Biometrics 68, 31–44
March 2012

DOI: 10.1111/j.1541-0420.2011.01630.x

Robust Estimation of Mean and Dispersion Functions in Extended Generalized Additive Models

Christophe Croux,[1,2,3] Irène Gijbels,[2,4,*] and Ilaria Prosdocimi[2,4]

[1]Faculty of Business and Economics, Katholieke Universiteit Leuven, Naamsestraat 69, Box 3555,
B-3000 Leuven, Belgium

[2]Leuven Statistics Research Center (LStat), Katholieke Universiteit Leuven, Celestijnenlaan 200B, Box 5307,
B-3001 Leuven (Heverlee), Belgium

[3]Tilburg University, CentER, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

[4]Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200B, Box 2400,
B-3001 Leuven (Heverlee), Belgium

[*]*email:* Irene.Gijbels@wis.kuleuven.be

Summary. Generalized linear models are a widely used method to obtain parametric estimates for the mean function. They have been further extended to allow the relationship between the mean function and the covariates to be more flexible via generalized additive models. However, the fixed variance structure can in many cases be too restrictive. The extended quasilikelihood (EQL) framework allows for estimation of both the mean and the dispersion/variance as functions of covariates. As for other maximum likelihood methods though, EQL estimates are not resistant to outliers: we need methods to obtain robust estimates for both the mean and the dispersion function. In this article, we obtain functional estimates for the mean and the dispersion that are both robust and smooth. The performance of the proposed method is illustrated via a simulation study and some real data examples.

Key words: Dispersion; Generalized additive modeling; M-estimation; Mean regression function; P-splines; Quasilikelihood; Robust estimation.

1. Introduction

Statistical modeling aims at describing how a phenomenon of interest changes with respect to some other quantities. Generally most of the modeling efforts focus on studying how the expected value of the dependent variable Y, denoted by μ, changes as a function of the covariates $\boldsymbol{X}_d = (X_1, \ldots, X_d)$. Generalized linear models (GLM, McCullagh and Nelder, 1989) are one of the most popular techniques to model the mean of different types of distributions belonging to the exponential family. Standard GLM though are not always most appropriate to model the data at hand; the assumption of a linear relationship between (a transformation of) μ and the covariates might be too restrictive. Also, GLM estimates are maximum likelihood estimates, which can be severely influenced by the presence of outliers. For both issues possible solutions have been proposed: we can allow the relationship between (a transformation of) μ and the covariates to be of a smooth unknown shape via generalized additive models (GAM; Hastie and Tibshirani, 1990) and we can obtain estimates that are robust via the approach proposed for GLMs by Cantoni and Ronchetti (2001a). Recently Alimadad and Salibian-Barrera (2011) propose a method for robust estimation of GAM.

In this article, we develop a statistical procedure to obtain smooth and robust estimates for *both* the *mean* and the *dispersion* function in a *multivariate* covariates setting. We thus allow for heteroscedasticity in the model. Estimating how the variance changes with respect to $\boldsymbol{X}_d$ can be in some cases of interest by itself, or it can be pursued to obtain a more appropriate fit. From the distributional assumptions made in GLM follows a fixed relationship between the shape of the variance and the mean, but the variance observed in real data often deviates from the theoretical model. Common deviations from the usual assumptions in GLM are heteroscedasticity in normal data and over- or underdispersion in count and proportion data. The estimation of the variance can indeed be a crucial point and different approaches have been proposed to tackle this problem: see the introduction in Gijbels, Prosdocimi, and Claeskens (2010) or Hinde and Demétrio (1998) for a review of possible methods. In particular, Rigby and Stasinopoulos (2005) propose a class of models generalized additive models for location, scale and shape (GAMLSS) that allows the user to obtain smooth functional estimates for different parameters of a given distribution (from which the data are assumed to be generated). To model the dispersion of, for example, count or proportion data, one needs to specify a distribution that also allows for dispersion modeling. Typically, one would assume data to come from a negative binomial or a beta binomial distribution, which extend the standard Poisson and binomial distribution via hierarchical modeling reasoning. These distributions only allow overdispersion modeling, and cannot be used in case of underdispersed data or when data show a combination of overdispersion and underdispersion.

In this work we use the extended quasilikelihood (EQL) approach (Nelder and Pregibon, 1987; McCullagh and Nelder, 1989) to obtain estimates for the dispersion function. The EQL framework allows for a very flexible type of modeling in which one also easily models both under- and overdispersion. Just as the standard quasilikelihood, EQL estimators can be severely affected by outliers, and we use the techniques proposed by Cantoni and Ronchetti (2001a) to robustly estimate both the mean and the dispersion function in our setting. Moreover, the methods presented here allow these estimates to be a flexible function of the covariates. We mostly use GAM combined with P-splines (Marx and Eilers, 1998). For simplicity of presentation we limit ourselves to P-splines. Different linear smoothers, possibly more appropriate than P-splines for a problem at hand, could be employed without undermining the main contribution of this article.

The remainder of the article is organized as follows: in Sections 2 and 3, standard and robust estimation methods within the EQL framework are presented. In Section 4, we introduce the generalized additive models framework to obtain smooth and robust estimates of the mean and the dispersion function. In Section 5, we discuss how to optimally choose the smoothing parameters. We show the performance of the proposed methods via a simulation study and real data examples in Sections 6 and 7, respectively. In Section 8, we provide a comparative study for different model choices, evaluating their impact on the final performance via a simulation study.

The methods presented in this article have been implemented in R. The files containing the implemented functions can be found at: `http://wis.kuleuven.be/stat/codes.html`.

2. Extended Quasilikelihood

To write a likelihood function for a certain model, we need to make assumptions on the distribution of the process of interest. In the quasilikelihood framework, rather than making a full distributional assumption, one only specifies the relationship between the mean and the variance of the process of interest. Estimates are obtained by maximizing a quasilikelihood function, which shares key properties with a likelihood function, but can be obtained with weaker assumptions (Wedderburn, 1974). We consider

$$E[Y \mid \boldsymbol{X}_d = \boldsymbol{x}_d] = \mu(\boldsymbol{x}_d) \text{ and } \mathrm{Var}[Y \mid \boldsymbol{X}_d = \boldsymbol{x}_d] = \phi V(\mu(\boldsymbol{x}_d)),$$

with $V(\cdot)$ a known function, and write the quasi-log-likelihood function as:

$$Q(y, \mu(\boldsymbol{x}_d)) = \int_y^{\mu(\boldsymbol{x}_d)} \frac{y-t}{\phi V(t)} dt \; .$$

We also introduce a monotone and twice differentiable function $\eta(\cdot)$, which transforms the expected value of $(Y \mid \boldsymbol{X}_d = \boldsymbol{x}_d)$ via a link function $g(\cdot)$, and is modeled as a linear combination of some generic functions of the covariates: $\eta(\boldsymbol{x}_d) = g(\mu(\boldsymbol{x}_d)) = \alpha_{\mu,0} + \eta_1(x_1) + \cdots + \eta_d(x_d)$. Furthermore, the quasideviance function

$$d(y, \mu(\boldsymbol{x}_d)) = -2Q(y, \mu(\boldsymbol{x}_d)) = 2\int_{\mu(\boldsymbol{x}_d)}^{y} \frac{y-t}{\phi V(t)} dt \; , \qquad (1)$$

measures the discrepancy between the value of y and the expected value of the original distribution.

In the quasilikelihood setting the relationship between the variance of $(Y \mid \boldsymbol{X}_d = \boldsymbol{x}_d)$ and the covariates is totally governed by the functional form of $V(\mu(\boldsymbol{x}_d))$. This relationship might however be too restrictive, and one might be interested in adding an extra dispersion parameter in the model, which varies as a function of the covariates (Nelder and Pregibon, 1987). We thus consider

$$\begin{aligned} E[Y \mid \boldsymbol{X}_d = \boldsymbol{x}_d] &= \mu(\boldsymbol{x}_d) \text{ and} \\ \mathrm{Var}[Y \mid \boldsymbol{X}_d = \boldsymbol{x}_d] &= \phi\gamma(\boldsymbol{x}_d)V(\mu(\boldsymbol{x}_d)), \end{aligned} \qquad (2)$$

with $\gamma(\boldsymbol{x}_d)$ an extra dispersion function. To model this dispersion function we take:

$$\begin{aligned} E[d(Y,\mu) \mid \boldsymbol{X}_d = \boldsymbol{x}_d] &= \gamma(\boldsymbol{x}_d) \text{ and} \\ \mathrm{Var}[d(Y,\mu) \mid \boldsymbol{X}_d = \boldsymbol{x}_d] &= 2\gamma^2(\boldsymbol{x}_d). \end{aligned} \qquad (3)$$

The structure used to model the dispersion is a mirror image of the mean modeling: the quasideviance is used as a response variable with mean function $\gamma(\boldsymbol{x}_d)$ and a suitable variance function is also assumed. Note how the chosen variance structure for the dispersion corresponds to assuming $(d(Y,\, \mu) \mid \boldsymbol{X}_d = \boldsymbol{x}_d) \sim \gamma(\boldsymbol{x}_d)\ \chi_1^2$. We introduce here a second monotone and twice differentiable function $\xi(\cdot)$, which transforms the expected value of $(d(Y,\, \mu) \mid \boldsymbol{X}_d = \boldsymbol{x}_d)$ via a link function h^{-1} $(\cdot)$, and is modeled as a linear combination of some generic functions of the covariates: $\xi(\boldsymbol{x}_d) = h^{-1}\ (\gamma(\boldsymbol{x}_d)) = \alpha_{\gamma,0} + \xi_1(x_1) + \cdots + \xi_d(x_d)$.

In the usual parametric approach one takes the relationship between the link functions and the covariates $\boldsymbol{X}_d$ to be linear: $\eta(\boldsymbol{x}_d) = \alpha_{\mu,0} + x_1\alpha_{\mu,1} + \cdots + x_d\alpha_{\mu,d}$ and $\xi(\boldsymbol{x}_d) = \alpha_{\gamma,0} + x_1\alpha_{\gamma,1} + \cdots + x_d\alpha_{\gamma,d}$ with $\boldsymbol{\alpha}_\mu = (\alpha_{\mu,0},\, \alpha_{\mu,1}, \ldots, \alpha_{\mu,d})^T$ and $\boldsymbol{\alpha}_\gamma = (\alpha_{\gamma,0},\, \alpha_{\gamma,1}, \ldots, \alpha_{\gamma,d})^T$ the vectors of parameters that need to be estimated.

For a given independent and identically distributed (i.i.d.) sample $(\boldsymbol{x},\, \boldsymbol{y}) = ((\boldsymbol{x}_{d,1},\, y_1)^T, \ldots, (\boldsymbol{x}_{d,n},\, y_n)^T)^T$, we take the $n \times (d+1)$ regression matrices $\boldsymbol{B}_\mu = [\boldsymbol{1}_n\ \ \boldsymbol{x}]$ and $\boldsymbol{B}_\gamma = [\boldsymbol{1}_n\ \ \boldsymbol{x}]$, where we denote $\boldsymbol{1}_n = (1, \ldots, 1)^T$ the unit vector of length n and model $\eta(\boldsymbol{x}) = \boldsymbol{B}_\mu\boldsymbol{\alpha}_\mu$ and $\xi(\boldsymbol{x}) = \boldsymbol{B}_\gamma\boldsymbol{\alpha}_\gamma$.

Denote by $\mu(\boldsymbol{x}) = (\mu(\boldsymbol{x}_{d,1},\, \boldsymbol{\alpha}_\mu), \ldots, \mu(\boldsymbol{x}_{d,n},\, \boldsymbol{\alpha}_\mu))^T$ the vector of computed $\mu(\cdot)$ values in each datapoint, by $V(\mu(\boldsymbol{x}))$ the vector of values of the $V(\cdot)$ function evaluated at each $\mu(\boldsymbol{x})$ point, and similarly $\boldsymbol{\gamma}(\boldsymbol{x}) = (\gamma(\boldsymbol{x}_{d,1},\, \boldsymbol{\alpha}_\gamma), \ldots, \gamma(\boldsymbol{x}_{d,n},\, \boldsymbol{\alpha}_\gamma))^T$ the vector of $\gamma(\cdot)$ values. The estimation of the mean and dispersion function is done via a two-step procedure that alternates between the estimation of $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_\gamma$ as in McCullagh and Nelder (1989):

Step 1: For a given $\boldsymbol{\alpha}_\gamma$ and $\boldsymbol{\gamma}$ (i.e., the estimated dispersion function from the previous iteration step) the EQL estimator of $\boldsymbol{\alpha}_\mu$ is obtained as the solution to

$$\boldsymbol{B}_\mu^T \left(\frac{\boldsymbol{y} - \mu(\boldsymbol{x})}{\phi\boldsymbol{\gamma}V(\mu(\boldsymbol{x}))}\ \frac{d\mu}{d\eta}(\boldsymbol{x}) \right) = \boldsymbol{0}, \qquad (4)$$

where $\boldsymbol{0}$ is the null vector. The multiplication of the vectors within the brackets is done element-wise. In the notation, we drop the dependence of $\mu(\boldsymbol{x})$ on $\boldsymbol{\alpha}_\mu$ to make the formulas more readable. Once $\boldsymbol{\alpha}_\mu$, and consequently $\mu(\boldsymbol{x})$, is estimated by solving (4), we compute the vector of deviances $\boldsymbol{d} = d(\boldsymbol{y}, \boldsymbol{\mu}) = (d(y_1, \mu(\boldsymbol{x}_{d,1})), \ldots, d(y_n, \mu(\boldsymbol{x}_{d,n})))^T$.

Step 2: For a given $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\mu}$ (i.e., the estimated mean function from the previous step) the EQL estimator of $\boldsymbol{\alpha}_\gamma$ is then obtained by solving:

$$\boldsymbol{B}_\gamma^T \left(\frac{\boldsymbol{d} - \gamma(\boldsymbol{x})}{2\gamma^2(\boldsymbol{x})} \frac{d\gamma}{d\xi}(\boldsymbol{x}) \right) = \boldsymbol{0}.$$

The estimation procedure alternates between the two steps till convergence.

3. Robust Estimation of Mean and Dispersion

The EQL estimators proposed in Section 2 can been shown to have an unbounded influence function. Outlying points, as well as bad leverage points, can have a severe effect on the performance of the estimator. To mitigate the effect of outliers and to obtain bounded influence functions, an M-type estimation procedure is followed similar as in Cantoni and Ronchetti (2001a).

The M-estimator for $\boldsymbol{\alpha}_\mu$ is obtained as the solution of the equation:

$$\Psi_s(\boldsymbol{y}, \mu(\boldsymbol{x})) = \boldsymbol{B}_\mu^T \left(s_\psi(\boldsymbol{y}, \mu(\boldsymbol{x})) w(\boldsymbol{x}) \boldsymbol{\mu}' - a(\boldsymbol{\alpha}_\mu) \right) = \boldsymbol{0}, \quad (5)$$

with $\boldsymbol{\mu}' = \frac{d\mu}{d\eta}(\boldsymbol{x})$. Robustness against outlying points is obtained if $\Psi_s(\cdot,\cdot)$ is a bounded function. For this, we take

$$s_\psi(\boldsymbol{y}, \mu(\boldsymbol{x})) = \psi_c \left(\frac{\boldsymbol{y} - \mu(\boldsymbol{x})}{\sqrt{\phi \boldsymbol{\gamma} V(\mu(\boldsymbol{x}))}} \right) \frac{1}{\sqrt{\phi \boldsymbol{\gamma} V(\mu(\boldsymbol{x}))}},$$

with ψ_c the Huber function defined as

$$\psi_c(x) = \begin{cases} x & \text{if } |x| \le c \\ c \ \mathrm{sign}(x) & \text{if } |x| > c. \end{cases} \quad (6)$$

Further $w(\cdot)$ in (5) is a weight function, which controls the effect of leverage points on the estimate.

The tuning constant c in (6) balances the robustness and the efficiency of the estimate; if $w(\boldsymbol{x}) = \boldsymbol{1}_n$ and $c = \infty$ (5) boils down to (4). Cantoni and Ronchetti (2001a) discuss procedures to choose c. Unless otherwise stated, we take $c = 1.345$, the standard value that ensures 95% efficiency for the normal model. Our experience shows that this value gives reasonable results for other models as well (see Section 8.2 for further discussion on the choice of this tuning parameter). In (5), the constant

$$a(\boldsymbol{\alpha}_\mu) = \bar{E}_n [s_\psi(\boldsymbol{y}, \mu(\boldsymbol{x})) \mid \boldsymbol{X}_d = \boldsymbol{x}_d] w(\boldsymbol{x}) \boldsymbol{\mu}'$$

ensures Fisher consistency. This means that the true parameter value $\boldsymbol{\alpha}_\mu$ is a solution of (5) for n tending to infinity (see, e.g., Heritier et al., 2009, p. 138). Here we used the shorthand notation $\bar{E}_n [s_\psi(\boldsymbol{y}, \mu(\boldsymbol{x})) \mid \boldsymbol{X}_d = \boldsymbol{x}_d]$ for $n^{-1} \sum_{i=1}^n E[s_\psi(Y, \mu(\boldsymbol{X}_d)) \mid \boldsymbol{X}_d = \boldsymbol{x}_{d,i}]$. This notation $\bar{E}_n$ will also be used in the next paragraph with a similar meaning involving a different quantity, as well as in Section 7.

Similarly, a robust estimate for $\boldsymbol{\alpha}_\gamma$ is obtained as the solution to:

$$\Psi_t(\boldsymbol{d}, \gamma(\boldsymbol{x})) = \boldsymbol{B}_\gamma^T \left(t_\psi(\boldsymbol{d}, \gamma(\boldsymbol{x})) w(\boldsymbol{x}) \boldsymbol{\gamma}' - b(\boldsymbol{\alpha}_\gamma) \right) = \boldsymbol{0}, \quad (7)$$

with $\boldsymbol{\gamma}' = \frac{d\gamma}{d\xi}(\boldsymbol{x})$. The estimate is robust against outliers, provided that we take $\Psi_t(\boldsymbol{d}, \gamma(\boldsymbol{x}))$ to be a bounded function. As for the estimation of $\boldsymbol{\alpha}_\mu$ we take

$$t_\psi(\boldsymbol{d}, \gamma(\boldsymbol{x})) = \psi_c \left(\frac{\boldsymbol{d} - \gamma(\boldsymbol{x})}{\sqrt{2}\gamma(\boldsymbol{x})} \right) \frac{1}{\sqrt{2}\gamma(\boldsymbol{x})},$$

with ψ_c the Huber function defined in (6). Again, the constant

$$b(\boldsymbol{\alpha}_\gamma) = \bar{E}_n [t_\psi(\boldsymbol{d}, \gamma(\boldsymbol{x})) \mid \boldsymbol{X}_d = \boldsymbol{x}_d] w(\boldsymbol{x}) \boldsymbol{\gamma}'$$

ensures Fisher consistency.

Both (5) and (7) cannot be solved analytically. We use the iteratively reweighted least squares algorithm (Hampel et al., 1986). The estimation procedure is done in two steps as in Section 2.

4. Robust Generalized Additive Models

In the models described in Sections 2 and 3, we have taken $\eta(\boldsymbol{x}_d)$ and $\xi(\boldsymbol{x}_d)$ to be linear combinations of the covariates. This functional form can in many cases be too restrictive and we would like to let the form of each $\eta_j(\cdot)$ and $\xi_j(\cdot)$ to be as unspecified as possible, assuming only that these are smooth functions. To obtain such smooth estimates we use generalized additive models for both the mean and the dispersion function. GAM (Hastie and Tibshirani, 1990) extend GLM by allowing the mean to be a flexible function of the covariates. Marx and Eilers (1998) have developed a way to estimate the smooth components $\eta_j(\cdot)$ via penalized B-splines (P-splines). In Ruppert, Wand, and Carroll (2003, 2009), nonparametric regression via splines is further discussed, although a complete and clear presentation of penalized splines and GAM can be found in Wood (2006). Gijbels and Prosdocimi (2011) extended GAM to estimate both the mean and the dispersion as smooth functions of the covariates. We intend to further develop these extended GAM to obtain estimates for both the mean and the dispersion that are both smooth and robust. Before introducing a robust extended version of GAM, we briefly introduce penalized splines and their use in GAM fitting.

4.1 *P-splines and P-GAM*

The use of penalized splines dates back at least to Reinsch (1967) and was further discussed in Silverman (1985). See also Wahba (1980). Eilers and Marx (1996) introduced what they call P-splines, in which the use of B-splines is combined with a discrete difference type of penalty on the coefficients. Thanks to their good numerical properties and easiness of implementation, modeling via P-splines quickly became a popular tool in nonparametric regression. We briefly explain how to use P-splines combined with GAM.

First consider the case of one covariate. For a given set of knots $\{\kappa_1, \ldots, \kappa_k\}$, B-spline basis functions of degree p, are composed of polynomial pieces of degree p, joined together in an appropriate way at each knot point κ_j. This leads to a B-spline basis of dimension $K = p + k + 1$. For a given i.i.d. sample $(\boldsymbol{x}, \boldsymbol{y})$ we build a large B-splines base matrix $\boldsymbol{B}_\mu$ and use this as regression matrix in a GLM, taking $\eta(\boldsymbol{x}) = \boldsymbol{B}_\mu \boldsymbol{\alpha}_\mu$. The central idea in P-splines is to take $\boldsymbol{B}_\mu$ to be a very large B-spline base that would overfit the data and to then avoid such overfitting by adding a penalty term which controls the smoothness of the curve in the quasilikelihood. Estimates for

$\boldsymbol{\alpha}_\mu$ are obtained as the solution to:

$$\boldsymbol{B}_\mu^T \left(\frac{\boldsymbol{y} - \mu(\boldsymbol{x})}{\phi V(\mu(\boldsymbol{x}))} \boldsymbol{\mu}' \right) - \lambda \boldsymbol{P}_\mu \boldsymbol{\alpha}_\mu = \boldsymbol{0}, \tag{8}$$

with $\boldsymbol{P}_\mu$ an appropriate matrix representation of the difference operator. The smoothing parameter $\lambda > 0$ governs the balance between the overfitting and the smoothness of the fitted function.

P-splines can be used also when we are interested in determining the relationship between the expected value of Y and *multiple* covariates. We take $\eta(\boldsymbol{x}_d) = \alpha_{\mu,0} + \eta_1(x_1) + \cdots + \eta_d(x_d)$ and fit the generic component $\eta_j(x_j)$ via P-splines. For a given i.i.d. sample $(\boldsymbol{x}, \boldsymbol{y})$ we now build d large B-splines base matrices $\boldsymbol{B}_{\mu,1}, \ldots, \boldsymbol{B}_{\mu,d}$ and model $\eta(\boldsymbol{x})$ as a linear combination of the B-splines matrices $\eta(\boldsymbol{x}) = \alpha_{\mu,0} + \boldsymbol{B}_{\mu,1}\boldsymbol{\alpha}_{\mu,1} + \cdots + \boldsymbol{B}_{\mu,d}\boldsymbol{\alpha}_{\mu,d} = \boldsymbol{B}_\mu \boldsymbol{\alpha}_\mu$, with $\boldsymbol{B}_\mu = [\boldsymbol{1}_n, \boldsymbol{B}_{\mu,1}, \ldots, \boldsymbol{B}_{\mu,d}]$ the design matrix and $\boldsymbol{\alpha}_\mu = (\alpha_{\mu,0}^T, \boldsymbol{\alpha}_{\mu,1}^T, \ldots, \boldsymbol{\alpha}_{\mu,d}^T)^T$ the column vector of parameters to be estimated. To avoid overfitting for each of the d components we build d penalty matrices $\boldsymbol{P}_{\mu,1}, \ldots, \boldsymbol{P}_{\mu,d}$ and take $\boldsymbol{\lambda}_\mu = (\lambda_{\mu,1}, \ldots, \lambda_{\mu,d})$ the smoothing parameters governing the smoothness of the components. Taking $\boldsymbol{P}_\mu = \mathit{blockdiag}\ [0, \lambda_{\mu,1} \boldsymbol{P}_{\mu,1}, \ldots, \lambda_{\mu,d} \boldsymbol{P}_{\mu,d}]$ a block-diagonal penalty matrix, we obtain an estimate of $\boldsymbol{\alpha}_\mu$ as the solution of

$$\boldsymbol{B}_\mu^T \left(\frac{\boldsymbol{y} - \mu(\boldsymbol{x})}{\phi V(\mu(\boldsymbol{x}))} \boldsymbol{\mu}' \right) - \boldsymbol{P}_\mu \boldsymbol{\alpha}_\mu = \boldsymbol{0}\ .$$

In the next section, we combine the GAM and the robust EQL framework presented in Section 2 to obtain robust and smooth estimates for both the mean and the dispersion function.

4.2 *Robust Extended P-GAM*

We next allow the dispersion function to be a smooth function of the covariates, and we propose estimators that are smooth as well as robust against outliers. Again we use the EQL framework and make assumptions only on the first two moments of $(Y \mid \boldsymbol{X}_d = \boldsymbol{x}_d)$ and $(d(Y, \mu) \mid \boldsymbol{X}_d = \boldsymbol{x}_d)$ just as in (2) and (3). Once more, $\eta(\cdot)$ and $\xi(\cdot)$ are two link functions for the mean and the dispersion, respectively. These link functions are of the form $\eta(\boldsymbol{x}_d) = \alpha_{\mu,0} + \eta_1(x_1) + \cdots + \eta_d(x_d)$ and $\xi(\boldsymbol{x}_d) = \alpha_{\gamma,0} + \xi_1(x_1) + \cdots + \xi_d(x_d)$, where the $\eta_j(x_j)$ and $\xi_j(x_j)$ components are modeled via P-splines.

As in Section 4.1, for a given i.i.d. sample $(\boldsymbol{x}, \boldsymbol{y})$ we build the regression matrix $\boldsymbol{B}_\mu$ and the blockdiagonal penalty matrix $\boldsymbol{P}_\mu$. In the same way we build the regression matrix $\boldsymbol{B}_\gamma$ and, given a set of smoothing parameters $\boldsymbol{\lambda}_\gamma = (\lambda_{\gamma,1}, \ldots, \lambda_{\gamma,d})$, the penalty matrix $\boldsymbol{P}_\gamma$ for the modeling of the dispersion function.

Smooth and robust estimates for $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_\gamma$ are obtained as the solution to

$$\begin{aligned} \Psi_s(\boldsymbol{y}, \mu(\boldsymbol{x})) - \boldsymbol{P}_\mu \boldsymbol{\alpha}_\mu &= \boldsymbol{B}_\mu^T (s_\psi(\boldsymbol{y}, \mu(\boldsymbol{x})) w(\boldsymbol{x}) \boldsymbol{\mu}' - a(\boldsymbol{\alpha}_\mu)) \\ &\quad - \boldsymbol{P}_\mu \boldsymbol{\alpha}_\mu = \boldsymbol{0} \end{aligned} \tag{9}$$

and

$$\begin{aligned} \Psi_t(\boldsymbol{d}, \gamma(\boldsymbol{x})) - \boldsymbol{P}_\gamma \boldsymbol{\alpha}_\gamma &= \boldsymbol{B}_\gamma^T (t_\psi(\boldsymbol{d}, \gamma(\boldsymbol{x})) w(\boldsymbol{x}) \boldsymbol{\gamma}' - b(\boldsymbol{\alpha}_\gamma)) \\ &\quad - \boldsymbol{P}_\gamma \boldsymbol{\alpha}_\gamma = \boldsymbol{0}. \end{aligned} \tag{10}$$

These estimates can be shown to have a bounded influence function when $\Psi_s(\cdot,\cdot)$ and $\Psi_t(\cdot,\cdot)$ are bounded, as in Section 3. Note how (9) and (10) differ from (5) and (7) because now $\boldsymbol{B}_\mu$ and $\boldsymbol{B}_\gamma$ represent a larger combination of matrices, and there is the penalty term that ensures the smoothness of the fit.

Estimates for $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_\gamma$ are obtained via iterative procedures. In particular, the rule to update the current estimate $\tilde{\boldsymbol{\alpha}}_\mu$ of $\boldsymbol{\alpha}_\mu$ is:

$$\boldsymbol{\alpha}_\mu = \left(\boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu \boldsymbol{B}_\mu + \boldsymbol{P}_\mu \right)^{-1} \boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu \tilde{\boldsymbol{z}}_\mu, \tag{11}$$

where $\tilde{\boldsymbol{W}}_\mu = \mathit{diag}(-E[\frac{d}{d\boldsymbol{\alpha}_\mu} \Psi_s(Y, \tilde{\mu}(\boldsymbol{X}_d)) \mid \boldsymbol{X}_d = \boldsymbol{x}_{d,i}])_i$, $\tilde{\boldsymbol{z}}_\mu = \boldsymbol{B}_\mu \tilde{\boldsymbol{\alpha}}_\mu + \tilde{\boldsymbol{W}}_\mu^{-1} \Psi_s(\boldsymbol{y}, \tilde{\mu}(\boldsymbol{x}))$ and $\tilde{\mu}(\cdot)$ is the vector of current estimates for $\mu(\cdot)$ which depends on $\tilde{\boldsymbol{\alpha}}_\mu$. Similarly $\boldsymbol{\alpha}_\gamma$ is updated with the following scheme:

$$\boldsymbol{\alpha}_\gamma = \left(\boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma \boldsymbol{B}_\gamma + \boldsymbol{P}_\gamma \right)^{-1} \boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma\ \tilde{\boldsymbol{z}}_\gamma\ , \tag{12}$$

where $\tilde{\boldsymbol{W}}_\gamma = \mathit{diag}(-E[\frac{d}{d\boldsymbol{\alpha}_\gamma} \Psi_t(d(Y, \mu), \tilde{\gamma}(\boldsymbol{X}_d)) \mid \boldsymbol{X}_d = \boldsymbol{x}_{d,i}])_i$, $\tilde{\boldsymbol{z}}_\gamma = \boldsymbol{B}_\gamma \tilde{\boldsymbol{\alpha}}_\gamma + \tilde{\boldsymbol{W}}_\gamma^{-1} \Psi_t(\boldsymbol{d}, \tilde{\gamma}(\boldsymbol{x}))$ and $\tilde{\gamma}(\boldsymbol{x})$ is the vector of current estimates for $\gamma(\boldsymbol{x})$, which depends on $\tilde{\boldsymbol{\alpha}}_\gamma$.

Once convergence is reached we take $\hat{\eta}(\boldsymbol{x}) = \boldsymbol{H}_\mu(\boldsymbol{\lambda}_\mu) \tilde{\boldsymbol{z}}_\mu$, with $\boldsymbol{H}_\mu(\boldsymbol{\lambda}_\mu) = \boldsymbol{B}_\mu (\boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu \boldsymbol{B}_\mu + \boldsymbol{P}_\mu)^{-1} \boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu$, the hat matrix for the mean model, which depends on the values of $\boldsymbol{\lambda}_\mu$. Similarly we have $\hat{\xi}(\boldsymbol{x}) = \boldsymbol{H}_\gamma(\boldsymbol{\lambda}_\gamma) \tilde{\boldsymbol{z}}_\gamma$ with $\boldsymbol{H}_\gamma(\boldsymbol{\lambda}_\gamma) = \boldsymbol{B}_\gamma (\boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma \boldsymbol{B}_\gamma + \boldsymbol{P}_\gamma)^{-1} \boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma$ the hat matrix for the dispersion model. Finally we take $\mathit{df}\ (\boldsymbol{\lambda}_\mu) = \mathit{tr}\ (\boldsymbol{H}_\mu(\boldsymbol{\lambda}_\mu))$ and $\mathit{df}\ (\boldsymbol{\lambda}_\gamma) = \mathit{tr}\ (\boldsymbol{H}_\gamma(\boldsymbol{\lambda}_\gamma))$ to be the equivalent number of degrees of freedom for the mean and the dispersion model.

The smoothing parameters $\boldsymbol{\lambda}_\mu$ and $\boldsymbol{\lambda}_\gamma$ strongly influence the final appearance of the fits and their values are chosen before and updated within each iteration. Methods to choose optimal values are discussed in the next section. The final algorithm to estimate the mean and dispersion function originates from the one presented at the end of Section 2 and iterates between the following steps (see also Gijbels et al., 2010):

Step 0: initializing the algorithm. In our implementation we took initial values $\boldsymbol{\alpha}_\mu^{(0)}$ and $\boldsymbol{\alpha}_\gamma^{(0)}$ based on the sample mean and standard deviation of the Y-observations.

Step 1.a: selection of the smoothing parameter $\boldsymbol{\lambda}_\mu$. Select an optimal value for $\boldsymbol{\lambda}_\mu$ by minimizing RGCV($\boldsymbol{\lambda}_\mu$) or RAIC($\boldsymbol{\lambda}_\mu$), defined in (14) and (16) respectively, in Section 5. In our implementation we used the **optim** function of `R` to numerically minimize RGCV($\boldsymbol{\lambda}_\mu$) or RAIC($\boldsymbol{\lambda}_\mu$).

Step 1.b: estimation of $\boldsymbol{\alpha}_\mu$. For the chosen $\boldsymbol{\lambda}_\mu$ values, estimates of $\boldsymbol{\alpha}_\mu$ are obtained by solving (5). Once an estimate for $\boldsymbol{\alpha}_\mu$, and hence for $\boldsymbol{\mu}$, has been obtained, we compute the vector of deviances $\boldsymbol{d}$, which is used to estimate the dispersion function.

Step 2.a: selection of the smoothing parameter $\boldsymbol{\lambda}_\gamma$. Select an optimal value for $\boldsymbol{\lambda}_\gamma$ by minimizing RGCV($\boldsymbol{\lambda}_\gamma$) or RAIC($\boldsymbol{\lambda}_\gamma$), as described in Section 5.

Step 2.b: estimation of $\boldsymbol{\alpha}_\gamma$. For the chosen $\boldsymbol{\lambda}_\gamma$ values, estimates of $\boldsymbol{\alpha}_\gamma$ are obtained by solving (10). The vector of estimated $\boldsymbol{\gamma}$ values can now be computed and used in the Step 1.a of the next iteration.

The algorithm iterates between Steps 1 and 2 till convergence.

5. Smoothing Parameter Selection

Different methods for choosing the smoothing parameters exist. A standard procedure is to minimize the generalized cross-validation (GCV) criterion (Craven and Wahba, 1979):

$$GCV(\boldsymbol{\lambda}_\mu) = \mathbf{1}_n^T \frac{d(\boldsymbol{y}, \hat{\mu}(\boldsymbol{x}, \boldsymbol{\lambda}_\mu))}{(n - df(\boldsymbol{\lambda}_\mu))^2}, \tag{13}$$

where with $\hat{\mu}(\boldsymbol{x}, \boldsymbol{\lambda}_\mu)$ we want to emphasize that the estimate of $\mu(\boldsymbol{x})$ depends on $\boldsymbol{\lambda}_\mu$.

The criterion above is widely used in standard GAM to choose optimal values for $\boldsymbol{\lambda}_\mu$. Nevertheless, the criterion needs to be slightly modified when the dispersion function is considered to be no longer constant as in Gijbels and Prosdocimi (2011). Moreover, as mentioned by Cantoni and Ronchetti (2001b), the choice of $\boldsymbol{\lambda}_\mu$ via GCV will no longer work well in presence of outliers, even when the estimation procedure is robust. Here we propose to choose optimal values for $\boldsymbol{\lambda}_\mu$ via a robust version of GCV:

$$\mathrm{RGCV}(\boldsymbol{\lambda}_\mu) = \mathbf{1}_n^T \frac{\psi_q\left(d(\boldsymbol{y}, \hat{\mu}(\boldsymbol{x}, \boldsymbol{\lambda}_\mu))/\boldsymbol{\gamma}\right)}{(n - df(\boldsymbol{\lambda}_\mu))^2},$$

where $\boldsymbol{\gamma}$ once more denotes the vector of estimated $\gamma(\cdot)$ values, which is kept fixed when estimating the mean function.

The choice of smoothing parameters for the dispersion function is much less discussed in the literature. Gijbels and Prosdocimi (2011) propose an appropriate form of GCV for the choice of $\boldsymbol{\lambda}_\gamma$. Here, we propose to use a robustified version of this criterion:

$$\mathrm{RGCV}(\boldsymbol{\lambda}_\gamma) = \mathbf{1}_n^T \frac{\psi_q\left(d_\gamma(\boldsymbol{d}, \hat{\gamma}(\boldsymbol{x}, \boldsymbol{\lambda}_\gamma))\right)}{(n - df(\boldsymbol{\lambda}_\gamma))^2}\ , \tag{14}$$

where $d_\gamma(\boldsymbol{d}, \hat{\gamma}(\boldsymbol{x}, \boldsymbol{\lambda}_\gamma))$ is the vector of deviance residuals for the dispersion model, with $d_\gamma(\cdot,\cdot)$ the deviance function defined as (see 1)

$$d_\gamma(d, \gamma(\boldsymbol{x}_d)) = 2\int_{\gamma(\boldsymbol{x}_d)}^{d} \frac{d-t}{2t^2}dt\ . \tag{15}$$

Akaike's information criterion (AIC) is also often used to choose a smoothing parameter value (among others in the original, Eilers and Marx paper, 1996). Similarly to what is done for GCV, AIC can also be appropriately modified for a robust selection of the smoothing parameters for both the mean and the dispersion function estimation. The two criteria to be minimized are then:

$$\mathrm{RAIC}(\boldsymbol{\lambda}_\mu) = \mathbf{1}_n^T \psi_q\left(\frac{d(\boldsymbol{y}, \hat{\mu}(\boldsymbol{x}, \boldsymbol{\lambda}_\mu))}{\boldsymbol{\gamma}}\right) + 2\ df(\boldsymbol{\lambda}_\mu), \tag{16}$$

and

$$\mathrm{RAIC}(\boldsymbol{\lambda}_\gamma) = \mathbf{1}_n^T \psi_q\left(d_\gamma(\boldsymbol{d}, \hat{\gamma}(\boldsymbol{x}, \boldsymbol{\lambda}_\gamma))\right) + 2\ df(\boldsymbol{\lambda}_\gamma).$$

Simulation results not presented here show that in many cases the two criteria (AIC and GCV; and RAIC and RGCV) perform comparably.

In all the RGCV $(\cdot)$ and RAIC $(\cdot)$ criteria we take ψ_q to be the Huber function defined in (6) with tuning constant q. Taking $q = \infty$ corresponds to using the standard GCV and AIC criteria. In our applications we take q to be equal to c, the value of the tuning constant in the estimating procedure, but other choices could be done. Also, bounded functions other than the Huber function could be employed both in the estimation and in the smoothing parameters selection.

6. Simulation Study

We investigate the performance of the proposed method through a simulation study. We simulated 1000 datasets of size $n = 250$ coming from a Poisson-like distribution with mean function $\mu(x) = \exp\left(\eta_0 + \eta_1(x_1) + \eta_2(x_2)\right)$ and dispersion function $\gamma(x) = \exp\left(\xi_0 + \xi_1(x_1) + \xi_2(x_2)\right)$ with:

$$\begin{aligned} &\eta_0 = 1, \quad \eta_1(x_1) = 1.8\ \sin\left(3.4x_1^2\right), \quad \eta_2(x_2) = 1.1\cos(8x_2) \\ &\xi_0 = -0.35, \quad \xi_1(x_1) = 2.3\ \sin(2x_1)x_1^2, \\ &\xi_2(x_2) = -1.35(\sin(x_2)\exp(1.5 - 0.8x_2))\ . \end{aligned}$$

When summarizing the simulation results we present centered curves. This means that we subtract from a curve its average over the values taken in all datapoints, i.e., for example for $\eta_1(x_1)$ we present $\eta_1(x_1) - n^{-1}\sum_{i=1}^{n}\eta_1(x_{1,i})$. The covariates x_1 and x_2 are generated from two independent $U(0,$ $1)$ distributions. We simulated data in three different settings, contaminating a growing percentage (0%, 3%, and 5%) of observations, uniformly located in $0.1 < x_1 < 0.2$ and $0.8 < x_2$ < 0.9, with Y-observations drawn from a discrete $U(25, 28)$ distribution.

In this Poisson-type modeling we take $V(\mu) = \mu$, and logarithmic link functions for both the mean and the dispersion: $\eta(\cdot) = \log\left(\mu(\cdot)\right)$ and $\xi(\cdot) = \log\left(\gamma(\cdot)\right)$. Also, we take $w(\cdot) = 1$.

For each simulated dataset we estimated both the mean and the dispersion function via the Robust extended GAM procedure proposed in Section 3 choosing the smoothing parameters both via the standard GCVs and the robust versions proposed in Section 5. We compare the performance of the proposed methods with the nonrobust extended GAM of Gijbels and Prosdocimi (2011) and the standard GAM with mean function estimation only. In this way we are able to investigate the differences between both robust and nonrobust methods, and between models in which only the mean function is estimated and the Double models in which both the mean and the dispersion functions are estimated.

For a given dataset we evaluate the performance of the estimation procedure via the approximate integrated squared error (AISE):

$$\mathrm{AISE} = \frac{\displaystyle\sum_{i=1}^{n}(\hat{f}(\boldsymbol{x}_{d,i}) - f_{\mathrm{true}}(\boldsymbol{x}_{d,i}))^2}{\displaystyle\sum_{i=1}^{n}(f_{\mathrm{true}}(\boldsymbol{x}_{d,i}))^2},$$

with $\hat{f}(\cdot)$ the estimated function and $f_{\mathrm{true}}(\cdot)$ the true function. In Figures 1 and 2 we summarize the results for the 0% and the 3% contamination setting. The results for the 5% contamination setting (not shown here) give results similar to these for the lower contamination setting. In each plot we show the

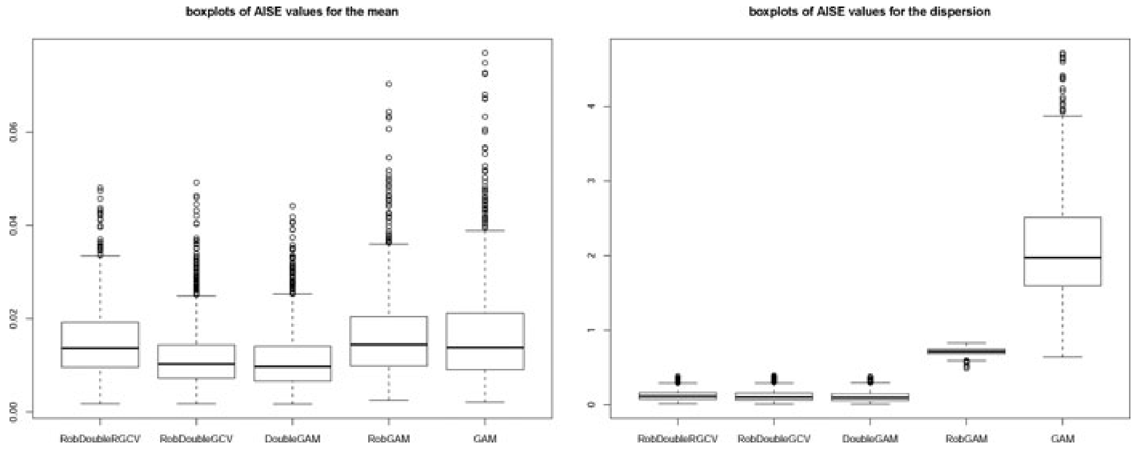

Figure 1. The 0% contamination setting: boxplots of AISE values for the mean (left) and dispersion (right) estimation.

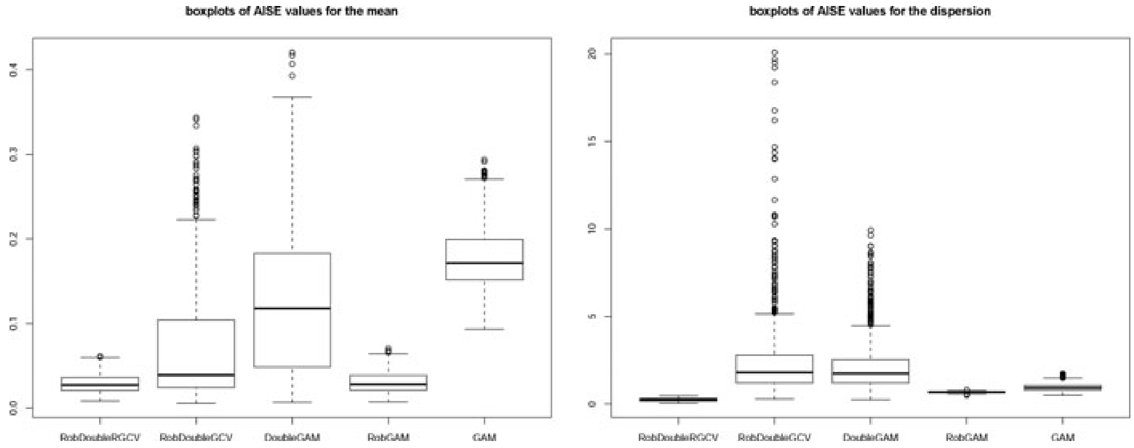

Figure 2. The 3% contamination setting: boxplots of AISE values for the mean (left) and dispersion (right) estimation.

boxplots of the AISE values for the mean and the dispersion function estimation for the different estimation procedures (boxplots from left to right):

- `RobDoubleRGCV`: the proposed robust estimation of mean and dispersion function, with smoothing parameter chosen via RGCV;
- `RobDoubleGCV`: the proposed robust estimation of mean and dispersion function, with smoothing parameter chosen via standard GCV;
- `DoubleGAM`: the nonrobust estimation of mean and dispersion function as in Gijbels and Prosdocimi (2011), with smoothing parameter chosen via GCV;
- `RobGAM`: the robust GAM estimation of the mean function, with smoothing parameter chosen via GCV;
- `GAM`: the standard GAM estimation of the mean function, with smoothing parameter chosen via GCV;

From Figure 1 it is seen that in case of no contamination the nonrobust and the robust methods have similar behavior, with nonrobust methods performing slightly better. We note that not taking into account the variability in the dispersion function can have a bad influence on the mean estimation as well. From Figure 2 we can see that as soon as outliers are present in the data, the nonrobust methods perform worse and worse: we even get lower AISE values for the dispersion when the dispersion function is not estimated rather than estimated in a nonrobust way. Note also that choosing the smoothing parameters with a robust criterion plays a crucial role: when using robust methods the optimal smoothing parameters need to be chosen with a robust criterion as well. In general the proposed method (`RobDoubleRGCV`) seems to perform quite well: not only the median AISE values are much lower than the ones of the other methods, but we also see little variability.

In Figure 3, we show a dataset simulated under the 3% contamination setting, together with nonrobust and robust estimates for the mean and dispersion functions. It is clearly seen that the robust methods are less affected by the presence of outliers.

7. Real Data Examples

In all examples we use a logarithmic link for the dispersion function, i.e., $\xi(\cdot) = \log(\gamma(\cdot))$, and take $w(\cdot) = 1$, i.e., no weighting function is employed to correct for leverage points.

7.1 *Influenza-Like Illness (ILI) Visits in the United States*

Alimadad and Salibian-Barrera (2011) study how the weekly counts of ILI visits in the United States change in the course of the influenza season (which lasts 33 weeks, from week 40 to the end of week 20 of the next year). They analyze data

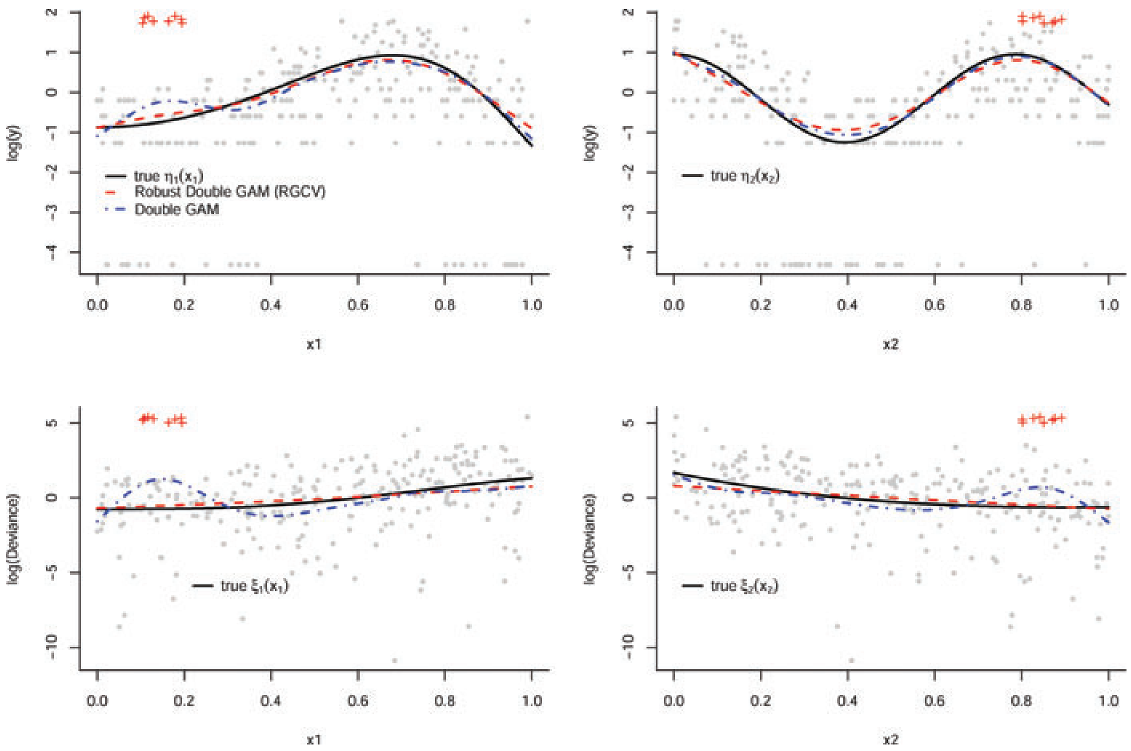

Figure 3. A simulated dataset from the 3% contamination setting. Outliers are indicated with crosses. Robust (dashed lines) and nonrobust (dashed–dotted lines) estimates, together with the true functions (solid lines), for both the mean (top panels) and the dispersion (lower panels) functions. This figure appears in color in the electronic version of this article.

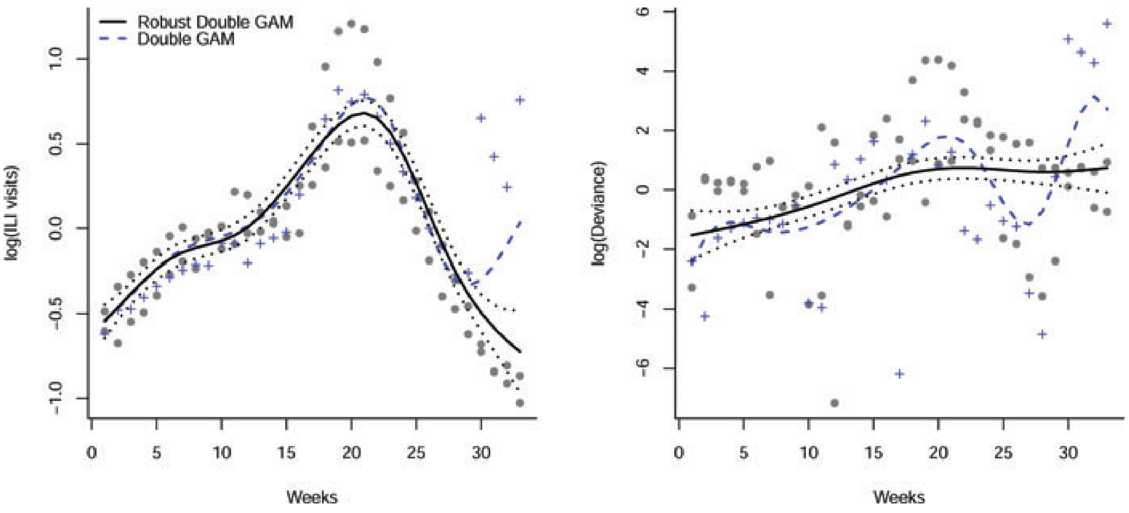

Figure 4. ILI visits in the United States (crosses indicate data of the 2008/2009 season): robust double GAM (solid line) with confidence intervals (dotted lines), and standard double GAM (dashed line) fits for the mean and dispersion function. This figure appears in color in the electronic version of this article.

regarding the influenza seasons of 2006/2007, 2007/2008, and 2008/2009. During the last weeks of the 2008/2009 season the H1N1 flu started spreading, and this resulted in a higher number of visits. Therefore, they suggest to analyze the data using robust methods which are less affected by the presence of high numbers of visits. What they do not take into account is that the variability of the data also seems to be changing over the weeks within the season. We propose that, to analyze these data properly, not only the mean but also the dispersion function should be estimated. The presence of extreme points in the data, suggests indeed that we should apply robust methods. For the mean modeling we take $V(\mu) = \mu$ and a logarithmic link function $\eta(\cdot) = \log\,(\mu(\cdot))$, as we would do for a Poisson regression.

In the left panel of Figure 4 we present the centered (log) data with a robust and a nonrobust fit of the mean, while the

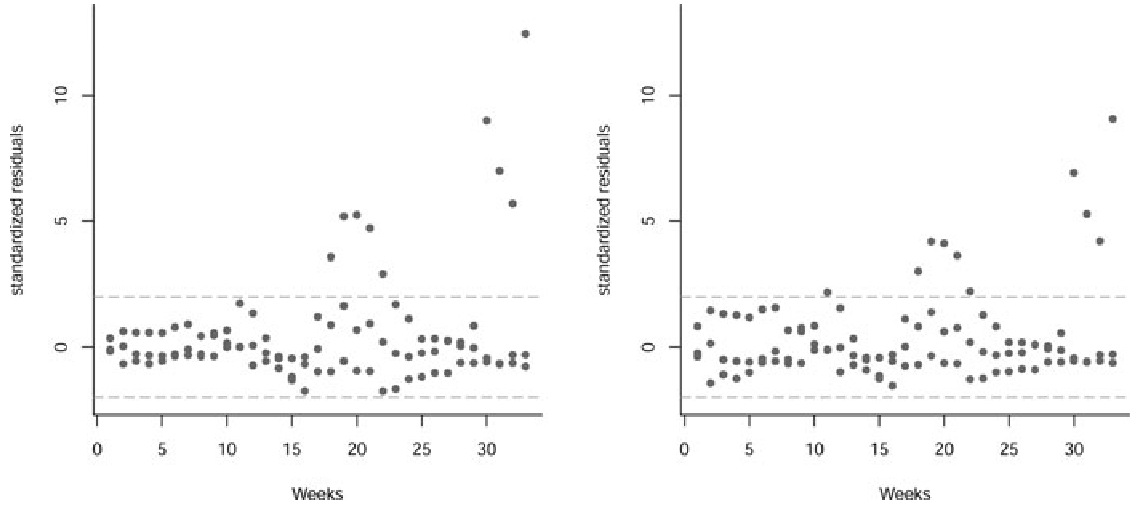

Figure 5. Standardized residuals for the fits to the ILI visits. On the left the residuals obtained when taking the dispersion as a constant, on the right the ones obtained when taking the dispersion to be a function of the covariate.

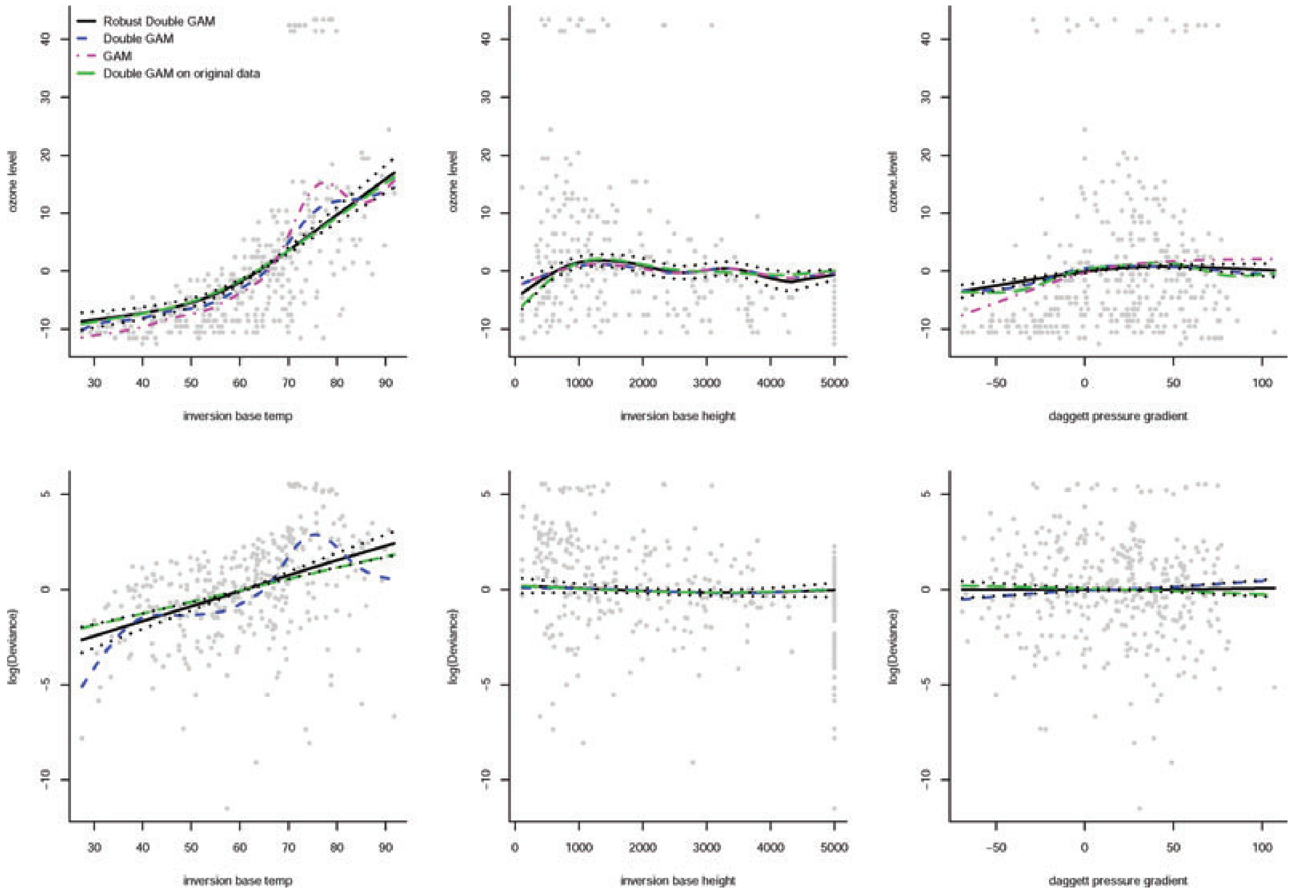

Figure 6. The ozone data with outliers. The robust double GAM (solid line) with confidence intervals (dotted lines) and the nonrobust double GAM (dashed line) fits for the mean and the dispersion (top and bottom panels, respectively). The long-dashed line represents the fit from a nonrobust double estimation of the mean and dispersion function on the original data. The dotted–dashed lines are standard GAM fits (top panels only). This figure appears in color in the electronic version of this article.

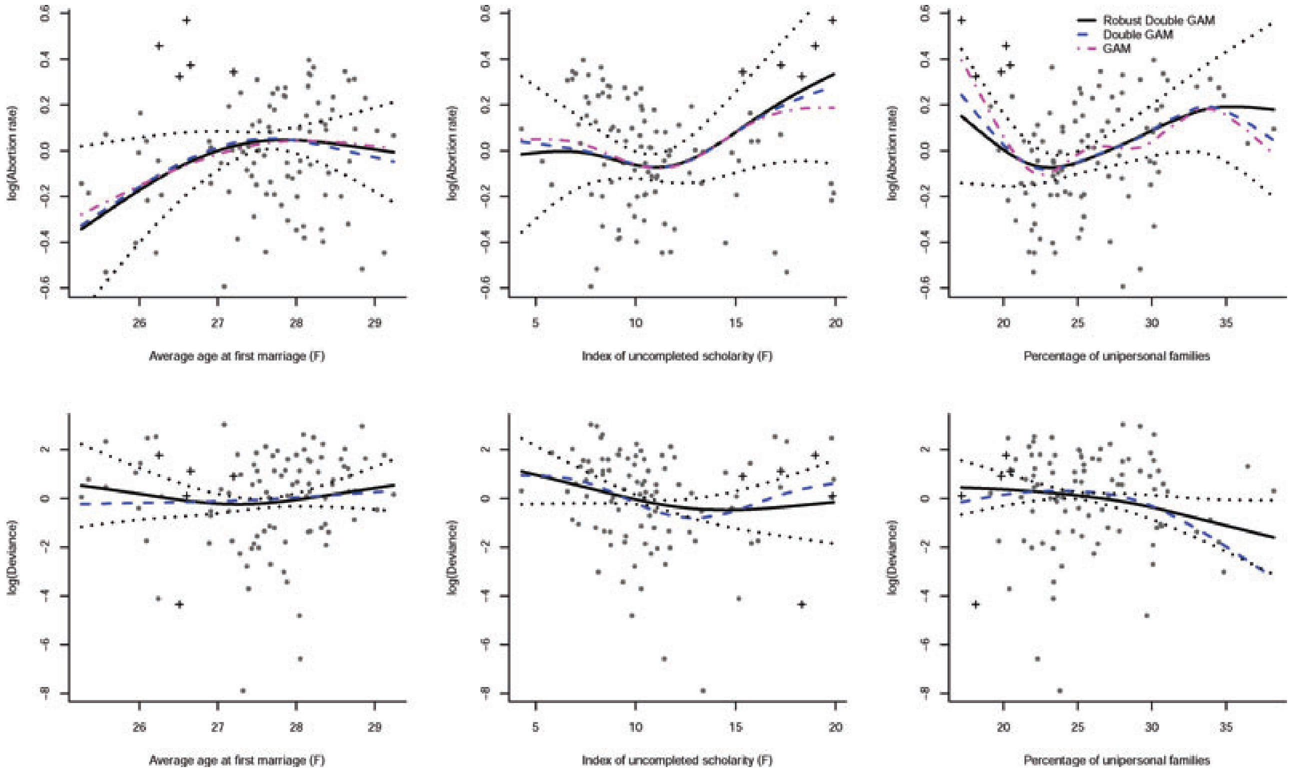

Figure 7. The abortion data: robust double GAM (solid lines) with confidence intervals (dotted lines), and standard double GAM (dashed lines) fits for the mean and dispersion function (top and bottom panels, respectively). Crosses indicate the provinces of Puglia. The dotted–dashed lines are standard GAM fits (top panels only). This figure appears in color in the electronic version of this article.

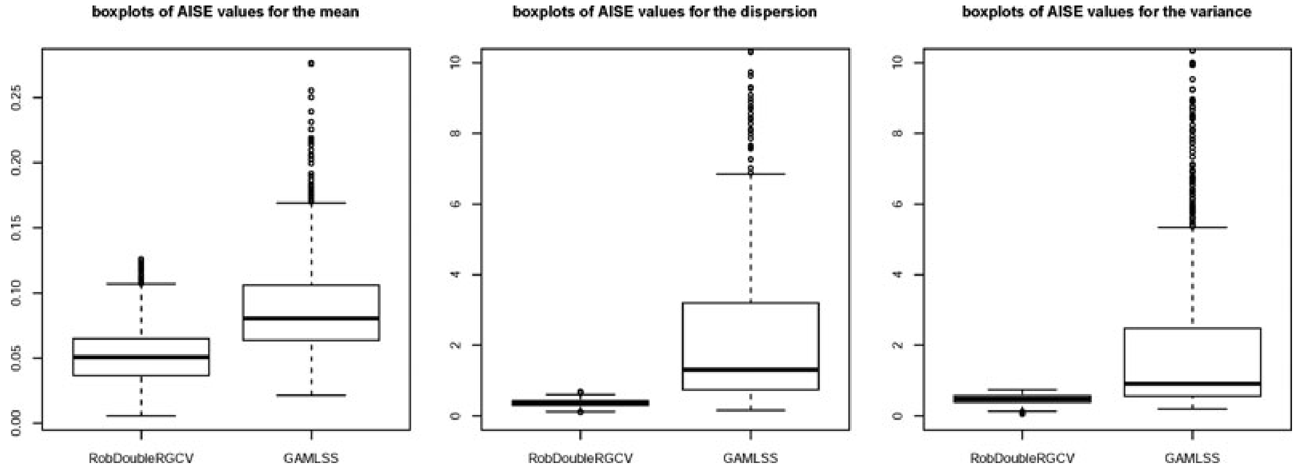

Figure 8. Overdispersed Poisson data: double robust GAM and GAMLSS estimation. For both estimation procedures boxplots of AISE values for the mean (left) dispersion (center) and variance (right) estimation are displayed.

centered (log) deviance residuals with a fit for the $\xi(weeks)$ component are shown in the right panel. We see that the robust methods are less affected by the presence of extreme values in the data, both for the mean and the dispersion estimation. The dotted lines in the plots are robust confidence intervals. They are obtained from the following asymptotic distributions, resulting from the theory of M-estimation (see Hampel et al., 1986),

$$\hat{\boldsymbol{\alpha}}_\mu \sim N\left(\boldsymbol{\alpha}_\mu, \tilde{\boldsymbol{W}}_\mu^{-1}\boldsymbol{Q}_\mu\tilde{\boldsymbol{W}}_\mu\right) \quad \text{with}$$
$$\boldsymbol{Q}_\mu = \bar{E}_n\left[\Psi_s(\boldsymbol{y},\mu(\boldsymbol{x}))(\Psi_s(\boldsymbol{y},\mu(\boldsymbol{x})))^T \mid \boldsymbol{X}_d = \boldsymbol{x}_d\right]$$
$$\hat{\boldsymbol{\alpha}}_\gamma \sim N\left(\boldsymbol{\alpha}_\gamma, \tilde{\boldsymbol{W}}_\gamma^{-1} Q_\gamma \tilde{\boldsymbol{W}}_\gamma^{-1}\right) \quad \text{with}$$
$$\boldsymbol{Q}_\gamma = \bar{E}_n\left[\Psi_t(\boldsymbol{d},\gamma(\boldsymbol{x}))(\Psi_t(\boldsymbol{d},\gamma(\boldsymbol{x})))^T \mid \boldsymbol{X}_d = \boldsymbol{x}_d\right],$$

where $\tilde{\boldsymbol{W}}_\mu$ and $\tilde{\boldsymbol{W}}_\gamma$ are the matrices defined in (11) and (12). Because for a large part of the covariate domain the confidence

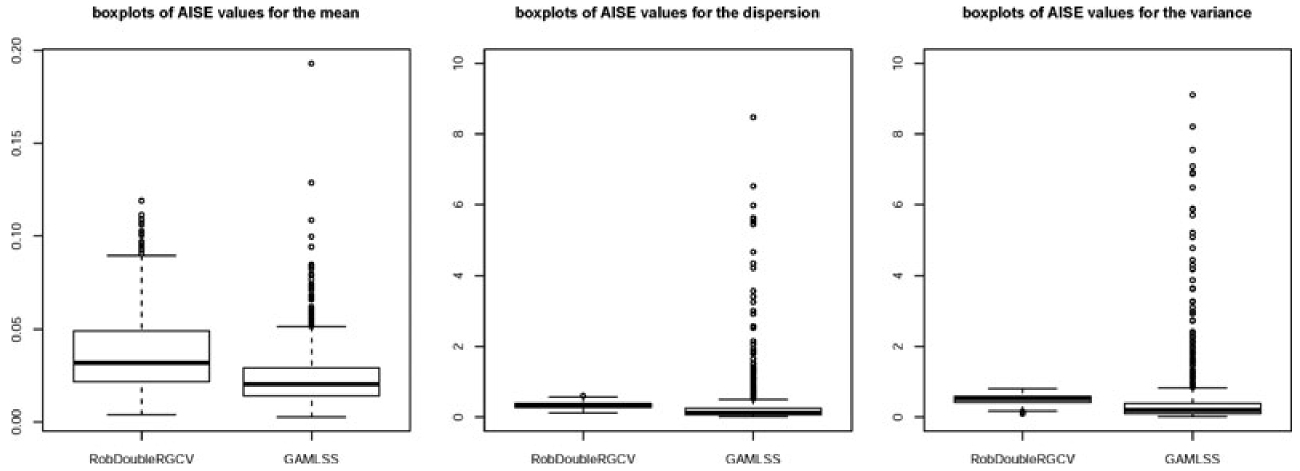

Figure 9. PIG data: double robust GAM and GAMLSS estimation. For both estimation procedures boxplots of AISE values for the mean (left) dispersion (center) and variance (right) estimation are displayed.

intervals for both functions do not contain the null value, we have an indication that a functional relationship between the mean/dispersion and the covariates is indeed present in the data. The dispersion of the model should then indeed be estimated as a varying function: in Figure 5 we see the standardized Pearson residuals we would obtain from a robust model with a constant dispersion ($r_P = (y - \mu)/(\phi\ \mu)$) and the ones we obtain when modeling also the dispersion in a robust way ($r_P = (y - \mu)/(\phi\ \gamma\ \mu)$). Clearly the shape present in the residuals in the left plot diminishes when estimating the dispersion and we also notice how the outlying points at the extreme right of the plot have now lower residuals.

A consequence of estimating the dispersion is that points which correspond to extremely high ILI visit counts and that have very large residuals in the left plot are scaled by the estimated dispersion function, which has higher values in the area where the more extreme counts are observed. Indeed in the right panel of Figure 5 the points with higher residuals are less sticking out, they appear to be less extreme. If we think that the process under study is prone to have parts of larger variability we should estimate the dispersion function and allow the process to be more variable in some parts, rather than interpreting high counts automatically as outliers.

7.2 *The Ozone Data*

In Figure 6 (top panels) data on the ozone level in Upland, California in 1976 (see Breiman and Friedman, 1985), are depicted. We are interested in modeling the ozone level as a flexible function of the inversion base temperature, the inversion base height and the daggett pressure gradient. To illustrate what the effect of outliers can be, we replaced 5% of the data by outliers scattered uniformly around (55,58). The datapoints to be substituted by outliers were selected among those with inversion base temperature values between 70 and 80. We take $V(\mu) = 1$ and the identity link $\eta(\cdot) = \mu(\cdot)$, the default choice in standard GAM for data assumed to be normal.

In Figure 6, we see how the robust techniques are much less influenced by the presence of the outliers in the data. The robust estimates obtained for the mean and the dispersion function resemble indeed much more the shape we would get when outliers are not present in the data. It is interesting to note that for the standard GAM (only estimation of the mean) the outliers in the data have a stronger impact on the final estimate than for the Double GAM. This is because one of the consequences of allowing for heteroscedasticity and to estimate the dispersion is that high values of the dispersion corresponding to the outlying points, result in those outlying points receiving a lower weight in the mean estimation (see (5)). They therefore influence less the final shape of the mean estimate. Estimating the dispersion in a nonrobust way can thus possibly be beneficial for obtaining better estimates for the mean, but when the interest lies in delivering robust estimates for both the mean and the dispersion functions, we advise using the robust methods presented in the previous sections.

7.3 *The Italian Abortion Data*

In Figure 7 data on the induced abortion rate for each Italian province are shown. This dataset was previously analyzed in Gijbels and Prosdocimi (2011). We are interested in studying how the abortion rate in the 98 Italian provinces changes as a function of the following socioeconomical covariates:

- The average age at first marriage for women;
- The index of nonfinished compulsory education for the female population between 15 and 52; and
- The percentage of unipersonal families.

As in Section 7.1, we take $V(\mu) = \mu$ and $\eta(\cdot) = \log\ (\mu(\cdot))$ in this example. Smoothing parameters for the standard and the robust fit are selected, respectively, via AIC and RAIC.

We know that the highest values present in the dataset are coming from the five provinces in one region (Puglia) in which the health care system, especially concerning induced abortion, is of higher quality than the one of the neighboring regions. It is suspected that the high abortivity rates observed in this region are due more to women from outside who travel to undergo the operation rather than from a real higher abortion rate among the women of the region. By using a robust estimate for the mean and dispersion value we are assured

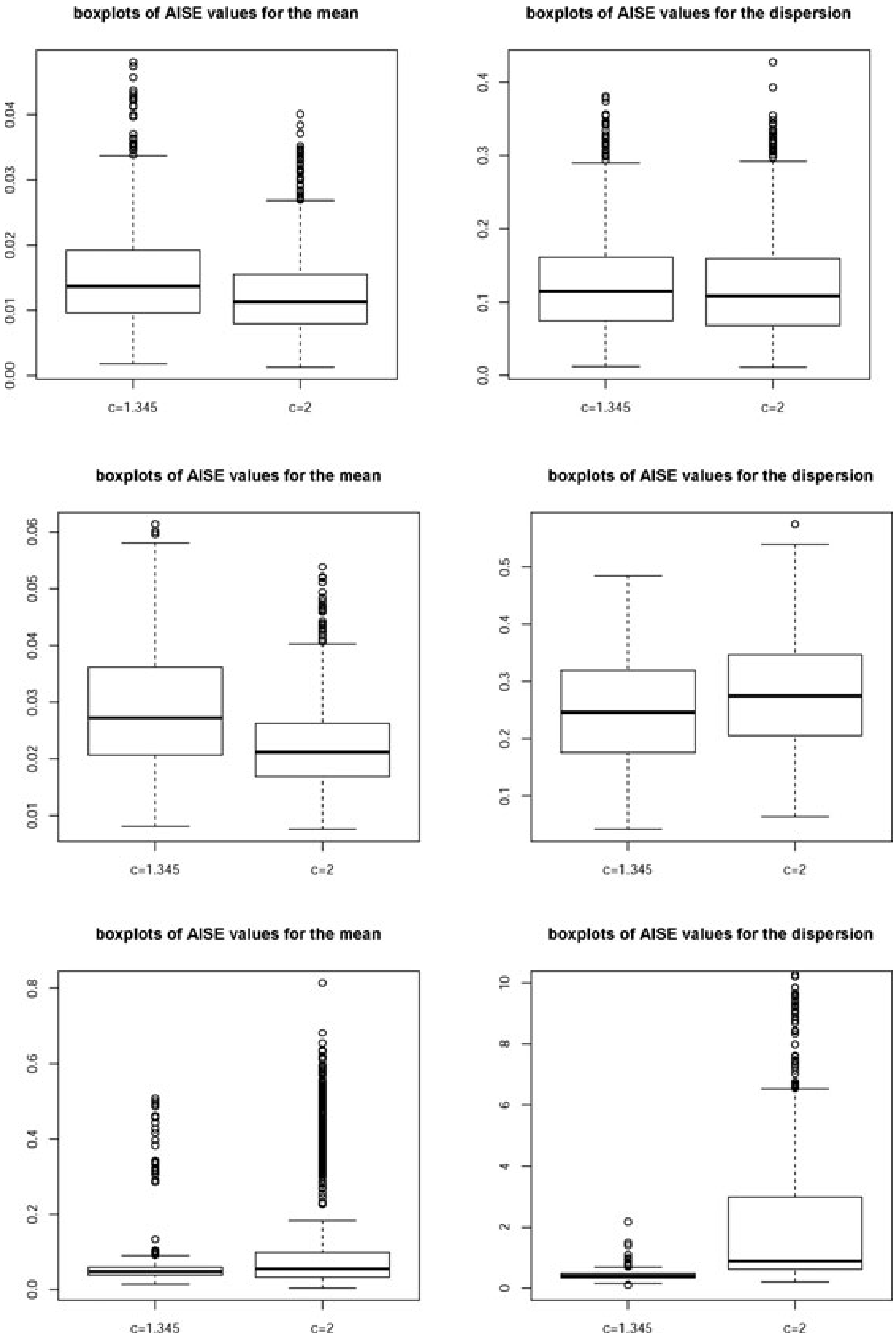

Figure 10. Sample size $n = 250$. Boxplots of AISE values for the mean (left) and dispersion (right) estimation in the 0%, 3%, and 5% contamination settings (top, middle and bottom panels).

that these outlying points will have less effect on the final estimates, as can be seen in Figure 7: indeed the robust fits are less affected by the presence of extreme points. Again, we see that the standard GAM estimate for the mean is most severely affected by extreme points.

8. Comparisons with Other Modeling Choices

In this section, we discuss different modeling choices and evaluate their performances. First we compare our modeling strategy with the GAMLSS approach of Rigby and Stasinopoulos (2005). Then, via an additional simulation study, we investigate how different choices of the tuning constant c affect the final estimate for the mean and dispersion function.

8.1 *A Comparison with a* GAMLSS Approach

As mentioned in Section 1, a possible different modeling approach to estimate the dispersion function of overdispersed data is to start from specific data-generating distributions which extend the standard distributions belonging to the exponential family of distributions. For example, to model the

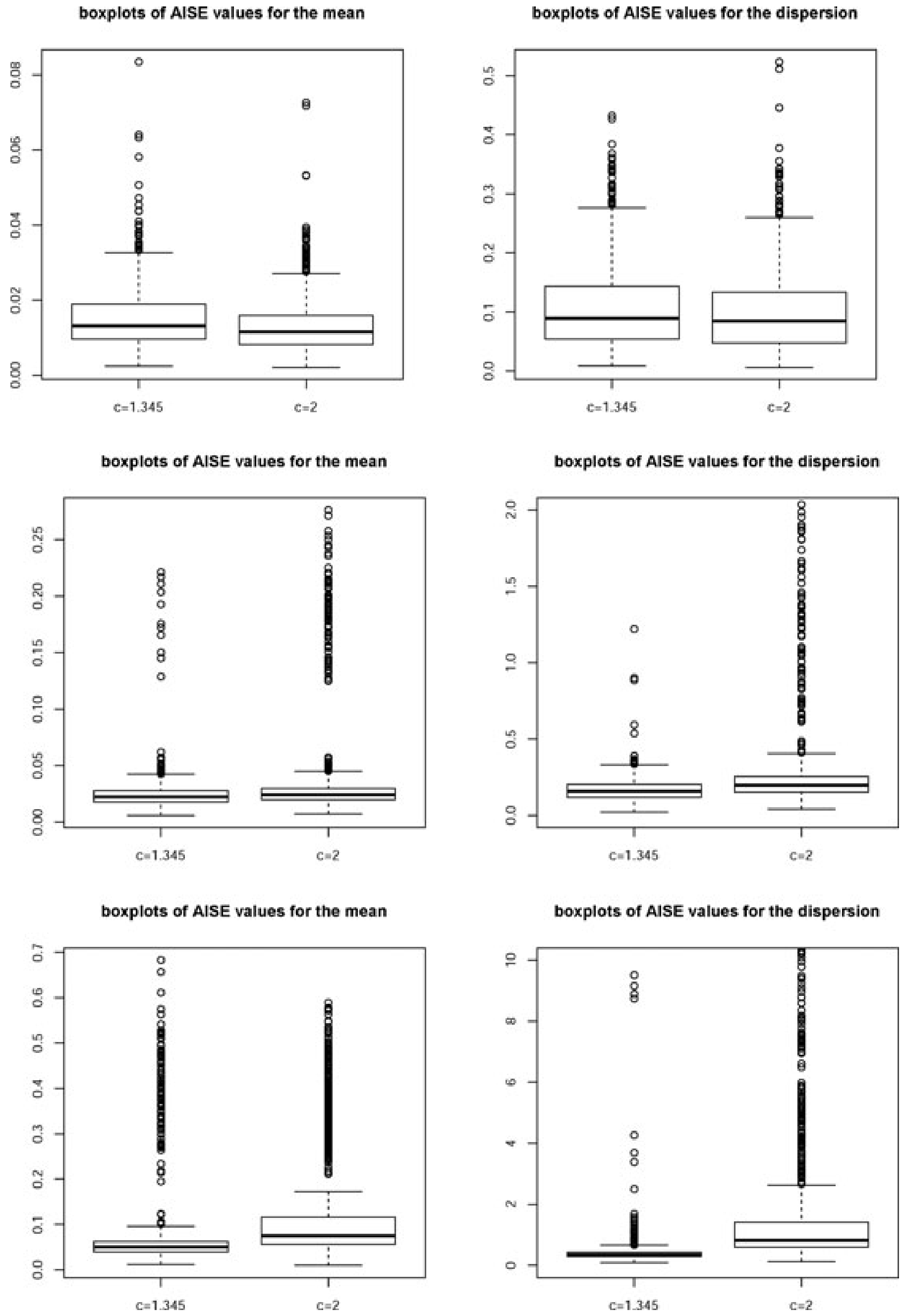

Figure 11. Sample size $n = 500$. Boxplots of AISE values for the mean (left) and dispersion (right) estimation in the 0%, 3%, and 5% contamination settings (top, middle and bottom panels).

mean and the dispersion function of count data, one could use a negative binomial distribution, and via the GAMLSS framework obtain smooth estimates for the mean and the dispersion function. These methods though, are also affected by outliers in the data, and as far as we know, no work has been done on robustifying the GAMLSS approach in a fashion like the one proposed in this article. In the presence of outliers, a possible approach within the GAMLSS class would be to use overdispersed long-tailed distributions, like the Poisson inverse Gaussian (PIG; Dean, Lawless, and Willmot, 1989). When $Y \mid \boldsymbol{X}_d = \boldsymbol{x}_d \sim$ PIG $(\lambda(\boldsymbol{x}_d), \tau(\boldsymbol{x}_d))$, we have that $E[Y \mid \boldsymbol{X}_d = \boldsymbol{x}_d] = \lambda(\boldsymbol{x}_d)$ and $Var\ [Y \mid \boldsymbol{X}_d = \boldsymbol{x}_d] = \lambda(\boldsymbol{x}_d)(1 + \lambda(\boldsymbol{x}_d)\ \tau(\boldsymbol{x}_d))$. Because $\tau(\boldsymbol{x}_d) > 0$ it is clear that this distribution can only be used in the case of overdispersed data. To have a fair comparison between the PIG approach and the methods presented in this article, we performed a new simulation study. The dispersion function used to generate the data of Section 6 in fact changes from underdispersion to overdispersion so assuming a PIG distribution for those data would not be realistic. We therefore generated data following either a PIG $(\lambda(x_1, x_2), \tau(x_1, x_2))$ or a 3% contaminated overdispersed Poisson distribution with mean $\lambda(x_1, x_2)$ and dispersion $\gamma(x_1,$

x_2). We took the mean to be as the one described in Section 6, $\lambda(x_1, x_2) = \exp\,(\eta_0 + \eta_1(x_1) + \eta_2(x_2))$, whereas for the dispersion function we took $\tau(x_1, x_2) = \exp\,(\xi_0^* + \xi_1^*(x_1) + \xi_2^*(x_2))$ and $\gamma(x_1, x_2) = (1 + \lambda(x_1, x_2)\ \tau(x_1, x_2))$ with

$$\xi_0^* = -1.75, \quad \xi_1^*(x_1) = 1.8\,\sin(2x_1)x_1^2,$$
$$\xi_2^*(x_2) = -0.95(\sin(x_2)\exp(1.5 - 0.8x_2)),$$

which results into a dispersion function $\gamma(\cdot,\cdot)$ similar in shape as that in Section 6, but now larger than one everywhere (only overdispersion). The data generation and the model fitting for the PIG data has been done using the `R` package `gamlss` of Stasinopoulos and Rigby (2007). In Figures 8 and 9, we present boxplots of the AISE values for the mean, the dispersion, and the variance estimation of the two estimation procedures. Clearly each one of the two methods performs (much) better in the case in which the (noncontaminated) data are generated from the distribution assumed in the estimation procedure, although we observe larger proportions of extremely high AISE values for the dispersion estimation when using the GAMLSS approach. Despite the fact that in our setting the added 3% outlying points do *not* fit the model, contrary to the setup when we draw data from the PIG model, we notice an overall very good performance of the proposed method.

8.2 *The Tuning Parameter c*

In the numerical studies in Sections 6, 7, and 8.1 we took the tuning parameter $c = 1.345$ for both the mean and the dispersion function estimation. The reasoning behind this particular choice is that, if using M-estimation for a location parameter on data coming from a normal distribution, using $c = 1.345$ would lead to an estimate that is 95% model efficient compared to the maximum likelihood estimate. Considering that the Pearson residuals are standardized quantities on which we apply the $\psi_c(\cdot)$ function, using a c value that has good properties for normal data seems to be a reasonable choice. Another sensible choice for both the mean and the dispersion function estimation would be to take $c = 2$, which, in case of normal data, allows for approximately 5% of the datapoints to be downweighted. The reasoning behind this choice is that we would downweight only values that are not very likely to be observed; to give an indication taking $c = 1.345$ would downweight approximately 20% of normally distributed datapoints.

We studied the effect of the choice of c on the final result via some simulation study. We repeated the simulation study of Section 6 (sample size $n = 250$) using this time a value $c = 2$ for the estimation of both the mean and the dispersion. In Figure 10 we compare the AISE values for both the mean and the dispersion function estimation for the different c-values. In Figure 11 simulation results are presented for sample size $n = 500$. The difference in performance when using the two different tuning parameter values is mostly visible in the case of higher levels of contamination, indicating that when the percentage of outliers in the data is relatively high we need to use a lower tuning constant c to ensure that the estimation procedure is indeed not too much influenced by extreme points. It should be noted that for the 5% contamination setting the worsening of the estimation performance affects the estimate of the dispersion more than the estimation of the mean. This is mostly because when a poorer estimate of the mean function is obtained, the deviance residuals which are computed based on this mean estimation will be less reliable and will not be a good response variable for the dispersion estimation. We would therefore advice the use of $c = 1.345$, to make sure that both functions estimation are not influenced by outlying points.

Acknowledgements

The authors thank the co-editor, an associate editor, and the reviewers for their very valuable comments, which led to a considerable improvement of the article. Support from the GOA/07/04-project of the Research Fund KULeuven is gratefully acknowledged. The second author also acknowledges support from the FWO-project G.0328.08N and the IAP research network P6/03 of the Federal Science Policy, Belgium.

References

Alimadad, A. and Salibian-Barrera, M. (2011). An outlier-robust fit for generalised additive models with applications to outbreak detection. *Journal of the American Statistical Association*, to appear.

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* **80,** 580–619.

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

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

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross validation. *Numerische Mathematik* **31,** 377–403.

Dean, C., Lawless, J. F., and Willmot, G. E. (1989). A mixed Poisson-inverse-Gaussian regression model. *Canadian Journal of Statistics* **17,** 171–181.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11,** 89–121.

Gijbels, I. and Prosdocimi, I. (2011). Smooth estimation of mean and dispersion function in extended generalized additive models with application to Italian induced abortion data. *Journal of Applied Statistics*, to appear. DOI: 10.1080/02664763.2010.550039.

Gijbels, I., Prosdocimi, I., and Claeskens, G. (2010). Nonparametric estimation of mean and dispersion functions in extended generalized linear models. *Test* **19,** 580–608.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*. West Sussex, Chichester, UK: Wiley & Sons.

Hinde, J. and Demétrio, C. G. B. (1998). Overdispersion: Models and estimation. *Computational Statistics & Data Analysis* **27,** 151–170.

Marx, B.D and Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis* **28,** 193–209.

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

Nelder, J. A. and Pregibon, D. (1987). An extended quasi likelihood function. *Biometrika* **74,** 221–232.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics* **54,** 507–554.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik* **10,** 177–183.

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

Ruppert, D., Wand, M., and Carroll, R. J. (2009). Semiparametric regression during 2003–2007. *Electronic Journal of Statistics* **3,** 1193–1256.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, Series B* **47,** 1–52.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software* **23,** 1–46.

Wahba, G. (1980). Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy dataA. In *Approximation Theory III*, W. Cheney (ed), 905–912. New York: Academic Press.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61,** 439–447.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, Florida: Chapman and Hall/CRC.

Received September 2010*. Revised April* 2011*.*
Accepted April 2011*.*