KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Robust estimation of mean and dispersion functions in extended generalized additive models

C. Croux, I. Gijbels and I. Prosdocimi

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 1017

Robust estimation of mean and dispersion functions in Extended Generalized Additive Models

C. Croux[1,3,4], I. Gijbels[2,3] and I. Prosdocimi[2,3]

[1] Faculty of Business and Economics
[2] Department of Mathematics
[3] Leuven Statistics Research Center (LStat)
Katholieke Universiteit Leuven, Belgium;
[4] Tilburg University, The Netherlands.

3rd September 2010

Abstract

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 Quasi-Likelihood (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 paper 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.

KEYWORDS: dispersion, generalized additive modelling, mean regression function, quasi-likelihood, M-estimation, P-splines, robust estimation.

1 Introduction

Statistical modelling aims at describing how a phenomenon of interest changes with respect to some other quantities. Generally most of the modelling 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 GLM's by Cantoni and Ronchetti (2001a). Recently Alimadad and Salibian-Barrera (2009) propose a method for robust estimation of GAM.

In this paper 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 pursuit 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 under- dispersion 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 *et al.* (2010) or Hinde and Demétrio (1998) for a review of possible methods. In this work we use the Extended Quasi-likelihood approach (Nelder and Pregibon (1987) and McCullagh and Nelder (1989)) to obtain estimates for the dispersion function. Just as the standard Quasi-likelihood, Extended Quasi-likelihood 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.

The remainder of the paper is organized as follows: in Sections 2 and 3 standard and robust estimation methods within the Extended Quasi-likelihood 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 for both the robust and the standard Generalized Additive Models. We show the performance of the proposed methods via a simulation study and real data examples in Sections 6 and 7 respectively.

2 Extended Quasi-likelihood

In order to write a likelihood function for a certain model, we need to make assumptions on the distribution of the process of interest. In the Quasi-likelihood 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 Quasi-likelihood function, which shares key properties with a likelihood function, but can be obtained with weaker assumptions (Wedderburn (1974)). We consider

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

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 $(Y|\boldsymbol{X}_d = \boldsymbol{x}_d)$ via a link function $g(\cdot)$, and is modelled as a linear combination of some generic functions of the covariates: $\eta(\boldsymbol{x}_d) = g(\mu(\boldsymbol{x}_d)) = \alpha_{\mu,0} + \eta_1(x_1) + \ldots + \eta_d(x_d)$. Furthermore, the quasi-deviance function

$$d(y, \mu(\boldsymbol{x}_d)) = -2Q(y, \mu(\boldsymbol{x}_d)) = 2\int_{\mu(\boldsymbol{x}_d)}^{y} \frac{y-t}{\phi V(t)} dt \;, \tag{2.1}$$

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

In the Quasi-likelihood setting the relationship between the variance of $(Y|\boldsymbol{X}_d = \boldsymbol{x}_d)$ 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

$$E[Y|\boldsymbol{X}_d = \boldsymbol{x}_d] = \mu(\boldsymbol{x}_d) \;\text{ and }\; \mathrm{Var}[Y|\boldsymbol{X}_d = \boldsymbol{x}_d] = \phi\gamma(\boldsymbol{x}_d) V(\mu(\boldsymbol{x}_d)) \;, \tag{2.2}$$

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

$$E[d(Y,\mu)|\boldsymbol{X}_d = \boldsymbol{x}_d] = \gamma(\boldsymbol{x}_d) \;\text{ and }\; \mathrm{Var}[d(Y,\mu)|\boldsymbol{X}_d = \boldsymbol{x}_d] = 2\gamma^2(\boldsymbol{x}_d) \;. \tag{2.3}$$

The structure used to model the dispersion is a mirror image of the mean modelling: the quasi-deviance 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 $(d(Y,\mu)|\boldsymbol{X}_d = \boldsymbol{x}_d) \sim \gamma(\boldsymbol{x}_d)\chi^2_1$. We introduce here a second monotone and twice differentiable function $\xi(\cdot)$, which transforms the expected value of $(d(Y,\mu)|\boldsymbol{X}_d = \boldsymbol{x}_d)$ via a link function $h^{-1}(\cdot)$, and is modelled as a linear combination of some generic functions of the covariates: $\xi(\boldsymbol{x}_d) = h^{-1}(\gamma(\boldsymbol{x}_d)) = \alpha_{\gamma,0} + \xi_1(x_1) + \ldots + \xi_d(x_d)$.

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

We propose to estimate the $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_\gamma$ in a two-steps procedures which alternates between the estimation of $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_\gamma$ as in McCullagh and Nelder (1989). For a given 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 matrix $\boldsymbol{B}_\mu = [\boldsymbol{1}_n\ \boldsymbol{x}]$ where we denote with $\boldsymbol{1}_n = (1, \ldots, 1)^T$ the unit vector of length n. We model $\eta(\boldsymbol{x}) = \boldsymbol{B}_\mu\boldsymbol{\alpha}_\mu$. The Extended Quasi-likelihood estimator of $\boldsymbol{\alpha}_\mu$ is then the solution to

$$\boldsymbol{B}_\mu^T \left(\frac{\boldsymbol{y} - \mu(\boldsymbol{x})}{\phi\boldsymbol{\gamma} V(\mu(\boldsymbol{x}))} \, \frac{d\mu}{d\eta}(\boldsymbol{x}) \right) = \boldsymbol{0} \ , \tag{2.4}$$

where $\boldsymbol{0}$ is the null vector, $\mu(\boldsymbol{x}) = (\mu(\boldsymbol{x}_{d,1}, \boldsymbol{\alpha}_\mu), \ldots, \mu(\boldsymbol{x}_{d,n}, \boldsymbol{\alpha}_\mu))^T$ is the vector of computed $\mu(\cdot)$ values for each data point and $V(\mu(\boldsymbol{x}))$ is the vector of values of the $V(\cdot)$ function evaluated at each $\mu(\boldsymbol{x})$ point. 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. By $\boldsymbol{\gamma}(\boldsymbol{x}) = (\gamma(\boldsymbol{x}_{d,1}, \boldsymbol{\alpha}_\gamma), \ldots, \gamma(\boldsymbol{x}_{d,n}, \boldsymbol{\alpha}_\gamma))^T$ we denote the vector of $\gamma(\cdot)$ values for each data point. While we estimate $\boldsymbol{\alpha}_\mu$ we take $\boldsymbol{\alpha}_\gamma$, and therefore $\boldsymbol{\gamma}$, to be fixed. Once that $\boldsymbol{\alpha}_\mu$, and consequently $\mu(\boldsymbol{x})$, is estimated by solving (2.4), we compute the vector of deviances $\boldsymbol{d} = d(\boldsymbol{y}, \boldsymbol{\mu}) = (d(y_1, \mu(\boldsymbol{x}_{d,1}))^T, \ \ldots, \ d(y_n, \mu(\boldsymbol{x}_{d,n}))^T)^T$, where $\boldsymbol{\mu} = \mu(\boldsymbol{x}) = (\mu(\boldsymbol{x}_{d,1}), \ldots, \mu(\boldsymbol{x}_{d,n}))$, is now taken to be a fixed vector. Taking $\boldsymbol{B}_\gamma = [\boldsymbol{1}_n\ \boldsymbol{x}]$ we model $\xi(\boldsymbol{x}) = \boldsymbol{B}_\gamma\boldsymbol{\alpha}_\gamma$ and estimate $\boldsymbol{\alpha}_\gamma$ by solving:

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

3 Robust estimation of mean and dispersion

The Extended Quasi-likelihood (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:

$$\Psi_s(\boldsymbol{y}, \mu(\boldsymbol{x})) = \boldsymbol{B}_\mu^T \left(s_\psi(\boldsymbol{y}, \mu(\boldsymbol{x}))w(\boldsymbol{x})\boldsymbol{\mu}' - a(\boldsymbol{\alpha}_\mu)\right) = \boldsymbol{0} \; , \tag{3.1}$$

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

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

with ψ_c the Huber function defined as

$$\psi_c(x) = \begin{cases} x & \text{if } |x| \leq c \\ c \; \text{sign}(x) & \text{if } |x| > c \; . \end{cases} \tag{3.2}$$

Further $w(\cdot)$ in (3.1) is a weight function which controls the effect of leverage points on the estimate.

The tuning constant c in (3.2) balances the robustness and the efficiency of the estimate; if $w(\boldsymbol{x}) = \boldsymbol{1}_n$ and $c = \infty$ (3.1) boils down to (2.4). Cantoni and Ronchetti (2001a) discuss procedures to choose c. Unless otherwise stated, we take $c = 1.345$, the standard value which ensures 95% efficiency for the normal model. Our experience shows that this value gives reasonable results for other models as well. In (3.1) the constant

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

ensures Fisher consistency. Here we used the shorthand notation $\bar{E}_n[s_\psi(\boldsymbol{y}, \mu(\boldsymbol{x}))|\boldsymbol{X}_d = \boldsymbol{x}_d]$ for $n^{-1}\sum_{i=1}^n E\left[s_\psi(Y, \mu(\boldsymbol{X}_d))|\boldsymbol{X}_d = \boldsymbol{x}_{d,i}\right]$. This notation $\bar{E}_n$ will also be used in the next paragraph with a similar meaning involving a different quantity.

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

$$\Psi_t(\boldsymbol{d}, \gamma(\boldsymbol{x})) = \boldsymbol{B}_\gamma^T \left(t_\psi(\boldsymbol{d}, \gamma(\boldsymbol{x}))w(\boldsymbol{x})\boldsymbol{\gamma}' - b(\boldsymbol{\alpha}_\gamma)\right) = \boldsymbol{0} \; , \tag{3.3}$$

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 (3.2). Again, the constant

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

ensures Fisher consistency.

Both (3.1) and (3.3) cannot be solved analytically and iterative methods like the iteratively reweighted least squares algorithm (Hampel *et al* (1986)) are used. 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. GAMs were originally introduced as an extension to GLM by Hastie and Tibshirani (1990) to obtain flexible estimates for the mean estimation. Marx and Eilers (1998) have developed a way to estimate the smooth components $\eta_j(\cdot)$ via penalized B-splines (P-splines). Gijbels and Prosdocimi (2010) extended GAM to estimate both the mean and the dispersion as smooth functions of the covariates. We intend to further develop these Extended GAM in order to obtain estimates for both the mean and the dispersion which 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

Penalized splines were introduced by Eilers and Marx (1996) and are a valid and widely used smoothing technique. In their seminal work Eilers and Marx introduce P-splines in the case of a univariate independent variable X, while in Marx and Eilers (1998) they propose ways of using P-splines in the GAM framework. We give here a brief overview, and refer to the above mentioned papers for a more extensive reading.

In Figure 4.1 we see a graphical representation of how P-splines work in practice. The top panel depicts a large B-spline base of degree $p = 3$ based on the set of knots $(\kappa_1, \ldots, \kappa_k)$ represented by dots. B-splines of degree p are bell-shaped smooth curves, $(p-1)$ times differentiable, joined together at each knot point. For a given point of the X domain, $(p+1)$ B-splines have positive values and the B-spline base has a dimension $K = p + k + 1$.

For a given i.i.d. sample $(\boldsymbol{x}, \boldsymbol{y})$ we build a large B-spline base matrix $\boldsymbol{B}_\mu$ and use this as regression matrix in a GLM, taking $\eta(\boldsymbol{x}) = \boldsymbol{B}_\mu \boldsymbol{\alpha}_\mu$. Each B-spline j ($j = 1, \dots, K$) is then multiplied by a certain amplitude $\boldsymbol{\alpha}_{\mu,j}$, and the linear combination of these amplified B-splines corresponds to the final estimate. In Figure 4.1 (b) the amplified B-splines are drawn as dashed lines, while the estimate that is obtained as their linear combination can be seen as a solid line. Clearly, since the regression matrix is so rich, the fit is too wiggly and it overfits the data. In order to avoid this overfitting, we add a difference order type of penalty which ensures that adjacent coefficients are not too different. Estimates for $\boldsymbol{\alpha}_\mu$ are obtained as the solution to:

$$\boldsymbol{B}_\mu^T \left(\frac{\boldsymbol{y} - \mu(\boldsymbol{x})}{\phi V(\mu(\boldsymbol{x}))} \boldsymbol{\mu}' \right) - \lambda \boldsymbol{P}_\mu \boldsymbol{\alpha}_\mu = \boldsymbol{0} \,, \tag{4.1}$$

with $\boldsymbol{P}_\mu$ an appropriate matrix representation of the difference operator (as in Eilers and Marx (1996)). The smoothing parameter $\lambda > 0$ governs the balance between the overfitting and the smoothness of the fitted function. In Figure 4.1 (c) we see the effect of the smoothing parameter: the solid line is obtained with a moderate value for λ while the dashed-dotted line is obtained with a too large value for λ, leading to a too smooth estimate. In Section 5 we discuss methods to optimally choose the smoothing parameter.

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

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

Gijbels and Prosdocimi (2010) further extended P-GAMs and present Double Generalized Additive Models to model both the mean and the dispersion function as flexible functions of the covariates. In the next section we combine this extended P-GAM setting with the robust EQL framework presented in Section 2 to obtain robust and smooth estimates for both the mean and the dispersion function.

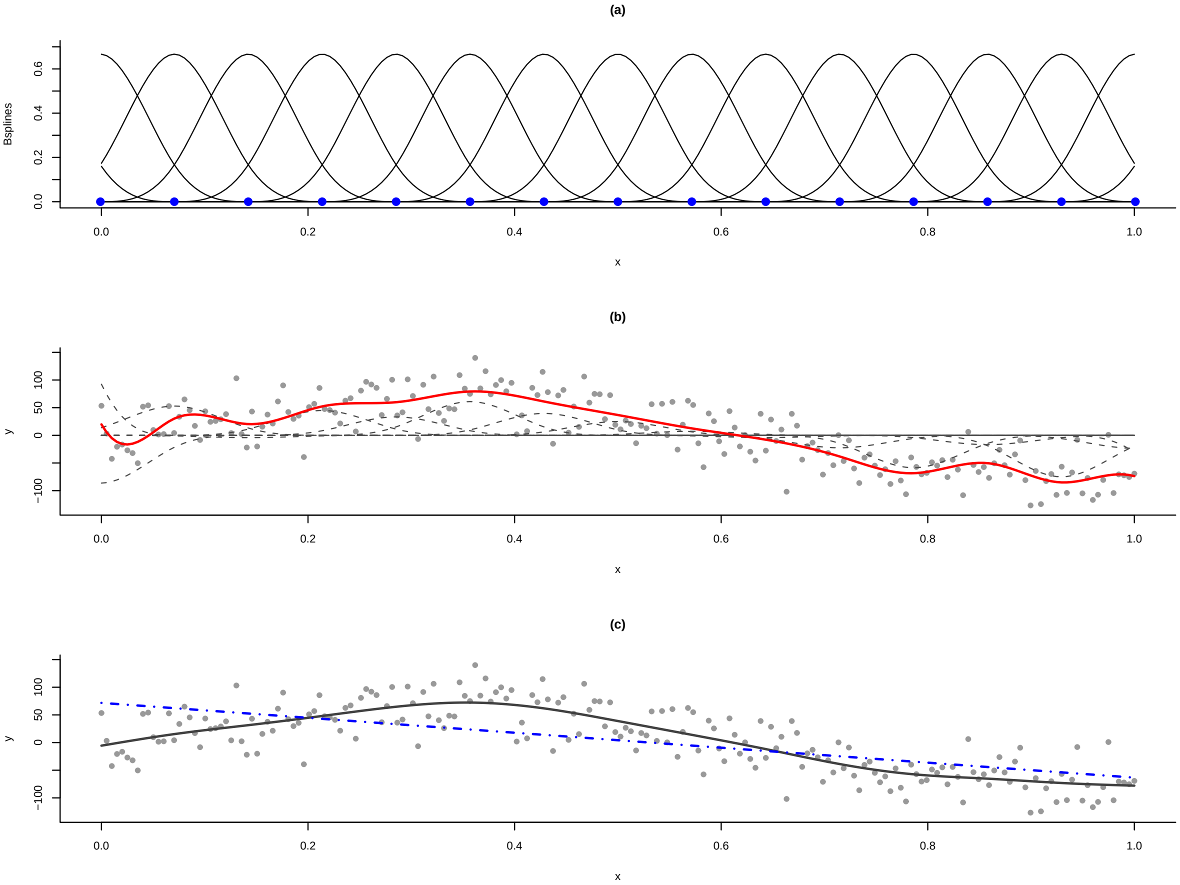

Figure 4.1: *P-splines in practice. (a) a B-spline base is shown; (b) the weighted B-splines (dashed lines) and the final fit obtained when using a large B-spline base (solid line). The effects of the smoothing parameter can be seen in (c).*

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 $(Y|\boldsymbol{X}_d = \boldsymbol{x}_d)$ and $(d(Y, \mu)|\boldsymbol{X}_d = \boldsymbol{x}_d)$ just as in (2.2) and (2.3). 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 $\eta(\boldsymbol{x}_d) = \alpha_{\mu,0} + \eta_1(x_1) + \ldots + \eta_d(x_d)$ and $\xi(\boldsymbol{x}_d) = \alpha_{\gamma,0} + \xi_1(x_1) + \ldots + \xi_d(x_d)$, where the $\eta_j(x_j)$ and $\xi_j(x_j)$ components are modelled via P-splines.

As in Section 4.1 for a given i.i.d. sample $(\boldsymbol{x}, \boldsymbol{y})$ we build the unique 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 modelling of the dispersion function.

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

$$\Psi_s(\boldsymbol{y}, \mu(\boldsymbol{x})) - \boldsymbol{P}_\mu \boldsymbol{\alpha}_\mu = \boldsymbol{B}_\mu^T \left(s_\psi(\boldsymbol{y}, \mu(\boldsymbol{x})) w(\boldsymbol{x}) \boldsymbol{\mu}' - a(\boldsymbol{\alpha}_\mu) \right) - \boldsymbol{P}_\mu \boldsymbol{\alpha}_\mu = \boldsymbol{0} \tag{4.2}$$

and

$$\Psi_t(\boldsymbol{d}, \gamma(\boldsymbol{x})) - \boldsymbol{P}_\gamma \boldsymbol{\alpha}_\gamma = \boldsymbol{B}_\gamma^T \left(t_\psi(\boldsymbol{d}, \gamma(\boldsymbol{x})) w(\boldsymbol{x}) \boldsymbol{\gamma}' - b(\boldsymbol{\alpha}_\gamma) \right) - \boldsymbol{P}_\gamma \boldsymbol{\alpha}_\gamma = \boldsymbol{0} \; . \tag{4.3}$$

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 (4.2) and (4.3) differ from (3.1) and (3.3) since now $\boldsymbol{B}_\mu$ and $\boldsymbol{B}_\gamma$ represent a larger combination of matrices, and there is the penalty term which 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:

$$\boldsymbol{\alpha}_\mu = (\boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu \boldsymbol{B}_\mu + \boldsymbol{P}_\mu)^{-1} \boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu \tilde{\boldsymbol{z}}_\mu \; , \tag{4.4}$$

where $\tilde{\boldsymbol{W}}_\mu = \text{diag}\left(-E\left[\frac{d}{d\boldsymbol{\alpha}_\mu} \Psi_s(Y, \tilde{\mu}(\boldsymbol{X}_d)) | \boldsymbol{X}_d = \boldsymbol{x}_{d,i} \right] \right)_i$, $\tilde{\boldsymbol{z}}_\mu = \boldsymbol{B}_\mu \tilde{\boldsymbol{\alpha}}_\mu + \tilde{\boldsymbol{W}}_\mu^{-1} \Psi_s(\boldsymbol{y}, \tilde{\mu}(\boldsymbol{x}))$ 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:

$$\boldsymbol{\alpha}_\gamma = (\boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma \boldsymbol{B}_\gamma + \boldsymbol{P}_\gamma)^{-1} \boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma \, \tilde{\boldsymbol{z}}_\gamma \; , \tag{4.5}$$

where $\tilde{\boldsymbol{W}}_\gamma = \text{diag}\left(-E\left[\frac{d}{d\boldsymbol{\alpha}_\gamma} \Psi_t(d(Y, \mu), \tilde{\gamma}(\boldsymbol{X}_d)) | \boldsymbol{X}_d = \boldsymbol{x}_{d,i} \right] \right)_i$, $\tilde{\boldsymbol{z}}_\gamma = \boldsymbol{B}_\gamma \tilde{\boldsymbol{\alpha}}_\gamma + \tilde{\boldsymbol{W}}_\gamma^{-1} \Psi_t(\boldsymbol{d}, \tilde{\gamma}(\boldsymbol{x}))$ 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 $\text{df}(\boldsymbol{\lambda}_\mu) = \text{tr}(\boldsymbol{H}_\mu(\boldsymbol{\lambda}_\mu))$ and $\text{df}(\boldsymbol{\lambda}_\gamma) = \text{tr}(\boldsymbol{H}_\gamma(\boldsymbol{\lambda}_\gamma))$ to be the equivalent number of degrees of freedom for the mean and the dispersion model.

The general estimation procedure iterates between the two iterative estimation steps for $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_\gamma$, as in Gijbels and Prosdocimi (2010). The smoothing parameters $\boldsymbol{\lambda}_\mu$ and $\boldsymbol{\lambda}_\gamma$ strongly influence the final appearance of the fits and are chosen before each of the iterative steps, using methods discussed in the next section.

5 Smoothing parameter selection

The solution to (4.4) and (4.5) can be found via iterative methods, although the smoothness of the final estimate depends on the values of the smoothing parameters $\boldsymbol{\lambda}_\mu$ and $\boldsymbol{\lambda}_\gamma$.

Different methods for choosing the smoothing parameters exist. A standard procedure is to minimize the Generalized Cross Validation (GCV) criterion (Wahba (1990)):

$$\mathrm{GCV}(\boldsymbol{\lambda}_\mu) = \mathbf{1}_n^T \frac{d(\boldsymbol{y}, \hat{\mu}(\boldsymbol{x}, \boldsymbol{\lambda}_\mu))}{(n - \mathrm{df}(\boldsymbol{\lambda}_\mu))^2} \ , \tag{5.1}$$

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 GAMs 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 (2010). 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:

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

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 (2010) propose an appropriate form of GCV for the choice of the $\boldsymbol{\lambda}_\gamma$. Here, we propose to use a robustified version of this criterion:

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

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 2.1)

$$d_\gamma(d, \gamma(\boldsymbol{x}_d)) = 2 \int_{\gamma(\boldsymbol{x}_d)}^{d} \frac{d - t}{2t^2} dt \ . \tag{5.2}$$

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:

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

and

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

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 (3.2) 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 $\mu(x) = \exp(\eta_0 + \eta_1(x_1) + \eta_2(x_2))$ and dispersion function $\gamma(x) = \exp(\xi_0 + \xi_1(x_1) + \xi_2(x_2))$ with:

$$\begin{array}{lll} \eta_0 = 1, & \eta_1(x_1) = 1.8\,\sin(3.4x_1^2)\,, & \eta_2(x_2) = 1.1\cos(8x_2) \\ \xi_0 = -.35, & \xi_1(x_1) = 2.3\,\sin(2x_1)x_1^2\,, & \xi_2(x_2) = -1.35(\sin(x_2)\exp(1.5 - 0.8x_2))\,. \end{array}$$

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 data points, 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, placing a growing percentage of outlying datapoints (0%, 3% and 5%) located in $.1 < x_1 < .2$ and $.8 < x_2 < .9$. In this Poisson-type modelling we take $V(\mu) = \mu$, and logarithmic link functions for both the mean and the dispersion: $\eta(\cdot) = \log(\mu(\cdot))$ and $\xi(\cdot) = \log(\gamma(\cdot))$. 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 4 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 non-robust Extended GAM of Gijbels and Prosdocimi (2010) and the standard GAM with mean function estimation only. In this way we are able to investigate the differences between both robust and non-robust 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 square error:

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

with $\hat{f}(\cdot)$ the estimated function and $f_{\mathrm{true}}(\cdot)$ the true function. In Figures 6.1 and 6.2 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 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 non-robust estimation of mean and dispersion function as in Gijbels and Prosdocimi (2010), 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 6.1 it is seen that in case of no contamination the non-robust and the robust methods have similar behavior, with non-robust 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 6.2 we can see that as soon as outliers are present in the data, the non-robust 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 non-robust 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

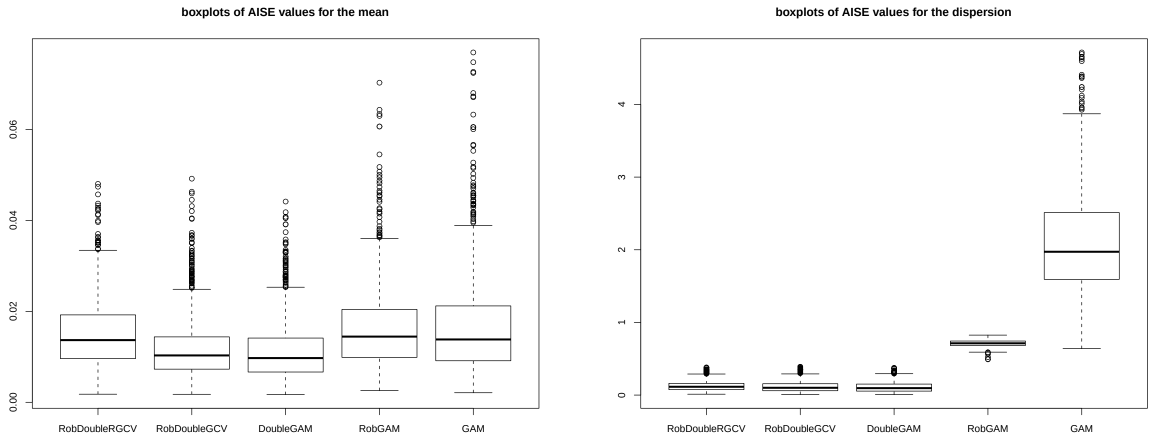

Figure 6.1: *The 0% contamination setting: boxplots of AISE values for the mean (left) and dispersion (right) estimation.*

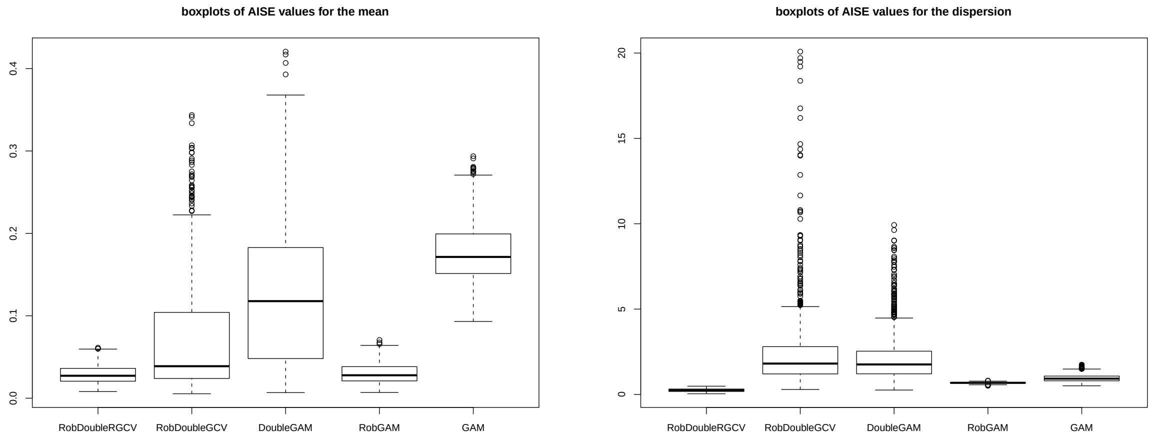

Figure 6.2: *The 3% contamination setting: boxplots of AISE values for the mean (left) and dispersion (right) estimation.*

well. In general the proposed method seem 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 6.3 we show a dataset simulated under the 3% contamination setting, together with non-robust 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 in the data.

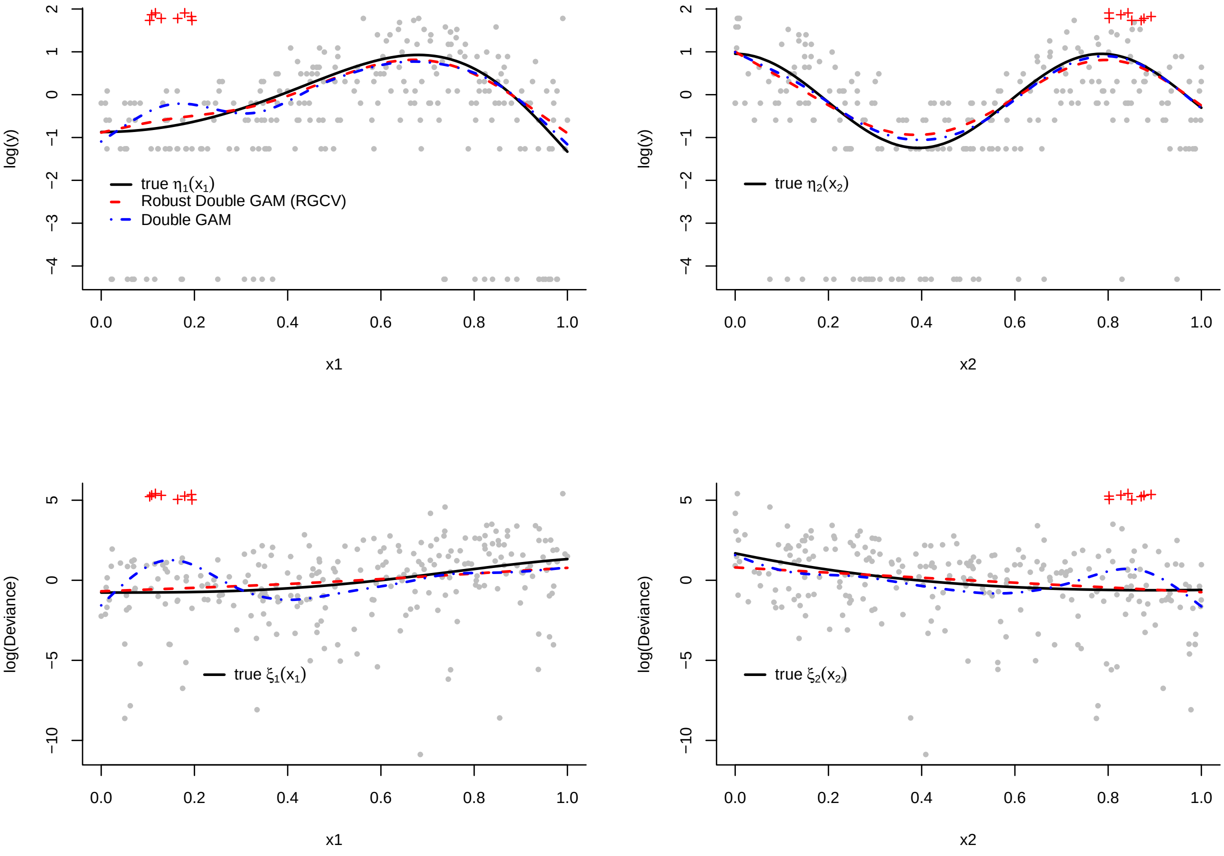

Figure 6.3: *A simulated dataset from the 3% contamination setting. Outliers are indicated with crosses and estimates for both the mean (top panels) and the dispersion (lower panels) functions are superimposed.*

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 visits in the US

Alimadad and Salibian-Barrera (2009) study how the weekly counts of Influenza-like Illness (ILI) visits in the US 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 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 modelling 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 right panel of Figure 7.1 we present the centered (log) data with a robust and a non-robust fit of the mean, while the centered (log) deviance residuals with a fit for the ξ(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. We claim that the dispersion of the model should indeed be estimated as a function: in Figure 7.2 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 modelling 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.

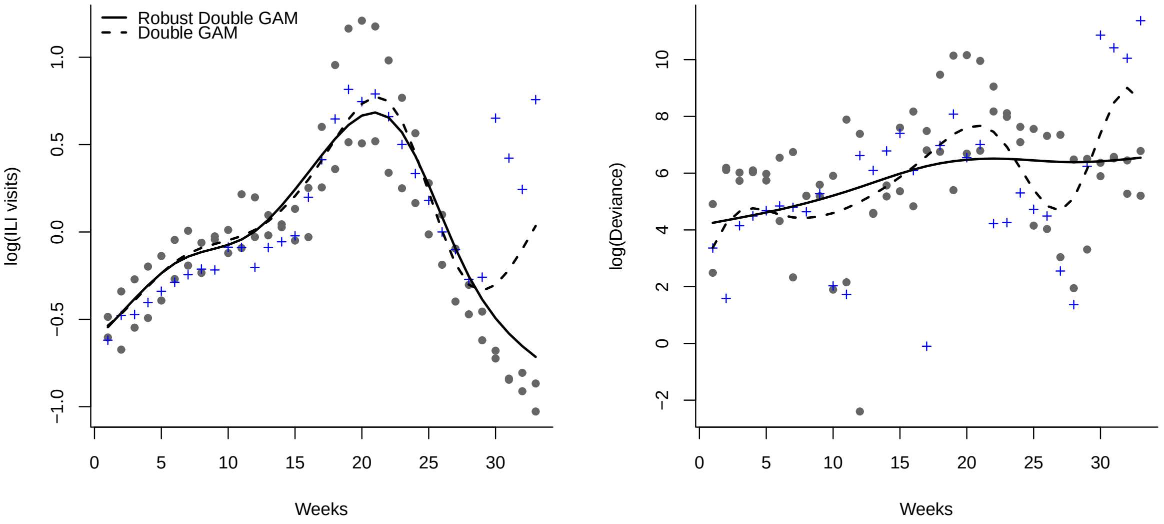

Figure 7.1: *ILI visits in the US (crosses indicate data of the 2008/2009 season): Robust Double GAM (solid) and standard Double GAM (dashed line) fits for the mean and dispersion function.*

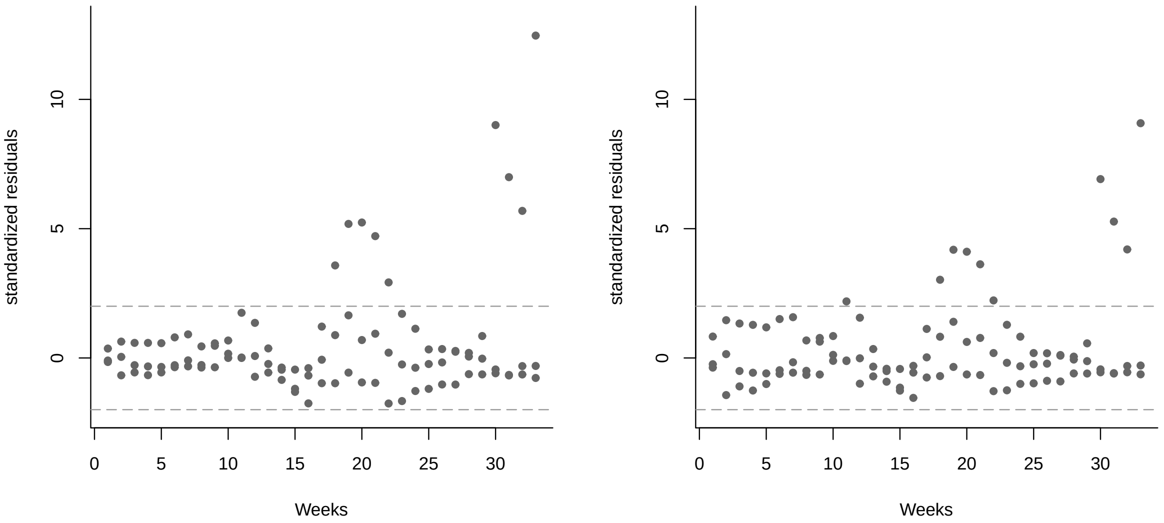

Figure 7.2: *Standardized residuals for the fits to the ILI visits. On the left the residuals obtained when taking the dispersion as a constant, on the right the ones obtained when taking the dispersion to be a function of the covariate.*

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 7.2 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 7.3 (top panels) data on the ozone level in Upland, California in 1976 (see Breiman and Friedman (1985)) are depicted. We are interested in modelling 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 data points to be substituted by outliers were selected among those with inversion base temperature values between (70,80). We take $V(\mu) = 1$ and the identity link $\eta(\cdot) = \mu(\cdot)$, the default

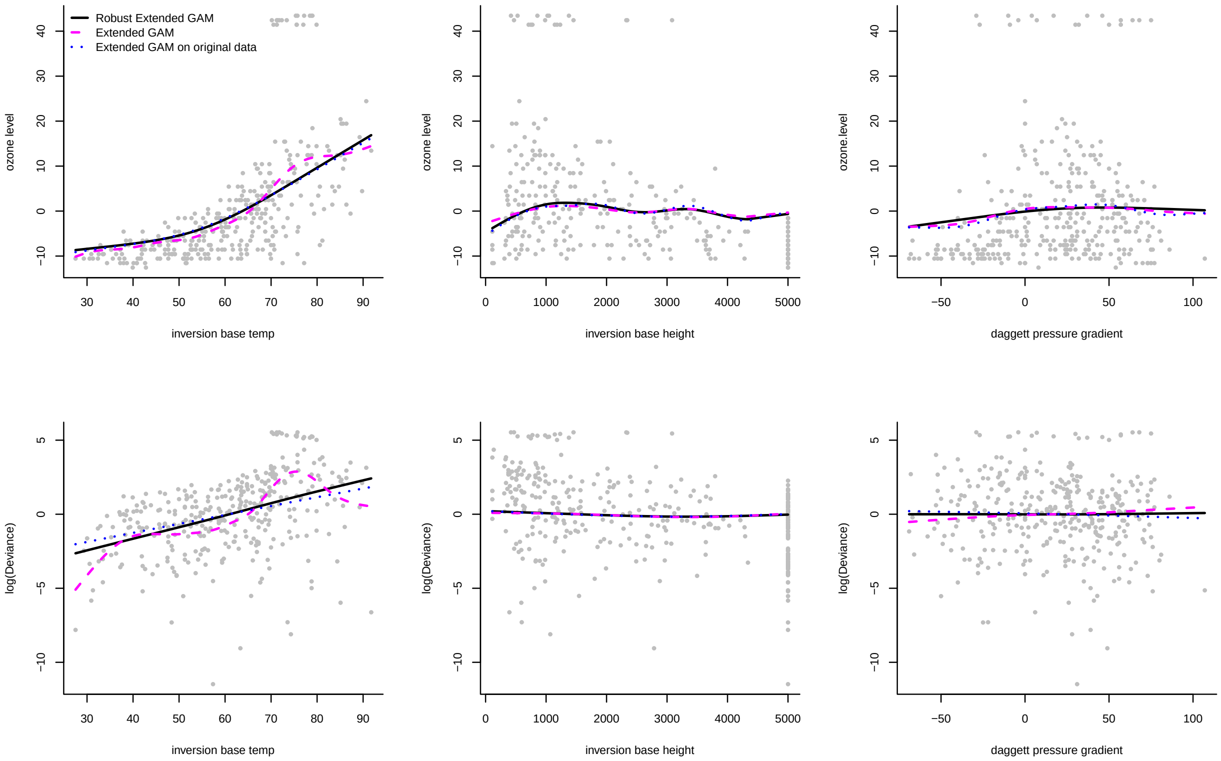

Figure 7.3: *The Ozone data with outliers: Robust Double GAM (solid) and non-robust Double GAM (dashed line) fits for the mean and the dispersion (top and bottom panels respectively). The dotted line represent the fit from a non-robust double estimation of the mean and dispersion function on the original data.*

choice in standard GAM for data assumed to be Normal.

In Figure 7.3 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.

7.3 The Italian abortion data

In Figure 7.4 data on the induced abortion rate for each Italian province are shown. This dataset was previously analyzed in Gijbels and Prosdocimi (2010). We are interested in studying how the abortion rate in the 98 Italian provinces changes as a function of the following socio-economical covariates:

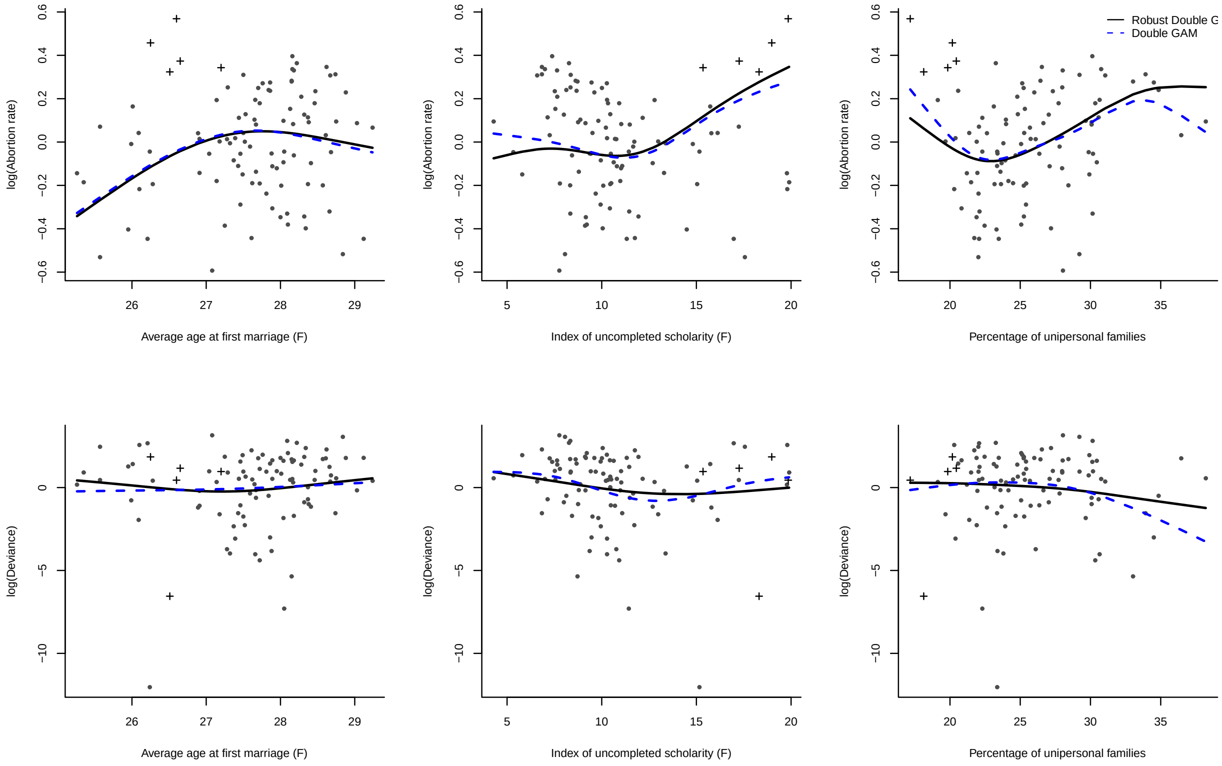

Figure 7.4: *The abortion data: Robust Double GAM (solid lines) and standard Double GAM (dashed lines) fits for the mean and dispersion function (top and bottom panels respectively). Crosses indicate the provinces of Puglia.*

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

As Section in 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 5 provinces in one region (Puglia) in which the health care system, specially 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 that these outlying points will have less effect on the final estimates, as can be seen in Figure 7.4: indeed the robust fits are less affected by the presence of extreme points.

Acknowledgements

Support from the GOA/07/04-project of the Research Fund KULeuven is gratefully acknowledged, as well as support from the IAP research network nr. P6/03 of the Federal Science Policy, Belgium.

References

Alimadad, A. and Salibian-Barrera, M. (2009). An outlier-robust fit for Generalised Additive Models with applications to outbreak detection. *Manuscript*, available at the second author's website.

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.

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. (2010). Smooth estimation of mean and dispersion function in extended Generalized Additive Models with application to Italian Induced Abortion data. *Manuscript.*

Gijbels I., Prosdocimi I. and Claeskens, G. (2010). Nonparametric estimation of mean and dispersion functions in extended Generalized Linear Models. *Test*, to appear. *DOI 10.1007/s11749-010-0187-1*

Hampel F., Ronchetti E., Rousseeuw P. and Stahel W. (1986). *Robust statistics: the approach based on influence functions.* Wiley, New York.

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

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*, Chapman and Hall: London.

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

Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM.

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