Robust Estimation for Generalized Additive Models

Raymond K. W. Wong* Fang Yao† Thomas C. M. Lee‡

November 20, 2011; revised: April 22, 2012

Abstract

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.

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

*Department of Statistics, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA, email: rkwwong@ucdavis.edu

†Department of Statistics, University of Toronto, 100 St. George Street, Toronto, Ontario M5S 3G3 Canada, email: fyao@utstat.toronto.edu.

‡Department of Statistics, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA, email: tcmlee@ucdavis.edu

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 general settings, Stefanski *et al.* (1986) and Künch *et al.* (1989) propose using bounded score functions to define robust estimates, Morgenthaler (1992) uses L_1 norm for likelihood calculations, and Preisser and Qaqish (1999) and Cantoni and Ronchetti (2001) construct robust estimating equations for conducting, respectively, robust estimation and robust inference procedures. For the robust estimation of GAMs, two recent papers are devoted to the subject: Alimadad and Salibian-Barrera (2011) and Croux *et al.* (2011). The estimation procedures developed in these two papers produce promising empirical results. However, they also have some minor shortcomings: the procedure of 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 of Croux *et al.* (2011).

Following the idea of Stefanski *et al.* (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 of 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 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, (1) reduces to $\eta_i = f(x_i)$.

One common nonparametric approach to estimating f is penalized basis expansion fitting. With a set of pre-specified 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 pre-specified 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 $\check{\boldsymbol{\beta}}$, is the solution of (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 *et al.*, 2003).

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

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, for example, 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 $\check{\boldsymbol{\beta}}$, we first note that $\check{\boldsymbol{\beta}}$ is an M-estimator defined by the score function

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

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 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

In order to achieve robust estimation for GAMs, one could modify the estimating equations (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 (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

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

In order 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:

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

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 numerically 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

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

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 (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 (5) by iteratively solving (3), taking the advantage that many fast methods and softwares are available for the solving of (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

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

Also define $\tilde{\boldsymbol{\beta}}$ as the solution to (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 (9).

Of course in practice $\tilde{y}_i$'s are unknown, but the above discussion suggests a fast iterative method for solving (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 (8), then plug in these new $\tilde{y}_i$'s into (9) and solve for the next set of estimates of μ_i's.

Many common GAM fitting methods, such as local scoring and iterative re-weighted least-squares, for solving (3) are iterative, with each iteration effectively as a weighted AM fitting. This means a direct application of the above idea for solving (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,
$$\tilde{z}_i^{(t+1)} = (\tilde{y}_i^{(t)} - \hat{\mu}_i^{(t)}) g'(\hat{\mu}_i^{(t)}) + \hat{\eta}_i^{(t)},$$
where
$$\tilde{y}_i^{(t)} = \left[\nu(y_i, \hat{\mu}_i^{(t)}) - E\left\{\nu(y_i, \hat{\mu}_i^{(t)})\right\}\right] \zeta(\mu_i^{(t)}) V(\hat{\mu}_i^{(t)}) + \hat{\mu}_i^{(t)}$$
and
$$\hat{\eta}_i^{(t)} = g(\hat{\mu}_i^{(t)}).$$

2. Fit a weighted additive model 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 (3); i.e., 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. Lastly, for problems with normal errors and identity link function, $\tilde{y}_i$ in (8) recovers the pseudo data derived by Oh *et al.* (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 (9) while $\hat{\boldsymbol{\beta}}$ is the solution to (5). Denote the corresponding estimates for f derived from $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ through (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 below 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 (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 $\mathbf{f}_w = (f_{w,1}, \ldots, f_{w,n})^{\mathrm{T}}$; i.e., $\mathbf{f}_w = \mathbf{W}^{1/2}\mathbf{f}$ 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

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

where $J(\mathbf{f}) = \mathbf{f}^{\mathrm{T}}\mathbf{R}^*\mathbf{f} = \mathbf{f}_w^{\mathrm{T}}\mathbf{R}\mathbf{f}_w$ is a reproducing kernel Hilbert space 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 $\rho(z_{w,i} - f_{w,i}) = [\phi(r_i) - E\{\phi(r_i)\}]\zeta(\mu_i)$. Now as we have shifted our focus from $\boldsymbol{\beta}$ to f, the score functions $\check{\boldsymbol{\psi}}(y_i, \boldsymbol{\beta})$ in (4) and $\boldsymbol{\psi}(y_i, \boldsymbol{\beta})$ in (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, i.e., $\sup_{-\infty<t<\infty}|f(t)|<\infty$.

($\mathcal{A}$.2) Assume that $\max_{1\le i\le n}\mathrm{var}\{\phi(r_i)\}<\infty$ 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 in Huber (1973), higher order derivatives are technically convenient, but hardly essential for the results to hold. It can be easily fulfilled by modify ϕ_c in (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 $\mathrm{tr}\{\mathbf{H}(\lambda_n)\}/\lambda_n<K_0$ 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 Nychka (1995) is actually equal to λ_n/n in this paper. In particular, ($\mathcal{A}$.3.b) involves balancing the rates of the smoothing parameter with the effective degrees of freedom, $\mathrm{tr}\{\mathbf{H}(\lambda_n)\}$, of the smoother. Based on the equivalent kernel theory in Nychka (1995), one expects that $\mathrm{tr}\{\mathbf{H}(\lambda_n)\}\sim(\lambda_n/n)^{-1/2m}$, 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 $0<\kappa\le 2m/(2m+1)$, 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 reproducing kernel Hilbert space. Let $\mathcal{C}=\{f\in\mathcal{H}:\|f\|_{\mathcal{H}}\le G\}$, 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 (A.1)–(A.4) hold, then a consistent robust estimator* $\hat{f}$ *exists in a neighborhood of* f *in* $\mathcal{C}$ *and* $C_n = E\{\|\tilde{\mathbf{f}} - \mathbf{f}\|_n^2\} \to 0$ *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 Appendix A.

5 Smoothing Parameter Selection

For nonrobust GAMs estimation, Wood (2004, 2008) has 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 (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

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

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. Appendix B shows 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

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

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; i.e., 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. In 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 $\mathbf{S} = \mathrm{diag}(2\lambda_1 \mathbf{D}_1, \ldots, 2\lambda_m \mathbf{D}_m)$, 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 (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 (14);
2. rgamRBIC: similar to rgamRAIC except λ is chosen by RBIC (15);
3. rgamRCV: similar to rgamRAIC except λ is chosen by RCV (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 *et al.* (2011), and
6. the method proposed by Alimadad and Salibian-Barrera (2011).

For the first four fitting procedures, we used the same radial basis of order 2 and 30 knots.[**Thomas: Raymond, please confirm.**] For $k = 1, \ldots, 30$, the knots were placed at the $(k/30)$th quantiles of the x_i's, with the smallest and the largest values removed.

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:

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

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 $g^{-1}\{t_k(x)\}$, $k = 1, 2$. Lastly, $p100\%$ 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.

7.1 Comparison with the Method of Alimadad and Salibian-Barrera (2011)

We obtained the code for the fitting method of Alimadad and Salibian-Barrera (2011) from one of the authors' website.[**Thomas: Raymond, please confirm.**] Since this method 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, $n = 100$,

Table 1: Averaged MSE values ($\times 10^4$) and standard errors ($\times 10^5$, in parentheses) from the simulation setting with Test Function t_1, $n = 100$, $p = 0.05$.

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	182 (2.02)	225 (17.8)

$p = 0.05$, 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 strong empirical evidence that the method of Alimadad and Salibian-Barrera (2011) is inferior to other robust methods. Consequently, this method will not be considered further.

7.2 Comparison with the Remaining Five Methods

In this subsection 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 rgamRCVwas 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 to 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 non-significantly different mean MSE values will share the same averaged rank. The resulting ranks are also reported in Tables 2 to 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. Lastly, we note that the method of Croux *et al.* (2011) also estimates the dispersion function, so the comparison here may not be entirely fair.

Table 2: Averaged MSE values ($\times 10^4$), standard errors ($\times 10^5$, 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: Similar to Table 2 but for Test Function t_2.

		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]

8 Real Data Example

Here we apply our methodology to analyze a two-covariate data set originated from a study conducted by the Deutsche Forschungsgemeinschaft (German research foundation). It was collected during the years 1960 to 1977 in a mechanical engineering plant in Munich, Germany. The aim is to study the relationship between chronic bronchitis and dust concen-

Table 5: Similar to Table 4 but for Test Function t_2.

		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]

tration. For more details, see for examples Kuchenhoff and Carroll (1997) and Kauermann and Opsomer (2004).

The data set contains records of 1,246 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 data set is plotted in Figure 1. A reasonable model is

$$g\{E(\texttt{cbr})\} = \texttt{f}_1(\texttt{dust}) + \texttt{f}_2(\texttt{expo}),$$

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

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, pp. 258).

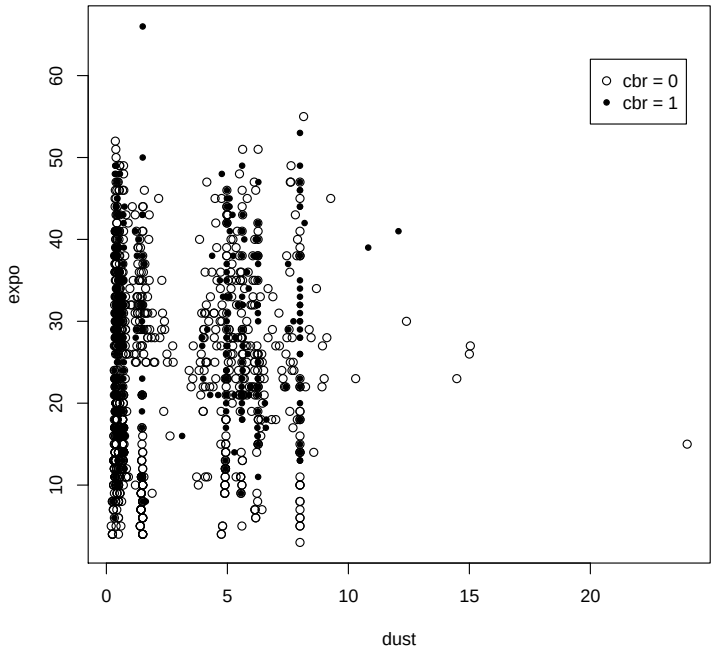

Figure 1: The Bronchitis data set.

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 are 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 4 observations with `dust` $>$ `13`. For the proposed robust fitting method, however, the effects of these 4 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`.

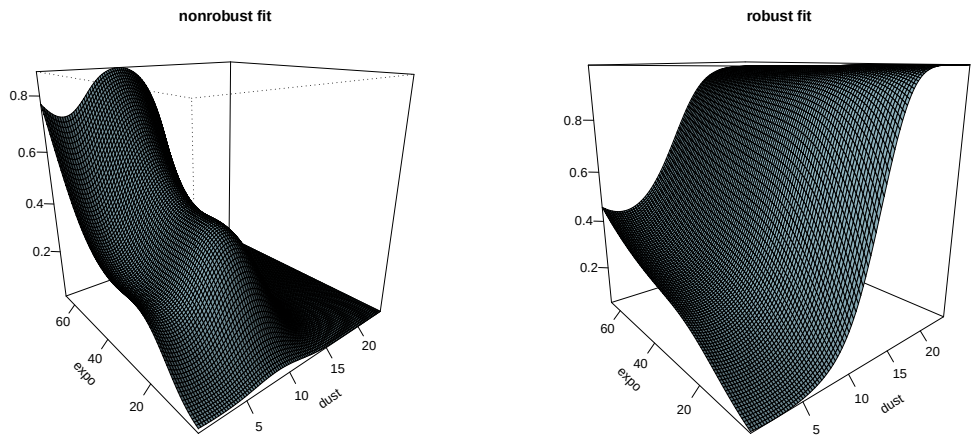

Figure 2: Fitted surfaces for the Bronchitis data set. Left: nonrobust fit. Right: robust fit.

9 Concluding Remarks

The methodology proposed in this paper 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.

Acknowledgement

The authors are most grateful to the referees and the associate editor for their constructive comments.

A Auxiliary Lemmas and Proofs

The proof of Theorem 1 partially follows the idea in Oh *et al.* (2007), with substantial changes made for the GAM framework. It is worth mentioning that the basic probability inequalities in Lemma A2 are required to hold uniformly over all neighborhoods of f. Major effort was devoted to prove the stronger uniform result. In fact the same technique used in this paper can be adopted to fix the proof of Oh *et al.* (2007) without altering the conclusion.

For convenience we abuse notation so that f is used to denote both the unknown function and the vector of function values sampled at the x_i's. Confusion should not arise as the correct interpretation can be clearly determined from the context. Similar abuses also exist for other symbols, such as $\hat{f}$ and $\tilde{f}_w$. We first present three lemmas and recall that $\tilde{f}_w$ is the penalized least squares estimate of f_w obtained by applying the smoothing matrix $\mathbf{H}(\lambda)$ to $\tilde{z}_{w,i}$'s.

Lemma A1. *(Consistency of roughness penalty and C_n). There exists $K_1 > 0$ such that*

$$E\left\{J(\tilde{f})\right\} \leq J(f) + \frac{K_1 \mathrm{tr}\left\{\mathbf{H}(\lambda)\right\}}{\lambda}$$

and

$$\frac{E\left\{\|\tilde{f}_w - f_w\|^2\right\}}{2\lambda} \leq J(f) + \frac{K_1 \mathrm{tr}\{\mathbf{H}(\lambda)\}}{\lambda},$$

which also implies that $E\{\|\tilde{f} - f\|_n^2\} \to 0$ as $n \to \infty$.

Proof of Lemma A1. It is sufficient to show that $\mathrm{var}(\tilde{z}_{w,i}) < K_1$ for all $i = 1, \ldots, n$, and the remaining arguments used in Oh *et al.* (2007) are valid. The assumption that $|f_i|$ is bounded from above implies that μ_i is bounded away from the singularities of $1/V^{1/2}(\cdot)$, including $\pm\infty$, and thus w_{ii}'s are uniformly bounded. From the relationship $z_{w,i} - f_{w,i} = (y_i - \mu_i)/V^{1/2}(\mu_i) = r_i$ and by the definition of ρ, it is easy to check $E\{\rho(z_{w,i} - f_{w,i})\} = 0$ and $\mathrm{var}\{\rho(z_{w,i} - f_{w,i})\} < \infty$ due to $(\mathcal{A}.1)$. By further noting that $\tilde{z}_{w,i} - f_{w,i} = \rho(z_{w,i} - f_{w,i})$, it follows that $E(\tilde{z}_{w,i}) = f_{w,i}$ and there exists K_1 such that $\mathrm{var}(\tilde{z}_{w,i}) < K_1$.

As $\tilde{f}$ is the penalized least squares estimator,

$$\begin{aligned} & \sum_{i=1}^{n}(\tilde{z}_{w,i}-\tilde{f}_{w,i})^2+2\lambda J(\tilde{f}) \;\leq\; \sum_{i=1}^{n}(\tilde{z}_{w,i}-f_{w,i})^2+2\lambda J(f) \\ \Rightarrow \quad & \|\{\mathbf{I}-\mathbf{H}(\lambda)\}\,\tilde{\mathbf{z}}_w\|^2+2\lambda J(\tilde{f}) \;\leq\; \|\tilde{\mathbf{z}}_w-f_w\|^2+2\lambda J(f). \end{aligned}$$

Taking expectation on both sides, we have

$$\begin{aligned} & \mathrm{tr}\left[\{\mathbf{I}-\mathbf{H}(\lambda)\}^2\,\mathrm{var}(\tilde{\mathbf{z}}_w)\right]+\|\{\mathbf{I}-\mathbf{H}(\lambda)\}\,f_w\|^2+2\lambda E\left\{J(\tilde{f})\right\} \\ & \qquad\qquad \leq \mathrm{tr}\left\{\mathrm{var}(\tilde{\mathbf{z}}_w)\right\}+2\lambda J(f) \\ \Rightarrow \quad & \mathrm{tr}\left\{\mathbf{H}^2(\lambda)\mathrm{var}(\tilde{\mathbf{z}}_w)\right\}+\|\{\mathbf{I}-\mathbf{H}(\lambda)\}\,f_w\|^2+2\lambda E\left\{J(\tilde{f})\right\} \\ & \qquad\qquad \leq 2\mathrm{tr}\left\{\mathbf{H}(\lambda)\mathrm{var}(\tilde{\mathbf{z}}_w)\right\}+2\lambda J(f) \\ \Rightarrow \quad & \mathrm{tr}\left\{\mathbf{H}^2(\lambda)\mathrm{var}(\tilde{\mathbf{z}}_w)\right\}+\|\{\mathbf{I}-\mathbf{H}(\lambda)\}\,f_w\|^2+2\lambda E\left\{J(\tilde{f})\right\} \\ & \qquad\qquad \leq 2K_1\mathrm{tr}\left\{\mathbf{H}(\lambda)\right\}+2\lambda J(f). \end{aligned}$$

By omitting the first and the second term on the left hand side, we prove the first inequality. Next writing $\|\tilde{f}_w-f_w\|^2=\|(\tilde{f}_w-\mathbf{H}(\lambda)f_w)-\{\mathbf{I}-\mathbf{H}(\lambda)\}f_w\|^2$ and taking expectations after expanding the r.h.s. lead to $E\left\{\|\tilde{f}_w-f_w\|^2\right\}=\mathrm{tr}\left\{\mathbf{H}^2(\lambda)\mathrm{var}(\tilde{\mathbf{z}}_w)\right\}+\|\{\mathbf{I}-\mathbf{H}(\lambda)\}\,f_w\|^2$, then the second inequality follows. Therefore, given ($\mathcal{A}$.1), ($\mathcal{A}$.3) and ($\mathcal{A}$.4) and some constant K_{11}, we arrive at

$$E\{\|\tilde{f}-f\|_n^2\}\leq 2K_{11}\left[J(f)+\frac{K_1\mathrm{tr}\{\mathbf{H}(\lambda_n)\}}{\lambda_n}\right]\frac{\lambda_n}{n}\longrightarrow 0.$$

Lemma A2. *(Score function approximation). Let $\mathcal{F}_L=\{h\in\mathcal{C}:\|h-f\|_n\leq L\}$ for $L>0$. For any $\epsilon>0$, $L_0>0$, there exists an N such that for $n>N$,*

$$\Pr\left[\sup_{h\in\mathcal{F}_L}\left\|\mathbf{W}^{-1}\mathbf{H}(\lambda)\left\{\boldsymbol{\psi}(h_w;\mathbf{z}_w)-\check{\boldsymbol{\psi}}(h_w;\tilde{\mathbf{z}}_w)\right\}\right\|_n>\epsilon L\right]<\epsilon, \tag{17}$$

$$\Pr\left[\sup_{h\in\mathcal{F}_L}\left\|\mathbf{W}^{-1}\mathbf{H}(\lambda)^{\frac{1}{2}}\left\{\boldsymbol{\psi}(h_w;\mathbf{z}_w)-\check{\boldsymbol{\psi}}(h_w;\tilde{\mathbf{z}}_w)\right\}\right\|_n>\epsilon L\right]<\epsilon, \tag{18}$$

hold uniformly for $0<L\leq L_0$.

Proof of Lemma A2. Write $h_{w,i} = w_{ii}^{1/2} h_i$ and $f_{w,i} = w_{ii}^{1/2} f_i$. Denote

$$\check{\boldsymbol{\psi}}(h_w; \tilde{\mathbf{z}}_w)_i - \boldsymbol{\psi}(h_w; \mathbf{z}_w)_i = w_{ii}^{1/2} \{(\tilde{z}_{w,i} - h_{w,i}) - \rho(z_{w,i} - h_{w,i})\} \equiv g(-h_{w,i}).$$

and note that $g(-f_{w,i}) = w_{ii}^{1/2} \{(\tilde{z}_{w,i} - f_{w,i}) - \rho(z_{w,i} - f_{w,i})\} = 0$. Applying Taylor's theorem to expand $g(-h_{w,i})$ around $(-f_{w,i})$ with the remaining term in an integral form,

$$\begin{aligned}
g(-h_{w,i}) &= g(-f_{w,i}) + (f_{w,i} - h_{w,i}) g'(-f_{w,i}) + (f_{w,i} - h_{w,i})^2 \\
&\quad \times \int_0^1 \int_0^1 t g''\{-h_{w,i} + ts(f_{w,i} - h_{w,i})\} ds dt \\
&= w_{ii}^{\frac{1}{2}} \Big[\left\{\rho'(r_i) - 1\right\} (f_{w,i} - h_{w,i}) + (f_{w,i} - h_{w,i})^2 \\
&\quad \times \int_0^1 \int_0^1 t \rho''(r_i + ts(f_{w,i} - h_{w,i})) ds dt \Big] \\
&= w_{ii} \left\{\rho'(r_i) - 1\right\} (f_i - h_i) + w_{ii}^{\frac{3}{2}} (f_i - h_i)^2 \\
&\quad \times \int_0^1 \int_0^1 t \rho''(r_i + ts w_{ii}^{\frac{1}{2}} (f_i - h_i)) ds dt \\
&\equiv w_{ii} u_{1,i} + w_{ii}^{\frac{3}{2}} u_{2,i}.
\end{aligned}$$

Then $\mathbf{W}^{-1}\mathbf{H}(\lambda)\{\check{\boldsymbol{\psi}}(h_w; \tilde{\mathbf{z}}_w) - \boldsymbol{\psi}(h_w; \mathbf{z}_w)\} = \mathbf{H}(\lambda)\mathbf{u}_1 + \mathbf{W}^{1/2}\mathbf{H}(\lambda)\mathbf{u}_2$, and let $T_1(h) = \|\mathbf{H}(\lambda)\mathbf{u}_1\|_n$ and $T_2(h) = \|\mathbf{W}^{1/2}\mathbf{H}(\lambda)\mathbf{u}_2\|_n$.

We first consider $E\{T_1^2(h)\}$. It is easy to see that w_{ii}, $\zeta(\mu_i)$ and thus $\mathrm{var}\{\rho'(r_i)\} = \mathrm{var}\{\phi'(r_i)\}\zeta(\mu_i)^2$ are uniformly bounded given $(\mathcal{A}.1)$ and $(\mathcal{A}.2)$, and that the eigenvalues of $\mathbf{H}(\lambda)$ are always between 0 and 1 implying $\mathbf{H}^2(\lambda)_{ii} < \mathbf{H}(\lambda)_{ii}$ and $\mathbf{a}^{\mathrm{T}}\mathbf{H}^2(\lambda)\mathbf{a} \leq \mathbf{a}^{\mathrm{T}}\mathbf{a}$ for any $\mathbf{a} \in \Re^n$. Hence we have, for some $K_{21} > 0$, noting $E\rho'(r_i) = E\phi'(r_i)\zeta(\mu)_i = 1$ by the definition $\zeta(\mu_i) = 1/E\phi'(r_i)$,

$$\begin{aligned}
nE\{T_1^2(h)\} = E\{\mathbf{u}_1^{\mathrm{T}}\mathbf{H}^2(\lambda)\mathbf{u}_1\} = \mathrm{tr}\big\{\mathrm{cov}(\mathbf{u}_1, \mathbf{u}_1)\mathbf{H}^2(\lambda)\big\} \\
\leq K_{21}\Big\{\sum_{i=1}^n \mathbf{H}(\lambda)_{ii}(h_i - f_i)^2\Big\}, \quad (19)
\end{aligned}$$

which leads to $E\{T_1^2(h)\} \leq K_{22} d_n \|h - f\|_n^2$ for some K_{22}, where $d_n = \max_{1 \leq i \leq n} \mathbf{H}(\lambda)_{ii}$.

We are now ready to characterize $\sup_{h \in \mathcal{F}_L} T_1(h)$. First fix $L_0 > 0$ and $\epsilon > 0$. Since $\mathbf{H}(\lambda)$ has eigenvalues restricted to $[0, 1]$ and $\rho'(\cdot)$ is bounded, there exists K_{23} such that

$|T_1(h_1) - T_1(h_2)| < K_{23}\|h_1 - h_2\|_n$. Choose a fixed $r_0 > 0$ such that $r_0/L_0 < \epsilon/(4K_{23})$, and find a collection of open balls $\{\mathcal{B}_s\}_{s\in S_0}$ defined by $\mathcal{B}_s = \{h \in \mathcal{H} : \|h - h_s\|_n < r_0\}$ centered at h_s such that $\mathcal{C} \subseteq \bigcup_{s\in S_0} \mathcal{B}_s$. Since $\mathcal{C}$ is compact from $(\mathcal{A}.4)$, there exists a finite subset $S_1 \subseteq S_0$ such that $\mathcal{F}_{L_0} \subseteq \mathcal{C} \subseteq \bigcup_{s\in S_1} \mathcal{B}_s$. Define $S_2 = S_1 \backslash \{s \in S_1 : \mathcal{B}_s \cap \mathcal{F}_{L_0} = \phi\}$ and denote the number of elements in S_2 by N_0. It is easy to see that $\|h_s - f\|_n < r_0 + L_0$ for all $s \in S_2$ and $f \in \mathcal{F}_{L_0}$. By $\mathcal{F}_{L_0} \subseteq \bigcup_{s\in S_2} \mathcal{B}_s$,

$$\begin{aligned}
\sup_{h\in\mathcal{F}_{L_0}} |T_1(h)| &\leq \max_{s\in S_2} \sup_{h\in\mathcal{B}_s} |T_1(h)| \leq \max_{s\in S_2} \sup_{h\in\mathcal{B}_s} \{|T_1(h) - T_1(h_s)| + |T_1(h_s)|\} \\
&\leq \max_{s\in S_2} |T_1(h_s)| + \max_{s\in S_2} \left[\sup_{h\in\mathcal{B}_s} \{|T_1(h) - T_1(h_s)|\}\right]. \qquad (20)
\end{aligned}$$

Consider the first term of (20), using Bonferroni's inequality, Markov's inequality and the bounds on $E\{T_1^2(h)\}$ and $\|h_s - f\|_n$,

$$\begin{aligned}
\Pr\left\{\max_{s\in S_2} T_1(h_s) < \frac{\epsilon L_0}{4}\right\} &\geq 1 - \sum_{s\in S_2} \Pr\left\{T_1(h_s) \geq \frac{\epsilon L_0}{4}\right\} \\
&\geq 1 - \sum_{s\in S_2} \frac{4E\{T_1(h_s)\}}{\epsilon L_0} \\
&\geq 1 - \frac{4(K_{22}d_n)^{\frac{1}{2}} N_0 \|h_s - f\|_n}{\epsilon L_0} \\
&> 1 - \frac{4(K_{22}d_n)^{\frac{1}{2}} N_0 (r_0 + L_0)}{\epsilon L_0}.
\end{aligned}$$

The second term of (20) is obvious, $\max_{s\in S_2}\left[\sup_{h\in\mathcal{B}_s}\{|T_1(h) - T_1(h_s)|\}\right] < K_{23}r_0$. Recall that r_0 is chosen by $K_{23}r_0 < \epsilon L_0/4$. Find an N_1 such that $4(K_{22}d_n)^{1/2}N_0(r_0 + L_0)/\epsilon L_0 < \epsilon$ for $n > N_1$, it follows that $\Pr\Big\{\sup_{h\in\mathcal{F}_{L_0}} T_1(h) < \epsilon L_0/2\Big\} > 1 - \epsilon$.

Now we need to prove that the above probability bound holds uniformly for all $\mathcal{F}_L = \{h \in \mathcal{C} : \|h - f\|_n \leq L\}$, $0 < L \leq L_0$, denoting $d_L = L_0/L$. Define a linear map $\mathcal{P}(h) = f + (h - f)/d_L$ that shrinks h towards the center f by a factor $0 < d_L^{-1} < 1$. It is easy to verify that the image $\mathcal{P}(\mathcal{F}_{L_0}) = \{\mathcal{P}(h) : h \in \mathcal{F}_{L_0}\}$ is $\mathcal{F}_L = \{h \in \mathcal{C} : \|h - f\|_n \leq L\}$, observing that $\|\mathcal{P}(h) - f\|_n \leq L$ and $J(\mathcal{P}(h))$ is finite, i.e., $\mathcal{P}(h) \in \mathcal{C}$. Moreover, apply the same map $\mathcal{P}$ on $\mathcal{B}_s = \{h \in \mathcal{H} : \|h - f\|_n < r_0\}$, $s \in S_2$. The resulting image is

$\mathcal{B}_s^* = \{h \in \mathcal{H} : \|h - \{f + (h_s - f)/d_L\}\|_n < r_0/d_L\}$ by noticing $\mathcal{P}(h_s) = f + (h_s - f)/d_L$ and $\|\mathcal{P}(h) - \mathcal{P}(h_s)\|_n = \|h - h_s\|_n/d_L \leq r_0/L_0$ for $h \in \mathcal{B}_s$. We next verify that the finite collection $\{\mathcal{B}_s^* : s \in S_2\}$ covers $\mathcal{F}_L$ in the same manner as $\{\mathcal{B}_s : s \in S_2\}$ covers $\mathcal{F}_{L_0}$. For any $h \in \mathcal{F}_L$, $\|\mathcal{P}^{-1}(h) - f\|_n = \|f + d_L(h-f) - f\|_n = d_L\|h - f\|_n \leq L_0$ and $J(\mathcal{P}^{-1}(h))$ is bounded, i.e., $\mathcal{P}^{-1}(h) \in \mathcal{C}$, thus $\mathcal{P}^{-1}(h) \in \mathcal{F}_{L_0}$. Then there exists some $s \in S_2$ such that $\mathcal{P}^{-1}(h) \in \mathcal{B}_s$, thus $\|\mathcal{P}^{-1}(h) - h_s\|_n = \|f + d_L(h-f) - h_s\|_n < r_0$, implying $\|h - \{f + (h_s - f)/d_L\|_n < r_0/d_L$ and thus $h \in \mathcal{B}_s^*$. Therefore we can use the same argument for $\Pr\Big\{ \sup_{h \in \mathcal{F}_L} T_1(h) > \epsilon L \Big\}$. It is important to note that the choice of N_1 only depends on N_0, r_0/L_0 determined by ϵ/K_{23}, ϵ and K_{22}, where N_0 is the size of S_2 and the radius of $\mathcal{B}_s^*$ is $r = r_0/d_L$ implying that $r/L = r_0/L_0$. Thus N_1 does not depends on L at all. Then we conclude that, for $n > N_1$, $\Pr\Big\{ \sup_{h \in \mathcal{F}_L} T_1(h) < \epsilon L/2 \Big\} > 1 - \epsilon$ holds uniformly for all $0 < L \leq L_0$.

For $T_2(h)$, note that w_{ii} and w_{ii}^{-1} and are uniformly bounded and the eigenvalues of $\mathbf{H}(\lambda)$ are restricted in $[0, 1]$, thus $T_2^2(h) \leq K_{24}\|\mathbf{u}_2\|_n^2$, where $u_{2,i} = (f_i - h_i)^2 \int_0^1 \int_0^1 t\rho''(r_i + tsw_{ii}^{1/2}(f_i - h_i))dsdt$. Without loss of generality, assume that $\phi''(t) = 0$, $t \notin \Delta_n$, for some Δ_n with its measure, denoted by $|\Delta_n|$, tends to 0 as $n \to \infty$. A typical construction is to modify ϕ_c in (6) using cubic splines for $c \leq |t| \leq c + \tau_n$ and set $\phi_{c,n}(t) = (c + \tau_n)\text{sign}(t)$ for $|t| \geq c + \tau_n$, i.e., $\Delta_n = [-c - \tau_n, -c] \cup [c, c + \tau_n]$. The integral $\int_0^1 \int_0^1 t\rho''(r_i + tsw_{ii}^{1/2}(f_i - h_i))dsdt$ is thus bounded by $K_{25} \int_0^1 \int_0^1 tI(\{r_i + tsw_{ii}^{1/2}(f_i - h_i) \in \Delta_n\})dsdt \leq K_{25}|\Delta_n|/|h_i - f_i|$, for some $K_{25} > 0$. Then $T_2^2(h) \leq K_{26}|\Delta_n|^2\|h - f\|_n^2$ for K_{26}, and there exists an N_2 such that, for $n > N_2$, $\sup_{h \in \mathcal{F}_L} T_2(h) \leq \epsilon L/2$ holds uniformly for all $0 < L \leq L_0$ with probability 1.

Combining both uniform bounds with $N = \max(N_1, N_2)$ yields (17). The proof of (18) follows the same arguments for (17) due to $\mathbf{H}^2(\lambda)_{ii} < \mathbf{H}(\lambda)_{ii}$. That is, one can replace $\mathbf{H}^2(\lambda)$ by $\mathbf{H}(\lambda)$ and all the steps are still valid.

Lemma A3. *(Bounds on score mapping). Let* $U(x) = x + \mathbf{W}^{-1}\mathbf{H}(\lambda) \times \boldsymbol{\psi}(x_w + f_w; \mathbf{z}_w)$ *and* $\mathcal{F}_n = \{h \in \mathcal{C} : \|h - f\|_n \leq L_n\}$ *with* $L_n = C_n^{1/2}/\delta$ *for an arbitrary* $\delta > 0$ *(notation abused*

for convenience). Then there is an N such that, for any $n > N$,

$$\Pr\Big\{U(x) \in \mathcal{F}_n - f\Big\} > 1 - \delta, \tag{21}$$

holds uniformly for all $x \in \mathcal{F}_n - f$.

Proof of Lemma A3. For any $h \in \Re^n$, denoting $x = h - f$, we have

$$\begin{aligned} U(x) &= \mathbf{W}^{-1}\mathbf{H}(\lambda)\boldsymbol{\psi}(h_w; \mathbf{z}_w) + (h - f) \\ &= \mathbf{W}^{-1}\mathbf{H}(\lambda)\left\{\boldsymbol{\psi}(h_w; \mathbf{z}_w) - \check{\boldsymbol{\psi}}(h_w; \tilde{\mathbf{z}}_w)\right\} \\ &\quad + \left\{\mathbf{W}^{-1}\mathbf{H}(\lambda)\check{\boldsymbol{\psi}}(h_w; \tilde{\mathbf{z}}_w) + (h - f)\right\}, \end{aligned}$$

where the second term in r.h.s simplifies to $(\tilde{f} - f)$, by noticing $\mathbf{H}(\lambda) = (\mathbf{I} + 2\lambda\mathbf{R})^{-1}$ and

$$\begin{aligned} \mathbf{W}^{-1}\mathbf{H}(\lambda)\check{\boldsymbol{\psi}}(h_w; \tilde{\mathbf{z}}_w) &= \mathbf{W}^{-\frac{1}{2}}\left\{\mathbf{H}(\lambda)\tilde{\mathbf{z}}_w - \mathbf{H}(\lambda)(\mathbf{I} + 2\lambda\mathbf{R})\mathbf{h}_w\right\} \\ &= \mathbf{W}^{-\frac{1}{2}}(\tilde{f}_w - \mathbf{h}_w) = \tilde{f} - h. \end{aligned}$$

Thus $\|U(x)\|_n \leq \|\mathbf{W}^{-1}\mathbf{H}(\lambda)\{\boldsymbol{\psi}(h_w; \mathbf{z}_w) - \check{\boldsymbol{\psi}}(h_w; \tilde{\mathbf{z}}_w)\}\|_n + \|\tilde{f} - f\|_n$. Define $\mathcal{F}_n$ as in Lemma A2 with $L_n = C_n^{1/2}/\delta$ for sufficiently large n such that $L_n \leq L_0$. Applying Lemma A2 with L_n and $\epsilon = \delta/2$ to the first term, we have

$$\Pr\left[\sup_{h \in \mathcal{F}_n} \left\|\mathbf{W}^{-1}\mathbf{H}(\lambda)\left\{\boldsymbol{\psi}(h_w; \mathbf{z}_w) - \check{\boldsymbol{\psi}}(h_w; \tilde{\mathbf{z}}_w)\right\}\right\|_n \leq \frac{C_n^{\frac{1}{2}}}{2}\right] > 1 - \frac{\delta}{2},$$

and recall $C_n = E\{\|\tilde{f} - f\|_n^2\}$, applying Chebyshev's inequality,

$$\Pr\left\{\|\tilde{f} - f\|_n \leq \left(\frac{2C_n}{\delta}\right)^{\frac{1}{2}}\right\} > 1 - \frac{\delta}{2}.$$

Combining the two probability statements, we have

$$\Pr\left[\sup_{x \in \mathcal{F}_n - f} \|U(x)\|_n \leq \left\{\frac{\delta}{2} + (2\delta)^{\frac{1}{2}}\right\}\left(\frac{C_n^{\frac{1}{2}}}{\delta}\right)\right] > 1 - \delta.$$

Letting $\{\delta/2 + (2\delta)^{1/2}\}$ strictly less than 1 leads to $\Pr\{\sup_{x \in \mathcal{F}_n - f} \|U(x)\|_n \leq C_n^{1/2}/\delta\} > 1 - \delta$ for sufficiently large n.

To show $U(x)+f \in \mathcal{C}$ for $x \in \mathcal{F}_n - f$ with large probability, it remains to verify that $J\{U(x)+f\}$ is bounded. Note that $J(f) = f^{\mathrm{T}}\mathbf{R}^* f$, $\mathbf{R}^* = \mathbf{W}^{1/2}\mathbf{R}\mathbf{W}^{1/2}$, for some $G_1 > 0$,

$$\begin{aligned} J\{U(x)+f\}^{\frac{1}{2}} &\leq \left\| \mathbf{R}^{*\frac{1}{2}}\mathbf{W}^{-1}\mathbf{H}(\lambda)\left\{\boldsymbol{\psi}(h_w;\mathbf{z}_w) - \check{\boldsymbol{\psi}}(h_w;\tilde{\mathbf{z}}_w)\right\}\right\| + \|\mathbf{R}^{*\frac{1}{2}}\tilde{f}\| \\ &= \left\| \mathbf{R}^{\frac{1}{2}}\mathbf{H}(\lambda)\mathbf{W}^{\frac{1}{2}}\mathbf{W}^{-1}\left\{\boldsymbol{\psi}(h_w;\mathbf{z}_w) - \check{\boldsymbol{\psi}}(h_w;\tilde{\mathbf{z}}_w)\right\}\right\| + J(\tilde{f})^{\frac{1}{2}} \\ &\leq G_1(2\lambda)^{-\frac{1}{2}}\left\|\mathbf{W}^{-1}\mathbf{H}(\lambda)^{\frac{1}{2}}\left\{\boldsymbol{\psi}(h_w;\mathbf{z}_w) - \check{\boldsymbol{\psi}}(h_w;\tilde{\mathbf{z}}_w)\right\}\right\| \\ &\quad + J(\tilde{f})^{\frac{1}{2}}. \end{aligned}$$

using the fact that $\|\mathbf{R}^{1/2}\mathbf{H}(\lambda)\| \leq (2\lambda)^{-1/2}\|\mathbf{H}(\lambda)^{1/2}\|$ and w_{ii}'s are bounded. Then applying Lemma A2 (18) with $L_n = C_n^{1/2}/\delta$ and $\epsilon = \delta/2$, for sufficiently large n, with probability greater than $(1-\delta/2)$,

$$\sup_{x\in\mathcal{F}_n - f}\left[G_1\left(\frac{n}{2\lambda}\right)^{\frac{1}{2}}\left\|\mathbf{W}^{-1}\mathbf{H}(\lambda)^{\frac{1}{2}}\left\{\boldsymbol{\psi}(h_w;\mathbf{z}_w) - \check{\boldsymbol{\psi}}(h_w;\tilde{\mathbf{z}}_w)\right\}\right\|_n\right] \leq \frac{G_1}{2\sqrt{2}}\left(\frac{nC_n}{\lambda}\right)^{\frac{1}{2}}.$$

To bound nC_n/λ, given ($\mathcal{A}$.3.b), we use the second inequality in Lemma A1, for some $G_2, G_3 > 0$,

$$\begin{aligned} \frac{nC_n}{\lambda} = \frac{E\left\{\left\|\mathbf{W}^{-\frac{1}{2}}(\tilde{f}_w - f_w)\right\|^2\right\}}{\lambda} &\leq \frac{G_2 E\left\{\|\tilde{f}_w - f_w\|^2\right\}}{\lambda} \\ &\leq 2G_2\left[J(f) + \frac{K_1\mathrm{tr}\left\{\mathbf{H}(\lambda)\right\}}{\lambda}\right] \leq G_3. \end{aligned}$$

The upper bound of $J(\tilde{f})$ in probability can be easily established by applying Markov's inequality $\Pr\big[J(\tilde{f}) \leq 2E\{J(\tilde{f})\}/\delta\big] > 1-\delta/2$. Thus, by Lemma A1 and ($\mathcal{A}$.3.b), there exists $G_4 > 0$ such that $\Pr\{J(\tilde{f}) \leq G_4\} > 1-\delta/2$. By combining the two probability statements, we show that $J\{U(x)+f\}$ is bounded and completes the proof of Lemma A3.

Proof of Theorem 1. The basic idea behind the proof is to use the fact that the difference between the score functions for the robust estimator and the penalized least squares estimator for $\tilde{f}_w$, $\boldsymbol{\psi}(h_w;\mathbf{z}_w) - \check{\boldsymbol{\psi}}(h_w;\tilde{\mathbf{z}}_w)$, is small. First we find uniform bounds on the score

functions (11) and (12) achieved by the content of Lemma A2. Now we apply a fixed point argument to $U(x) = x + \mathbf{W}^{-1}\mathbf{H}(\lambda)\boldsymbol{\psi}(x_w + f_w; \mathbf{z}_w)$. Recall the definition of $\mathcal{C}$ in $(\mathcal{A}.4)$, let $\mathcal{F}_n = \{h \in \mathcal{C} : \|h - f\|_n \leq C_n^{1/2}/\delta\}$. Using Lemma A3 and Brouwer's fixed point theorem, there must exist at least one point $\hat{x} \in \mathcal{F}_n - f$ such that $\Pr\{U(\hat{x}) = \hat{x}\} > 1 - \delta$. Then it is easy to verify that $\boldsymbol{\psi}(\hat{x}_w + f_w; \mathbf{z}_w) = 0$, i.e., a robust estimate $\hat{f} = \hat{x} + f$ exists in a neighborhood of f with probability greater than $(1 - \delta)$.

It remains to bound the quantity $\|\hat{f} - \tilde{f}\|_n$. In view of $\tilde{f}_w = \mathbf{H}(\lambda)\tilde{\mathbf{z}}_w$, $\boldsymbol{\psi}(\hat{f}_w; \mathbf{z}_w) = \mathbf{0}$, $\tilde{f} = \mathbf{W}^{-1/2}\tilde{f}_w$, $\hat{f} = \mathbf{W}^{-1/2}\hat{f}_w$, $\check{\boldsymbol{\psi}}(\hat{f}_w; \tilde{\mathbf{z}}_w) = \mathbf{W}^{1/2}\{(\tilde{\mathbf{z}}_w - \hat{f}_w) - 2\lambda\mathbf{R}\hat{f}_w\}$ and $\mathbf{H}(\lambda) = (\mathbf{I} + 2\lambda\mathbf{R})^{-1}$, one has $\|\mathbf{W}^{-1}\mathbf{H}(\lambda)\{\boldsymbol{\psi}(\hat{f}_w; \mathbf{z}_w) - \check{\boldsymbol{\psi}}(\hat{f}_w; \tilde{\mathbf{z}}_w)\}\|_n = \|\mathbf{W}^{-1}\mathbf{H}(\lambda)\check{\boldsymbol{\psi}}(\hat{f}_w; \tilde{\mathbf{z}}_w)\|_n = \|\mathbf{W}^{-1/2}(\tilde{f}_w - \hat{f}_w)\|_n = \|\tilde{f} - \hat{f}\|_n$. Therefore, for sufficiently large n such that $L_n = C_n^{1/2}/\delta \leq L_0$, applying Lemma A2 with $L_n = C_n^{1/2}/\delta$ and $\epsilon = \delta^2 < \delta$, we have

$$\Pr\left[\|\hat{f} - \tilde{f}\|_n \leq \delta C_n^{\frac{1}{2}}\right] > 1 - \delta^2 \geq 1 - \delta,$$

which completes the proof.

B Derivation of (14)

The GIC formula of Konishi and Kitagawa (1996) contains two terms: a data fidelity term and a penalty term. For the current problem, it is straightforward to show that the data fidelity term is $-2\sum_i q(y_i, \hat{\mu}_{i\lambda})$. To derive the penalty term, we first note that $\hat{\boldsymbol{\beta}}$ is an M-estimator with influence function $\mathrm{IF}(y; \boldsymbol{\psi}, F) = \mathbf{P}(\boldsymbol{\psi}, F)^{-1}\boldsymbol{\psi}\{y, T(F)\}$, where

$$\mathbf{P}(\boldsymbol{\psi}, F)^{\mathrm{T}} = -\int \left.\frac{\partial \boldsymbol{\psi}(z, \boldsymbol{\beta})^{\mathrm{T}}}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=T(F)} dF(z, x).$$

For any M-estimator, Konishi and Kitagawa (1996) provide a general mechanism for deriving the penalty term. For our $\hat{\boldsymbol{\beta}}$, the derived penalty term is

$$2 \times \mathrm{tr}\left[\mathbf{P}(\boldsymbol{\psi}, F)^{-1}\int \boldsymbol{\psi}\{z, T(F)\}\left.\frac{\partial q(z, \mu)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right|_{\boldsymbol{\beta}=T(F)} dF(z, x)\right].$$

Direct calculations show that

$$\mathbf{P}(\boldsymbol{\psi}, F)^{\mathrm{T}} = -\int \frac{\partial \boldsymbol{\gamma}(z, \boldsymbol{\beta})^{\mathrm{T}}}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=T(F)} dF(z, x) + \frac{1}{n}\mathbf{S}^{\mathrm{T}}, \tag{22}$$

where $\boldsymbol{\gamma}(z, \boldsymbol{\beta}) = \nu(z, \mu)\zeta(\mu)\partial\mu/\partial\boldsymbol{\beta} - a(\boldsymbol{\beta})$. Also,

$$\begin{aligned} & \int \boldsymbol{\psi}\{z, T(F)\}\frac{\partial q(z, \mu)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=T(F)} dF(z, x) \\ = & \int \boldsymbol{\psi}\{z, T(F)\}\boldsymbol{\gamma}\{z, T(F)\}^{\mathrm{T}} dF(z, x) \\ = & \int \boldsymbol{\gamma}\{y, T(F)\}\boldsymbol{\gamma}\{y, T(F)\}^{\mathrm{T}} dF(y, x). \end{aligned} \tag{23}$$

Both the first term on the right hand side of (22) and (23) are unknown. It is shown in (Cantoni and Ronchetti, 2001) that these terms can be estimated by $\mathbf{X}^{\mathrm{T}}\mathbf{B}\mathbf{X}/n$ and $\mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X}/n - a(\boldsymbol{\beta})a(\boldsymbol{\beta})^{\mathrm{T}}$ respectively. Combining the above results we obtain (14).

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.

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.

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

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.

Copas, J. B. (1988) Binary regression models for contaminated data. *Journal of the Royal Statistical Society Series B*, **50**, 225–265.

Croux, C., Gijbels, I. and Prosdocimi, I. (2011) Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics*. To appear.

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

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.

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

Huber, P. J. (1973) Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, **1**, 799–821.

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.

Konishi, S. and Kitagawa, G. (1996) Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.

Kuchenhoff, H. and Carroll, R. J. (1997) Segmented regression with errors in predictors: semi-parametric and parametric methods. *Statistics in Medicine*, **16**, 169–188.

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.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*. London: Chapman & Hall, 2 edn.

Morgenthaler, S. (1992) Least-absolute-deviations fits for generalized linear models. *Biometrika*, **79**, 747–754.

Nychka, D. (1995) Splines as local smoothers. *The Annals of Statistics*, **23**, 1175–1197.

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.

Peng, R. D. and Welty, L. J. (2004) The NMMAPSdata package. *R News*, **4**, 10–14.

Preisser, J. S. and Qaqish, B. F. (1999) Robust regression for clustered data with applications to binary responses. *Biometrics*, **55**, 574–579.

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

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

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.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B*, **36**, 111–147.

Wahba, G. (1990) *Spline models for observational data*. Pennsylvania: Society for Industrial and Applied Mathematics.

Wood, S. (2006) *Generalized Additive Models: An Introduction with R*. Boca Raton: Chapman & Hall.

Wood, S. N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, **99**, 673–686.

— (2008) Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society Series B*, **70**, 495–518.