcal{H}$ is called a RKHS if $\mathcal{H}$ is a Hilbert space of real-valued

functions on an index set $\mathcal{T}$, and there exists a bivariate symmetric, nonnegative definite function $K(\cdot,\cdot)$ defined on $\mathcal{T}\times\mathcal{T}$ such that the following two conditions are satisfied: (i) $K(t,\cdot)\in\mathcal{H}$, for all $t\in\mathcal{T}$, and (ii) the inner product $\langle K(t,\cdot), f(\cdot)\rangle_{\mathcal{H}} = f(t)$, for all $t\in\mathcal{T}$ and $f\in\mathcal{H}$. With this setup, the penalty matrix $\mathbf{D}$ is defined through $K(\cdot,\cdot)$ (for details, please see Wahba 1990).

In the following discussion, we use $J(\mathbf{f})$ to denote such a penalty term. Without loss of generality, we shall present the theory for a single covariate model. The Euclidean norm is denoted by $||\mathbf{x}||^2 = \sum_{i=1}^n x_i^2$ for $\mathbf{x}\in\Re^n$, while the normalized version is $||\mathbf{x}||_n^2 = ||\mathbf{x}||^2/n$.

We begin by noting that the solution of Equation (3) can be obtained by iteratively solving a sequence of weighted least-squares problems, as follows. Let $f_i = f(x_i)$, $w_{ii} = [V(\mu_i)\{g'(\mu_i)\}^2]^{-1}$, $z_i = f_i + g'(\mu_i)(y_i-\mu_i)$, $z_{w,i} = w_{ii}^{1/2} z_i$, and $f_{w,i} = w_{ii}^{1/2} f_i$; here the z_i's are typically known as the working data used during the fitting process, while $f_{w,i}$ and $z_{w,i}$ are the weighted versions of f_i and z_i, respectively. Further write $\mathbf{W} = \text{diag}\{w_{ii} : i = 1,\ldots,n\}$, $\mathbf{z} = (z_1,\ldots,z_n)^{\mathrm{T}}$, $\mathbf{f} = (f_1,\ldots,f_n)^{\mathrm{T}}$, $\mathbf{z}_w = (z_{w,1},\ldots,z_{w,n})^{\mathrm{T}}$, and $\mathbf{f}_w = (f_{w,1},\ldots,f_{w,n})^{\mathrm{T}}$; that is, $\mathbf{f}_w = \mathbf{W}^{1/2}\mathbf{f}$ and $\mathbf{z}_w = \mathbf{W}^{1/2}\mathbf{z}$. Then, given $\mathbf{z}$ and $\mathbf{z}_w$, in each iteration the next estimates for $\mathbf{f}$ and $\mathbf{f}_w$ are given, respectively, as the minimizers of

$$\frac{1}{2}(\mathbf{z}-\mathbf{f})^{\mathrm{T}}\mathbf{W}(\mathbf{z}-\mathbf{f}) + \lambda\mathbf{f}^{\mathrm{T}}\mathbf{R}^*\mathbf{f}, \quad \text{that is,} \quad \frac{1}{2}\|\mathbf{z}_w - \mathbf{f}_w\|^2 + \lambda\mathbf{f}_w^{\mathrm{T}}\mathbf{R}\mathbf{f}_w,$$

where $J(\mathbf{f}) = \mathbf{f}^{\mathrm{T}}\mathbf{R}^*\mathbf{f} = \mathbf{f}_w^{\mathrm{T}}\mathbf{R}\mathbf{f}_w$ is a RKHS representation of the penalty $\lambda\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta}$ with $\mathbf{R}^* = \mathbf{W}^{1/2}\mathbf{R}\mathbf{W}^{1/2}$. It can be shown that the estimate for $\mathbf{f}_w$ is $\check{\mathbf{f}}_w = \mathbf{H}(\lambda)\mathbf{z}_w$, where the smoothing matrix is $\mathbf{H}(\lambda) = (\mathbf{I}+2\lambda\mathbf{R})^{-1}$.

For technical convenience, define

$$\rho(z_{w,i}-t) = [\phi(z_{w,i}-t) - E\{\phi(z_{w,i}-t)\}]\zeta(\mu_i), \tag{10}$$

and $r_i = z_{w,i} - f_{w,i} = (y_i-\mu_i)/V^{1/2}(\mu_i)$. Then we have $\rho(z_{w,i}-f_{w,i}) = [\phi(r_i) - E\{\phi(r_i)\}]\zeta(\mu_i)$. We remark that $\rho(z_{w,i}-t)$ depends on μ_i, so a more explicit notation would be $\rho(z_{w,i}-t,\mu_i)$. However, for simplicity, we dropped μ_i from the notation.

Now as we have shifted our focus from $\boldsymbol{\beta}$ to f, the score functions $\check{\boldsymbol{\psi}}(y_i,\boldsymbol{\beta})$ in Equation (4) and $\boldsymbol{\psi}(y_i,\boldsymbol{\beta})$ in Equation (5) are now written as, respectively, $\check{\boldsymbol{\psi}}(\mathbf{f}_w;\mathbf{z}_w)$ and $\boldsymbol{\psi}(\mathbf{f}_w;\mathbf{z}_w)$. Their elements are

$$\check{\boldsymbol{\psi}}(\mathbf{f}_w;\mathbf{z}_w)_i = \{(z_{w,i}-f_{w,i}) - 2\lambda(\mathbf{R}\mathbf{f}_w)_i\}w_{ii}^{\frac{1}{2}} \tag{11}$$

and

$$\boldsymbol{\psi}(\mathbf{f}_w;\mathbf{z}_w)_i = \{\rho(z_{w,i}-f_{w,i}) - 2\lambda(\mathbf{R}\mathbf{f}_w)_i\}w_{ii}^{\frac{1}{2}} \tag{12}$$

respectively. Further, denote $\tilde{z}_i = f_i + g'(\mu_i)(\tilde{y}_i-\mu_i)$ and $\tilde{z}_{w,i} = w_{ii}^{1/2}\tilde{z}_i$. Write $\tilde{\mathbf{z}} = (\tilde{z}_1,\ldots,\tilde{z}_n)^{\mathrm{T}}$ and $\tilde{\mathbf{z}}_w = (\tilde{z}_{w,1},\ldots,\tilde{z}_{w,n})^{\mathrm{T}}$. With these notations, we have $\check{\boldsymbol{\psi}}(\mathbf{f}_w;\tilde{\mathbf{z}}_w) = \boldsymbol{\psi}(\mathbf{f}_w;\mathbf{z}_w)$. We shall show that, with the assumptions below, the M-type robust estimator $\hat{\mathbf{f}} = \mathbf{W}^{-1/2}\hat{\mathbf{f}}_w$ satisfying $\boldsymbol{\psi}(\hat{\mathbf{f}}_w,\mathbf{z}_w) = \mathbf{0}$ can be approximated arbitrarily well by the traditional estimator $\tilde{\mathbf{f}} = \mathbf{W}^{-1/2}\tilde{\mathbf{f}}_w = \mathbf{W}^{-1/2}\mathbf{H}(\lambda)\tilde{\mathbf{z}}_w$ satisfying $\check{\boldsymbol{\psi}}(\tilde{\mathbf{f}}_w;\tilde{\mathbf{z}}_w) = \mathbf{0}$.

($\mathcal{A}$.1) The function f is bounded, that is, $\sup_{-\infty<t<\infty}|f(t)| < \infty$.

($\mathcal{A}$.2) Assume that $\max_{1\le i\le n}\text{var}\{\phi(r_i)\} < \infty$ for all n, where $r_i = (y_i-\mu_i)/V^{1/2}(\mu_i)$, and that ϕ possesses bounded first and second derivatives.

Note that ($\mathcal{A}$.1) is to ensure that $\mu_i = g^{-1}(f_i)$'s are bounded away from singularities (including $\pm\infty$) of the functions g, g', $1/g$, $1/g'$, and $1/V$, and thus avoid unboundedness of w_{ii}, $\zeta(\mu_i)$ and $\partial\mu_i/\partial\beta$. Regarding ($\mathcal{A}$.2), as mentioned by Huber (1973), higher-order derivatives are technically convenient, but hardly essential for the results to hold. It can be easily fulfilled by modifying ϕ_c in Equation (6) with cubic splines for small intervals around $\pm c$.

($\mathcal{A}$.3) To address the dependence of λ on n, we may write $\lambda = \lambda_n$ if necessary.

(a) Let $d_n = \max_i\{\mathbf{H}(\lambda_n)_{ii}\}$, assume that $\lambda_n/n \to 0$ and $d_n \to 0$, as $n \to \infty$.

(b) There exists $K_0 < \infty$ such that $\text{tr}\{\mathbf{H}(\lambda_n)\}/\lambda_n < K_0$ for all n.

One can easily verify ($\mathcal{A}$.3) for smoothing splines based on the equivalent kernel representations (Nychka 1995). Note that, as a result of the normalization by n^{-1} in the sum of squares, the "λ" appearing in the article by Nychka (1995) is actually equal to λ_n/n in this article. In particular, ($\mathcal{A}$.3.b) involves balancing the rates of the smoothing parameter with the effective degrees of freedom, $\text{tr}\{\mathbf{H}(\lambda_n)\}$, of the smoother. Based on the equivalent kernel theory by Nychka (1995), one expects that $\text{tr}\{\mathbf{H}(\lambda_n)\} \sim (\lambda_n/n)^{-1/2m}$, where m is the order of the spline. So ($\mathcal{A}$.3.b) holds with a wide range of the smoothing parameter $\lambda_n/n \sim n^{-\kappa}$ for $0 < \kappa \leq 2m/(2m+1)$, while the fastest one $\kappa = 2m/(2m+1)$ corresponds to the optimal convergence rate of the resulting estimator.

($\mathcal{A}$.4) The space of all f's, denoted as $\mathcal{H}$, is a RKHS. Let $\mathcal{C} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq G\}$, where $\|f\|^2_{\mathcal{H}} = J(f)$ and $G > 0$ is some constant. Assume that $\mathcal{C}$ is compact with respect to the L_2 norm.

Theorem 1. If the assumptions ($\mathcal{A}$.1)–($\mathcal{A}$.4) hold, then a consistent robust estimator $\hat{f}$ exists in a neighborhood of f in $\mathcal{C}$ and $C_n = E\{\|\tilde{\mathbf{f}} - \mathbf{f}\|^2_n\} \to 0$ as $n \to \infty$, moreover,

$$C_n^{-1/2}\|\tilde{\mathbf{f}} - \hat{\mathbf{f}}\|_n \xrightarrow{P} 0.$$

This theorem implies that the robust estimate $\hat{\mathbf{f}}$ can be well approximated by $\tilde{\mathbf{f}}$. It also suggests that $\hat{\mathbf{f}}$ shares the same asymptotic squared error properties as $\tilde{\mathbf{f}}$. The proof of this theorem can be found in the online Appendix.

5. SMOOTHING PARAMETER SELECTION

For nonrobust GAMs estimation, Wood (2004, 2008) developed fast, stable, and efficient methods for smoothing parameter selection. However, most of these nonrobust selection methods cannot be directly applied to the robust GAM setting. In the context of nonparametric regression, it is known that classical smoothing parameter selection methods could be badly affected by outlying data. In this section, we develop three smoothing parameter selection procedures that are capable of handling such outliers. The first one is based on the cross-validation idea. It can be applied to any smoothing methods but it is computationally expensive. The last two procedures are much less computationally demanding, but can only be applied to the penalized smoothers (2). Although our presentation below is for

the case with one covariate, all three methods can be extended straightforwardly to select multiple smoothing parameters for multiple covariates. In general, we denote the estimate of μ_i computed using the smoothing parameter λ as $\hat{\mu}_{i\lambda}$.

5.1 ROBUST CROSS-VALIDATION

Cross-validation (Stone 1974) is a widely applicable method for choosing smoothing parameter. It uses the so-called "leave-one-out" strategy to approximate the best λ that minimizes the loss function under consideration. For the current problem, a natural loss function is the following Kullback-Leibler distance between the true and estimated μ_i's:

$$\mathrm{KL}(\lambda) = E\left\{\sum_{i=1}^{n} q(y_i, \mu_i)\right\} - E\left\{\sum_{i=1}^{n} q(y_i, \hat{\mu}_{i\lambda})\right\},$$

where q is the robust quasi-likelihood defined in Equation (7). As the first term is a constant with respect to λ, it can be ignored in the minimization. Denote the leave-one-out estimate of μ_i as $\hat{\mu}_{i\lambda}^{-i}$. The second term of $\mathrm{KL}(\lambda)$ can then be estimated by the following robust cross-validation (RCV) criterion

$$\mathrm{RCV}(\lambda) = -\frac{1}{n}\sum_{i=1}^{n} q\big(y_i, \hat{\mu}_{i\lambda}^{-i}\big), \tag{13}$$

and λ is chosen as its minimizer.

One shortcoming about this procedure is that it is computationally expensive. Although k-fold cross-validation can be applied to alleviate this problem, it could still be impractical when n and/or m (number of covariates) are large. Thus, we seek faster alternatives.

5.2 ROBUST INFORMATION CRITERIA

Generalized information criterion (GIC) was introduced by Konishi and Kitagawa (1996) for estimating the Kullback-Leibler distance between a true and a fitted model. It can be viewed as a generalization of the Akaike information criterion (AIC), as it relaxes the AIC's assumption that the model parameters are estimated with maximum likelihood.

Recall the basis functions for representing f are $b_1, \ldots, b_p$. Write $\mathbf{b}(x) = \{b_1(x), \ldots, b_p(x)\}^{\mathrm{T}}$ and $\mathbf{X} = \{\mathbf{b}(x_1), \ldots, \mathbf{b}(x_n)\}^{\mathrm{T}}$, and denote the conditional density $y_i|x_i$ as h. In the online Appendix, it is shown that, for the current problem, applying the GIC methodology will result in selecting λ as the minimizer of the following robust AIC (RAIC) formula:

$$\mathrm{RAIC}(\lambda) = -2\sum_{i=1}^{n} q(y_i, \hat{\mu}_{i\lambda}) + 2 \times \mathrm{tr}(\mathbf{P}^{-1}\mathbf{Q}), \tag{14}$$

where

$$\mathbf{P} = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{B}\mathbf{X} + \frac{1}{n}\mathbf{S} \quad \text{and} \quad \mathbf{Q} = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X} - a(\boldsymbol{\beta})a(\boldsymbol{\beta})^{\mathrm{T}}.$$

In the above $\mathbf{A}$ and $\mathbf{B}$ are diagonal matrices with elements, respectively,

$$a_i = E\left[\phi_c\left\{\frac{y_i - \hat{\mu}_{i\lambda}}{V^{\frac{1}{2}}(\hat{\mu}_{i\lambda})}\right\}^2\right]\left\{\frac{\zeta^2(\mu_i)}{V(\hat{\mu}_{i\lambda})}\right\}\left(\frac{\partial}{\partial\eta_i}\mu_i\Big|_{\mu_i=\hat{\mu}_{i\lambda}}\right)^2$$

and

$$b_i = E\left[\phi_c\left\{\frac{y_i - \mu_i}{V^{\frac{1}{2}}(\mu_i)}\right\}\frac{\partial}{\partial\mu_i}\log h(y_i|x_i,\mu_i)\right]\Bigg|_{\mu_i=\hat{\mu}_{i\lambda}}\left\{\frac{\zeta(\mu_i)}{V^{\frac{1}{2}}(\hat{\mu}_{i\lambda})}\right\}\left(\frac{\partial}{\partial\eta_i}\mu_i\Big|_{\mu_i=\hat{\mu}_{i\lambda}}\right)^2.$$

For many model selection problems, it has been observed that AIC tends to select over-parameterized models, and this issue may carry over to RAIC(λ). One common method to overcome this is to increase the penalty (e.g., see Bhansali and Downham 1977). Typically, the constant 2 in the penalty term is changed to $\log(n)$, which coincides with the penalty of the Bayesian information criterion (BIC). Following this practice we obtain our third criterion, robust BIC (RBIC), for selecting λ:

$$\text{RBIC}(\lambda) = -2\sum_{i=1}^{n} q(y_i, \hat{\mu}_{i\lambda}) + \log(n) \times \text{tr}(\mathbf{P}^{-1}\mathbf{Q}). \tag{15}$$

6. MULTIPLE COVARIATES

This section returns to the case when there is more than one covariate; that is, when $m > 1$. Recall that the goal is to estimate $f_1, \ldots, f_m$ in $\eta_i \equiv \sum_{j=1}^{m} f_j(x_{ji})$, where $\{x_{1i}, \ldots, x_{mi}\}$ are the observed covariate values. Since there is no interaction term, each f_j can be modeled independently, and we allow different f_j's to have different basis functions. Let the number of bases for f_j be p_j, and the bases be $\{b_1^{(j)}, \ldots, b_{p_j}^{(j)}\}$. Then we have the following representation for f_j:

$$f_j(x;\boldsymbol{\beta}_j) = \sum_{k=1}^{p_j} b_k^{(j)}(x)\beta_k^{(j)},$$

where $\boldsymbol{\beta}_j = (\beta_1^{(j)}, \ldots, \beta_{p_j}^{(j)})^{\mathrm{T}}$ are the basis coefficients. To keep the model identifiable, it is customary to impose the constraint that, except for f_1, all f_j's have zero mean. This constraint can be automatically achieved by applying a suitable transformation to the coefficients, basis matrix, and penalty matrix (see, e.g., Wood 2006, for details). Below we assume that this transformation has been applied.

Let λ_j and $\mathbf{D}_j$ be the smoothing parameter and penalty matrix, respectively, for f_j. Similarly to the case when $m = 1$, the robust estimate of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_m^{\mathrm{T}})^{\mathrm{T}}$ is defined as the maximizer of

$$\sum_{i=1}^{n} q(y_i, \mu_i) - \sum_{j=1}^{m} \lambda_j \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{D}_j\boldsymbol{\beta}_j. \tag{16}$$

As mentioned before, the proposed algorithm can be applied to approximate this maximizer. Also, if we let $\mathbf{S} = \text{diag}(2\lambda_1\mathbf{D}_1, \ldots, 2\lambda_m\mathbf{D}_m)$, we can re-express the above penalty term

$\sum_j \lambda_j \boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{D}_j \boldsymbol{\beta}_j$ as $\boldsymbol{\beta}^{\mathrm{T}} \mathbf{S} \boldsymbol{\beta}/2$, making Equation (16) in the same form as for the single covariate case. The robust smoothing parameter selection criteria RAIC(λ) and RBIC(λ) can then be straightforwardly applied.

7. SIMULATION STUDY

A simulation study was conducted to evaluate the practical performance of the proposed methodology. All together six different fitting procedures are compared. They are

1. rgamRAIC: the algorithm proposed in Section 3.2 with λ chosen by RAIC (Equation (14));
2. rgamRBIC: similar to rgamRAIC except that λ is chosen by RBIC (Equation (15));
3. rgamRCV: similar to rgamRAIC except that λ is chosen by RCV (Equation (13)); and
4. gamAIC: a nonrobust GAM fitting procedure available in the R package *mgcv* (Wood 2006) with λ chosen by AIC.
5. the method proposed by Croux, Gijbels, and Prosdocimi (2011) (codes downloaded from one of the authors' web site), and
6. the method proposed by Alimadad and Salibian-Barrera (2011) (codes available in the R package *rgam*).

For the first four fitting procedures, we used thin plate regression spline basis of order 2.

Two types of error distributions were considered: the binomial and the Poisson families. For the former the logit link was used while for the latter the log link was used. For the three robust procedures, we followed Cantoni and Ronchetti (2001) and set $c = 1.2$ for binomial and $c = 1.6$ for Poisson. We considered two univariate test functions:

$$t_1(x) = 4\cos\{2\pi(1-x)^2\} \quad \text{and} \quad t_2(x) = -10x^2 - 2x + 5, \quad t \in [0, 1].$$

A bivariate example will be given in Section 8. Three sample sizes were tested: $n = 100$, 200, and 500.

The noisy data were generated in the following manner. First a covariate value x was drawn from Uniform[0, 1]. Then the response y was simulated from the distribution under consideration with mean $g^{-1}\{t_k(x)\}$, $k = 1, 2$. Finally, $p100\%$ of the simulated (x, y)'s were randomly selected and changed to outliers in the following manner. For binomial data, y is set to 0 if the original value of y is 1, and vice versa. For Poisson data, y is set to the nearest integer to $yu_1^{u_2}$, where u_1 is generated from Uniform(2, 5) and u_2 is drawn randomly from $\{-1, 1\}$. Altogether, three values of p were tested: 0, 0.05, and 0.1.

The mean squared error (MSE) $\sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2/n$ was used to measure the quality of the estimates. The number of replicates used in each experimental configuration was 500.

Table 1. Averaged MSE values and standard errors (in parentheses) from the simulation setting with test function t_1, $n = 100$, $p = 0.05$. All numbers from the binomial setting were multiplied by 10^4, while numbers from the Poisson setting were multiplied by 10

Fitting method	Binomial	Poisson
rgamRAIC	87.7 (2.74)	95.8 (10.4)
rgamRBIC	88.2 (2.77)	103 (10.7)
rgamRCV	91.1 (2.64)	62.1 (8.82)
Croux et al.	129 (3.83)	76.3 (11.5)
Alimada & Salibian-Barrera	300 (18.1)	147.5 (37.0)

7.1 COMPARISON WITH THE METHOD BY ALIMADAD AND SALIBIAN-BARRERA (2011)

Since the fitting method by Alimadad and Salibian-Barrera (2011) employs brute-force cross-validation for smoothing parameter selection, it is computationally very slow and discouraged us from testing it with all the simulation settings described above. Instead, we tested for the simulation setting with test function t_1, $n = 100$, $p = 0.05$, for both binomial and Poisson cases. The resulting averaged MSE values, together with their estimated standard errors, are reported in Table 1. From this table, there is some empirical evidence that the method by Alimadad and Salibian-Barrera (2011) is inferior to other robust methods. Also, as this method is computationally very slow, it will not be considered further.

7.2 COMPARISON WITH THE REMAINING FIVE METHODS

In this section, the first five fitting methods listed at the beginning of this section are tested for all the simulation settings described above, with the exception that the computationally expensive method rgamRCV was only considered for $n = 100$. For each simulation setting, the averaged MSEs together with their estimated standard errors were computed and reported in Tables 2–5.

To facilitate comparison, except for rgamRCV, for all possible pairs of fitting procedures, we applied paired t-tests to test if the averaged MSEs are significantly different. The significance level was adjusted with Bonferroni's method and the overall family-wise error rate was 0.05. The fitting procedures were then ranked in the following manner. If the mean MSE value of a procedure is significantly less than the remaining two, it will be assigned a rank 1. If the mean MSE value of a procedure is significantly larger than one but less than the other one, it will then be assigned a rank 2, and similarly for rank 3. Procedures having nonsignificantly different mean MSE values will share the same averaged rank. The resulting ranks are also reported in Tables 2–5.

From these tables, one can see that no method is universally the best. One can also see that rgamRBIC never performed worse than any other methods in the contaminated cases. It also performed well when there was no contamination except for the Poisson family with test function t_1. For rgamRAIC, from the averaged ranks, it seems to be slightly superior to gamAIC but inferior to rgamRBIC. As for rgamRCV, it performed well in most cases and its results are comparable to those from rgamRBIC. However, its huge computational expenses significantly lower its practical values. Finally, we note that the method by Croux,

Table 2. Averaged MSE values ($\times 10^4$), standard errors ($\times 10^4$, in parentheses), and paired t-test rankings (in square brackets) from the simulation setting with test function t_1 and binomial data

		Fitting method								
p	n	gamAIC		rgamRAIC		rgamRBIC		rgamRCV	Croux et al.	
0	100	79 (3.56)	[2.5]	90.1 (5.45)	[2.5]	90.7 (5.45)	[2.5]	74.1 (3.05)	107 (3.7)	[2.5]
	200	36.2 (1.24)	[1]	45.1 (1.3)	[2]	46 (1.29)	[3]	—	59.6 (1.83)	[4]
	500	14.4 (0.442)	[2.5]	24.6 (0.466)	[2.5]	26.1 (0.424)	[2.5]	—	25.7 (0.653)	[2.5]
0.05	100	110 (2.81)	[3]	87.7 (2.74)	[1.5]	88.2 (2.77)	[1.5]	91.1 (2.64)	129 (3.83)	[4]
	200	68.1 (1.44)	[3]	56.8 (1.38)	[1.5]	57.2 (1.39)	[1.5]	—	84.3 (2.18)	[4]
	500	41.9 (0.688)	[2.5]	40.2 (0.719)	[1]	40.8 (0.707)	[2.5]	—	49.5 (0.886)	[4]
0.1	100	182 (3.42)	[3.5]	140 (12.4)	[1.5]	140 (12.4)	[1.5]	149 (3.21)	186 (4.25)	[3.5]
	200	139 (2.01)	[3.5]	102 (2.52)	[1.5]	101 (2.57)	[1.5]	—	145 (2.46)	[3.5]
	500	104 (1.11)	[3.5]	86.5 (1.33)	[2]	84.6 (1.37)	[1]	—	106 (1.19)	[3.5]
Averaged rank			[2.78]		[1.78]		[1.94]			[3.5]

Table 3. Averaged MSE values ($\times 10^4$), standard errors ($\times 10^4$, in parentheses), and paired t-test rankings (in square brackets) from the simulation setting with test function t_2 and binomial data

		Fitting method								
p	n	gamAIC		rgamRAIC		rgamRBIC		rgamRCV	Croux et al.	
0	100	44.3 (2.3)	[2.5]	39.3 (2.01)	[2.5]	38.7 (2.01)	[2.5]	40.6 (1.86)	49.8 (2.94)	[2.5]
	200	18.2 (0.94)	[2.5]	18 (0.745)	[2.5]	17.6 (0.732)	[2.5]	—	21.7 (1.15)	[2.5]
	500	6.89 (0.344)	[2.5]	7.43 (0.337)	[2.5]	7.1 (0.311)	[2.5]	—	9.69 (0.436)	[2.5]
0.05	100	72.6 (2.72)	[2.5]	62.7 (2.03)	[2.5]	60.9 (1.99)	[2.5]	67.9 (2.13)	68.8 (2.74)	[2.5]
	200	45.1 (1.16)	[3.5]	39.2 (1.13)	[2]	37.1 (1.05)	[1]	—	45.9 (1.34)	[3.5]
	500	31.7 (0.517)	[3.5]	25.7 (0.553)	[2]	24.8 (0.529)	[1]	—	32.3 (0.623)	[3.5]
0.1	100	135 (2.89)	[3.5]	114 (2.62)	[2]	112 (2.62)	[1]	121 (2.65)	135 (3.46)	[3.5]
	200	110 (1.8)	[3.5]	94.2 (1.74)	[2]	92.6 (1.69)	[1]	—	109 (1.94)	[3.5]
	500	90.1 (0.894)	[3.5]	81 (0.937)	[2]	79.6 (0.922)	[1]	—	88.5 (0.901)	[3.5]
Averaged rank			[3.06]		[2.22]		[1.67]			[3.06]

Table 4. Averaged MSE values ($\times 10$), standard errors ($\times 10$, in parentheses), and paired t-test rankings (in square brackets) from the simulation setting with test function t_1 and Poisson data

		Fitting method								
p	n	gamAIC		rgamRAIC		rgamRBIC		rgamRCV	Croux et al.	
0	100	29.5 (0.761)	[1]	84 (8.22)	[3.5]	82.9 (7.87)	[3.5]	29.4 (0.767)	38.3 (0.945)	[2]
	200	16 (0.371)	[1]	19.8 (1.41)	[3]	20.1 (0.934)	[3]	—	21.7 (0.439)	[3]
	500	6.72 (0.146)	[2.5]	7.98 (0.547)	[2.5]	9.05 (0.602)	[2.5]	—	9.23 (0.18)	[2.5]
0.05	100	312 (20.5)	[4]	95.8 (10.4)	[2]	103 (10.7)	[2]	62.1 (8.82)	76.3 (11.5)	[2]
	200	165 (9.82)	[2.5]	31.9 (4.17)	[2.5]	33.5 (4.18)	[2.5]	—	36.7 (4.54)	[2.5]
	500	76.1 (2.84)	[4]	12.1 (0.918)	[2]	13.5 (1)	[2]	—	11.4 (0.256)	[2]
0.1	100	631 (30.8)	[4]	171 (25.1)	[2]	181 (24.8)	[2]	112 (17.6)	126 (16.9)	[2]
	200	363 (16.9)	[4]	50.7 (7.37)	[2]	51.5 (7.37)	[2]	—	52.5 (7.42)	[2]
	500	182 (6.09)	[4]	18.1 (1.37)	[2]	18.5 (1.34)	[2]	—	14.8 (0.759)	[2]
Averaged rank			[3]		[2.39]		[2.39]			[2.22]

Table 5. Averaged MSE values ($\times 10$), standard errors ($\times 10$, in parentheses), and paired t-test rankings (in square brackets) from the simulation setting with test function t_2 and Poisson data

		Fitting method								
p	n	gamAIC		rgamRAIC		rgamRBIC		rgamRCV	Croux et al.	
0	100	23.3 (1.01)	[1.5]	24.8 (1.1)	[3]	22.1 (0.994)	[1.5]	27.1 (1.2)	42.2 (1.47)	[4]
	200	11.1 (0.485)	[2.5]	11.6 (0.484)	[2.5]	10.3 (0.435)	[1]	—	21.3 (0.685)	[4]
	500	4.19 (0.193)	[1.5]	4.57 (0.209)	[3]	4.07 (0.198)	[1.5]	—	9.35 (0.296)	[4]
0.05	100	804 (77.6)	[4]	32.2 (2.66)	[2]	25.6 (1.19)	[1]	32.5 (2.13)	48.6 (2.55)	[3]
	200	573 (58.2)	[4]	15.9 (0.812)	[2]	13.9 (0.689)	[1]	—	27.5 (1.05)	[3]
	500	246 (14.8)	[4]	5.54 (0.236)	[2]	4.87 (0.214)	[1]	—	11.2 (0.335)	[3]
0.1	100	1930 (151)	[4]	121 (50.3)	[2]	75.7 (38.1)	[2]	44.2 (2.17)	127 (42)	[2]
	200	991 (71.1)	[4]	22.6 (2.53)	[2]	16.7 (0.933)	[1]	—	35.1 (3.65)	[3]
	500	562 (24.5)	[4]	7.62 (0.335)	[2]	6.54 (0.297)	[1]	—	14.1 (0.444)	[3]
Averaged rank			[3.28]		[2.28]		[1.22]			[3.22]

Gijbels, and Prosdocimi (2011) also estimates the dispersion function, so the comparison here may not be entirely fair.

8. REAL DATA EXAMPLE

Here we apply our methodology to analyze a two-covariate dataset that originated from a study conducted by the Deutsche Forschungsgemeinschaft (German research foundation). It was collected during the years 1960–1977 in a mechanical engineering plant in Munich, Germany. The aim is to study the relationship between chronic bronchitis and dust concentration (for more details, see e.g., Kuchenhoff and Carroll 1997; Kauermann and Opsomer 2004).

The dataset contains records of 1246 workers. The response `cbr` is binary: occurrence of chronic bronchitis (`cbr` $=$ `1` for yes, `cbr` $=$ `0` for no). The covariates are `dust`, dust concentration in mg/m^3, and `expo`, duration of exposure in years. This dataset is plotted in Figure 1. A reasonable model is

$$g\{E(\texttt{cbr})\} = f_1(\texttt{dust}) + f_2(\texttt{expo}),$$

where f_1 and f_2 are smooth functions and g is the logit link.

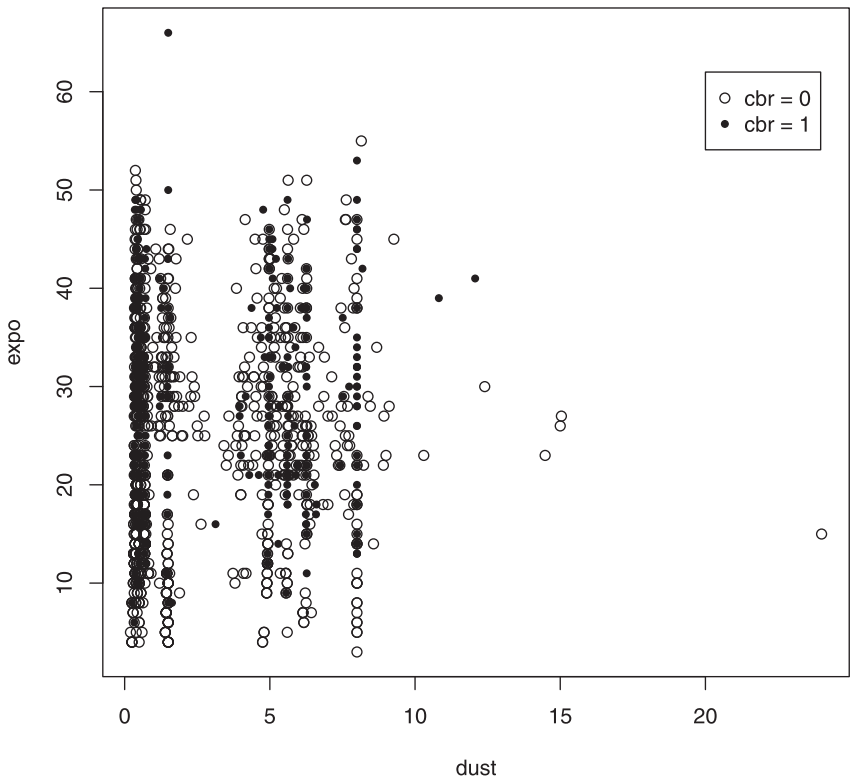

Figure 1. The bronchitis dataset.

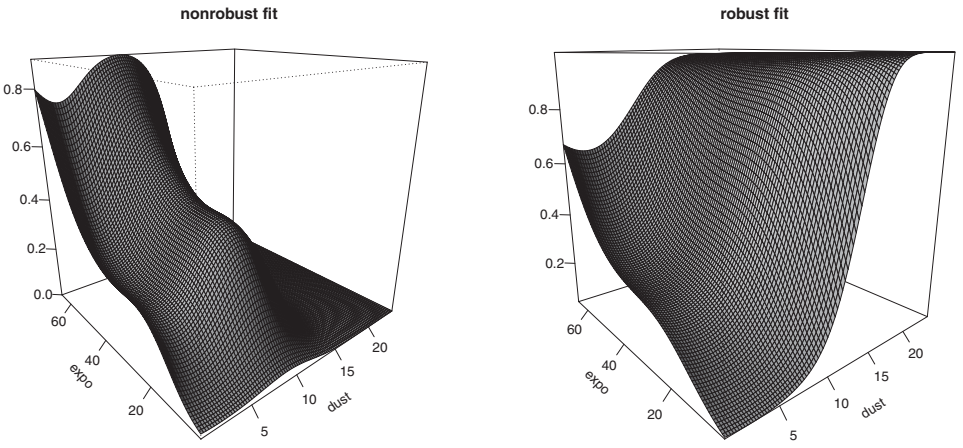

Figure 2. Fitted surfaces for the bronchitis dataset. Left: nonrobust fit. Right: robust fit.

A quick inspection of Figure 1 reveals a few potential high-leverage observations (those with `dust` > 13). These high-leverage observations may induce undesirable effects on our estimation, and the idea discussed by Cantoni and Ronchetti (2001) can be used to reduce such effects. We follow this idea and modify the robust score function (5) by replacing $\zeta(\mu_i)$ with $\zeta^*(\mu_i, \mathbf{x}_i) = \zeta(\mu_i)\xi(\mathbf{x}_i)$, where ξ is chosen to down-weigh those high-leverage observations. We used $\xi(\mathbf{x}_i) = \{1 + (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\mathbf{x}}^{\mathrm{T}}\hat{\mathbf{S}}_{\mathbf{x}}\hat{\boldsymbol{\mu}}_{\mathbf{x}})\}^{(-1/2)}$ where $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{S}}_{\mathbf{x}}$ are robust estimates of the mean and variance of $\mathbf{x}_i$'s, respectively (for other choices of $\xi(\mathbf{x}_i)$, see Rousseeuw and Leroy 1987, p. 258).

We applied the proposed robust fitting method rgamRBIC to estimate f_1 and f_2. For comparative purposes, we also estimated f_1 and f_2 with gamAIC (i.e., nonrobust fitting). The choice of basis functions and other user-specific parameters such as knot locations is the same as those used in Section 7. The resulting fitted functions are displayed in Figure 2.

The left panel of Figure 2 shows a counter-intuitive phenomenon in the nonrobust fit: it seems to suggest that the higher the dust concentration, the lower the chance of contracting chronic bronchitis. By inspecting Figure 1, this counter-intuitive phenomenon is most likely due to the four observations with `dust` > 13. For the proposed robust fitting method, however, the effects of these four observations have been down-weighted. The corresponding fitted surface does provide a reasonable qualitative conclusion: the chance of contracting chronic bronchitis increases with both `expo` and `dust`.

9. CONCLUDING REMARKS

The methodology proposed in this article provides automatic methods for fitting GAMs in the presence of high-leverage points and outliers. It contains three main ingredients: the use of robust estimating equations to define robust estimates, a practical algorithm for calculating these estimates, and three new selection methods for choosing the smoothing parameter. Overall, rgamRBIC is the recommended default procedure if estimation of the dispersion function is not needed. It is relatively fast, backed up with theoretical justification

for equivalence results, and gave promising empirical performance in both simulations and real data analysis. *R* codes implementing rgamRBIC can be obtained from the authors.

SUPPLEMENTARY MATERIALS

R-package `robustgam` **implementing the proposed methods**: can be obtained at the official R web site: *http://cran.r-project.org/web/packages/robustgam/*

R scripts for repeating the numerical experiments: file name: `sim-code.zip`. The main file is `sim.R`. It also contains codes for the method by Croux, Gijbels, and Prosdocimi (2011).

Online Appendix: contains technical details of the article.

ACKNOWLEDGMENTS

The authors are most grateful to the referees and the associate editor for their constructive comments. The work of Yao was partially supported by NSERC Individual Discovery and Discovery Accelerator Supplement grants. The work of Lee was partially supported by the National Science Foundation grants 1007520, 1209226, and 1209232. This article was accepted prior to Thomas Lee being named Editor of *JCGS*.

[Received November 2011. Revised November 2012.]

REFERENCES

Alimadad, A., and Salibian-Barrera, M. (2011), "An Outlier-Robust Fit for Generalized Additive Models With Applications to Disease Outbreak Detection," *Journal of the American Statistical Association*, 106, 719–731. [271,280]

Bhansali, R. J., and Downham, D. Y. (1977), "Some Properties of the Order of an Autoregressive Model Selected by a Generalization of Akaike's FPE Criterion," *Biometrika*, 64, 547–551. [279]

Cantoni, E., and Ronchetti, E. (2001), "Robust Inference for Generalized Linear Models," *Journal of the American Statistical Association*, 96, 1022–1030. [271,280,287]

Carroll, R. J., and Pederson, S. (1993), "On Robustness in the Logistic Regression Model," *Journal of the Royal Statistical Society,* Series B, 55, 693–706. [270]

Copas, J. B. (1988), "Binary Regression Models for Contaminated Data," *Journal of the Royal Statistical Society,* Series B, 50, 225–265. [270]

Croux, C., Gijbels, I., and Prosdocimi, I. (2011), "Robust Estimation of Mean and Dispersion Functions in Extended Generalized Additive Models," *Biometrics*, 68, 31–44. [271,280,286,288]

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, London: Chapman & Hall. [272]

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley. [272]

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman & Hall. [270]

Huber, P. J. (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799–821. [277]

Kauermann, G., and Opsomer, J. D. (2004), "Generalized Cross-Validation for Bandwidth Selection of Backfitting Estimates in Generalized Additive Models," *Journal of Computational and Graphical Statistics*, 13, 66–89. [286]

Konishi, S., and Kitagawa, G. (1996), "Generalised Information Criteria in Model Selection," *Biometrika*, 83, 875–890. [271,278]

Kuchenhoff, H., and Carroll, R. J. (1997), "Segmented Regression With Errors in Predictors: Semi-Parametric and Parametric Methods," *Statistics in Medicine*, 16, 169–188. [286]

Künch, 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. [271]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall. [270]

Morgenthaler, S. (1992), "Least-Absolute-Deviations Fits for Generalized Linear Models," *Biometrika*, 79, 747–754. [271]

Nychka, D. (1995), "Splines as Local Smoothers," *The Annals of Statistics*, 23, 1175–1197. [277]

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. [275]

Preisser, J. S., and Qaqish, B. F. (1999), "Robust Regression for Clustered Data With Applications to Binary Responses," *Biometrics*, 55, 574–579. [271]

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley. [287]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge: Cambridge University Press. [272]

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, 413–424. [271]

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society,* Series B, 36, 111–147. [278]

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia, PA: Society for Industrial and Applied Mathematics. [276]

Wood, S. N. (2004), "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models," *Journal of the American Statistical Association*, 99, 673–686. [277]

——— (2006), *Generalized Additive Models: An Introduction With R*, Boca Raton, FL: Chapman & Hall. [279,280]

——— (2008), "Fast Stable Direct Fitting and Smoothness Selection for Generalized Additive Models," *Journal of the Royal Statistical Society,* Series B, 70, 495–518. [277]