Flexible mean and dispersion function estimation in extended Generalized Additive Models.

I. Gijbels and I. Prosdocimi

Department of Mathematics, and Leuven Statistics Research Center (LStat), Katholieke Universiteit Leuven, Celestijnenlaan 200B, Box 2400, B-3001 Leuven (Heverlee), Belgium.

April 8, 2021

Abstract

Real data may expose a larger (or smaller) variability than assumed in an exponential family modeling, the basis of Generalized linear models and additive models. To analyze such data smooth estimation of the mean and the dispersion function has been introduced in extended generalized additive models using P-splines techniques. This methodology is here further explored by allowing to model some of the covariates parametrically and some nonparametrically. The main contribution in this paper is a simulation study investigating the finite-sample performance of the P-spline estimation technique in these extended models, including comparisons with a standard generalized additive modeling approach, as well as with a hierarchical modeling approach.

1 Introduction

A convenient way to describe the mean effect of two or more covariates on a variable of interest (the response) is via Generalized Additive Models (GAM), see Hastie and Tibshirani (1986, 1990) and Wood (2006a), among others.

Within the GAM framework one assumes that the data at hand come from a distribution belonging to the exponential family of distributions, which results in assuming a specific relationship between the variance and the mean function. In many cases though, data actually show a variability which deviates from the one expected from the theoretical model. This is for example the case for heteroscedastic normal data, for which the variability is not constant as assumed by the model, but changes as a function of the covariates. Another notable example is the case of over-dispersed (respectively under-dispersed) count or proportion data in which the variability shown by the data is higher (respectively lower) than the one assumed by the Poisson and Binomial distribution usually employed to analyze such data.

Some early work on dispersion function estimation can be found in Efron (1986), who introduces a double exponential family of distributions, which extends the standard

exponential family by adding an extra dispersion parameter which governs the extra variability observed in the data. Efron (1986) also presents methods to model this extra parameter as a parametric function of the covariates.

In this paper we focus on the case when the over-dispersion (or under-disperion) in the data is changing as a function of the covariates. Moreover, the mean and dispersion functions might be entirely unknown, or partially modeled in a parametric way and partially in a nonparametric way. In the univariate case a similar estimation method, in a fully nonparametric setting, can be found in Gijbels *et al.* (2010). An application to Italian abortion rate data in a multivariate setting is provided in Gijbels and Prosdocimi (2011). That paper also includes a small simulation study for a normal model in which all covariates enter in a nonparametric way. The aim of the current paper is to further investigate this flexible technique for estimation of mean and dispersion functions, and to provide a simulation study in which the quality of mean and dispersion function estimation for normal and counts data is investigated. To work in full generality we allow for a subset of the covariates to be modeled through a parametric function, whereas the effect of the remaining set of covariates is fully unknown and requires nonparametric estimation techniques. Both, the quality of estimation of the parametric part as well as the nonparametric part are investigated. A comparison with an alternative approach, in case of over-dispersed data, is also included.

Such an alternative approach is to start from the class of models called Generalized Additive Models for Location Scale and Shape (GAMLSS). See for example Rigby and Stasinopoulos (2005) and Stasinopoulos and Rigby (2007). This class of models allows the user to obtain smooth functional estimates for different parameters of a given distribution which is assumed to be generating the data. To model the dispersion of, for example, count data, one needs to specify a distribution which also allows for dispersion modeling. In this context, a typical extension of the standard Poisson distribution is the Negative Binomial distribution, which extends the standard Poisson via hierarchical modeling reasoning. Nevertheless the Negative Binomial distribution only allows for over-dispersion modeling. In our approach the modeling of under-dispersed data or over-dispersed data, or even the modeling of data showing a combination of over-dispersion and under-dispersion, are all easily done in one unique framework. Some comparisons with alternative approaches of hierarchical modeling type are available in the literature. See Gijbels *et al.* (2010) for such comparisons in the univariate covariate case, and Croux *et al.* (2011) for a comparison in a robust modeling context. In the current paper we in addition provide a comparison in the multivariate covariate setting in which some covariates enter the

model in a parametric way, and others in a nonparametric way. All these comparisons together reveal that the presented modeling framework is very appealing due to mainly (i) its easiness to capture various departures from the original theoretical model, and (ii) its computational convenience by using a sparse estimation method.

The paper is organized as follows. In Section 2 we briefly introduce the statistical modeling framework used in this paper. The method for a flexible estimation of the multivariate mean and dispersion function is exposed in Section 3. A detailed study investigating the finite-sample performance of the method is provided in Section 4. This section also presents comparisons with other methods. An illustration with a real data example is given in Section 5.

2 The statistical framework

Let Y be the response variable of interest and consider a set of d covariates $\boldsymbol{X}_d = (X_1, \ldots, X_d)$. In the settings of Generalized Linear Models (GLM), see for example McCullagh and Nelder (1989), and GAM one assumes that the response variable Y given $\boldsymbol{X}_d = \boldsymbol{x}_d$, with $\boldsymbol{x}_d = (x_1, \ldots, x_d) \in I\!R^d$, has a distribution belonging to the exponential family of distributions, i.e. the conditional density of Y given $\boldsymbol{X}_d = \boldsymbol{x}_d$, is

$$e_Y(y; \theta(\boldsymbol{x}_d), \phi) = \exp\left\{ \frac{y\theta(\boldsymbol{x}_d) - b(\theta(\boldsymbol{x}_d))}{\phi} + c(y; \phi) \right\}, \tag{2.1}$$

where $b(\cdot)$ and $c(\cdot; \phi)$ are known functions, identifying specific distributions and ϕ is a scale parameter. For short, we denote this as $(Y|\boldsymbol{X}_d = \boldsymbol{x}_d) \sim \mathrm{EF}(b(\theta(\boldsymbol{x}_d)), \phi)$. It can be shown that

$$\mu(\boldsymbol{x}_d) = \mathrm{E}[Y|\boldsymbol{X}_d = \boldsymbol{x}_d] = b'(\theta(\boldsymbol{x}_d)) \qquad \text{and} \qquad \mathrm{Var}[Y|\boldsymbol{X}_d = \boldsymbol{x}_d] = \phi b''(\theta(\boldsymbol{x}_d)) \; . \tag{2.2}$$

In GLM and GAM, a transformation of the mean function $\eta(\boldsymbol{x}_d) = g\left(\mu(\boldsymbol{x}_d)\right)$, with $g(\cdot)$ the link function, is then further modeled. A link function is called a canonical link when $\eta(\boldsymbol{x}_d) = g(b'(\theta(\boldsymbol{x}_d))) = \theta(\boldsymbol{x}_d)$, i.e. when $g(\cdot) = (b')^{-1}(\cdot)$. In this paper we work with canonical link functions.

In a GLM setting the function $\eta(\boldsymbol{x}_d)$ is modeled as a parametric linear function of the covariates. In GAM the unknown multivariate predictor $\eta(\boldsymbol{x}_d)$ is modeled as a linear combination of unknown univariate functions:

$$g(\mu(\boldsymbol{x}_d)) = \eta(\boldsymbol{x}_d) = \alpha_{\mu 0} + \eta_1(x_1) + \ldots + \eta_d(x_d) = \alpha_{\mu 0} + \sum_{j=1}^{d} \eta_j(x_j) \; , \tag{2.3}$$

with $\alpha_{\mu 0}$ an intercept parameter. The estimation task consists of estimating the parameter $\alpha_{\mu 0}$ and all functional univariate components $\eta_j(\cdot)$. Note that GLM models are essentially parametric in nature whereas GAM are essentially nonparametric. In general, the model in (2.3) is defined up to a constant and is not identifiable. Indeed, one could add and subtract the same constant α from two component functions $\eta_i(\cdot)$ and $\eta_j(\cdot)$ (with $i \neq j$), without affecting the final model. To avoid this identifiability issue, constraints $\mathrm{E}[\eta_j(X_j)] = 0$, are imposed on each component.

Of particular interest in this modeling context is to allow for some variables entering the model in a parametric linear fashion (e.g. as a polynomial), and others entering in a nonparametric fashion via additive modeling as in (2.3). This then leads to a semiparametric model.

Under the exponential family modeling it is assumed that the variance behaves as in (2.2). Such a structure of the variance can be too restrictive though. Data, and in particular counts or proportion data, sometimes show a variance that is smaller (respectively larger) than the variance imposed by the theoretical model in (2.2). One refers to this as under-dispersion (respectively over-dispersion). The amount of over- or under-dispersion may also vary with different values taken by the covariates. In a normal model with variance σ^2, $b(\theta) = \theta^2/2$, $b''(\theta) = 1$, the canonical link function is $g(t) = t$, and $\phi = \sigma^2$, and hence the theoretical model assumes a constant variance. However, in practice, data following a normal model may exhibit heteroscedasticity, i.e. a variance that is different for different values of the covariates.

Various approaches have been proposed to analyze over-dispersed data. A review of common methods can be found in Hinde and Demétrio (1998). For a comparison of different approaches see Nelder and Lee (1992) and Davidian and Carroll (1988), among others. See also Gijbels *et al.* (2010) and references therein.

A unique framework for modeling both over-dispersed and under-dispersed data as well as heteroscedastic data is obtained by starting from the double exponential family of distributions introduced by Efron (1986). For ease of presentation we focus first on the case with no covariates involved. In the exponential family model (2.1), take θ_S to be the choice of θ corresponding to the saturated one-parameter model, which maximizes $e_Y(y;\theta,\phi)$ over all possible values of θ ($\theta_S = (b')^{-1}(y)$). The corresponding double exponential family is

$$\tilde{f}_Y(y;\theta,\phi,\gamma) = c(\theta,\gamma)\gamma^{-\frac{1}{2}} e_Y(y;\theta,\phi)^{\frac{1}{\gamma}} e_Y(y;\theta_S,\phi)^{1-\frac{1}{\gamma}} , \tag{2.4}$$

where $c(\theta,\gamma)$ is a normalizing constant, such that $\int_{-\infty}^{\infty} \tilde{f}_Y(y;\theta,\phi,\gamma)dy = 1$. As discussed

in Efron (1986) and Nelder and Lee (1992) this normalizing constant can be approximated (in first order) by 1. The deviance for a one-parameter exponential family is defined as $d(y,\theta) = 2[\log(e_Y(y;\theta_S,\phi)) - \log(e_Y(y;\theta,\phi))]$, so that an approximation of (2.4) can be written as

$$f_Y(y;\theta,\phi,\gamma) = \gamma^{-\frac{1}{2}} e_Y(y;\theta,\phi)^{\frac{1}{\gamma}} e_Y(y;\theta_S,\phi)^{1-\frac{1}{\gamma}} = \gamma^{-\frac{1}{2}} \left\{ \exp\left[\frac{1}{2} d(y,\theta)\right] \right\}^{-\frac{1}{\gamma}} e_Y(y;\theta_S,\phi). \tag{2.5}$$

We denote this density as $Y \sim \text{DEF}(b(\theta),\phi,\gamma)$. Efron (1986) shows that for such a Y the approximate mean and variance are respectively $E(Y) = \mu = b'(\theta)$ and $\text{Var}[Y] = \gamma\phi b''(\theta)$. The role of the extra parameter γ is clear:

$\gamma = 1$: back to the one-parameter exponential family (2.1)
$\gamma > 1$: over-dispersion
$\gamma < 1$: under-dispersion.

In the case when Y is normally distributed γ coincides with the variance parameter (denoted with σ^2) when taking $\phi = 1$, and the normalizing constant $c(\theta,\gamma)$ has exactly value 1. For other distributions, the variance is the product of the variance we would have in the one-parameter exponential family framework multiplied by the value of the γ parameter. Estimation of γ thus refers to estimation of the variance (in a normal model) as well as to estimation of the dispersion.

Coming back to our covariate setting, we assume that $(Y|\boldsymbol{X}_d = \boldsymbol{x}_d) \sim \text{DEF}(b(\theta(\boldsymbol{x}_d)), \phi, \gamma(\boldsymbol{x}_d))$, where ϕ is assumed to be constant and known. Similarly as in (2.3) we use additive modeling for the (transformed) unknown dispersion function $\gamma(\boldsymbol{x}_d)$

$$h^{-1}(\gamma(\boldsymbol{x}_d)) = \xi(\boldsymbol{x}_d) = \alpha_{\gamma 0} + \xi_1(x_1) + \ldots + \xi_d(x_d)\ , \tag{2.6}$$

where $h(\cdot)$ is a given link function, and $\alpha_{\gamma 0}$ an intercept parameter. Since $\gamma(\boldsymbol{x}_d)$ is a dispersion function it must be nonnegative and a natural link function is $h(t) = \exp(t)$.

3 Flexible estimation of mean and dispersion

Looking back to (2.3) and (2.6) the task is now to estimate both intercept parameters $\alpha_{\mu 0}$ and $\alpha_{\gamma 0}$ as well as all unknown univariate functions $\eta_j(\cdot)$ and $\xi_j(\cdot)$, $j = 1,\ldots,d$. An appealing method for estimating these univariate functions is penalized splines. Eilers and Marx (1996) used B-splines as a starting basis and used a specific form of penalty, leading to the P-splines estimation technique. For explaining briefly this method and for introducing some notation, consider for a moment the case $d = 1$ and absence of

an intercept parameter in (2.3). The task is then to estimate a univariate function $\eta(\cdot)$. Eilers and Marx (1996) propose to model the linear predictor $\eta(x)$ as a linear combination of B-spline basis functions. 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 at each knot point κ_j, such that the resulting function is $(p-1)$ times differentiable with a continuous $(p-1)$th derivative. This results into a basis of dimension $K = k + p + 1$, and the linear predictor $\eta(\cdot)$ can be approximated in this space of B-spline basis functions:

$$\eta(x) = \sum_{j=1}^{K} \alpha_j B_j(x) = \boldsymbol{B}^T(x)\boldsymbol{\alpha} \,, \tag{3.1}$$

denoting $\boldsymbol{B}(x) = (B_1(x), \ldots, B_K(x))^T$ the B-splines base and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^T$, the unknown vector of parameters. The superscript T denotes the transpose of a vector or a matrix. Obviously, taking a large set of B-spline basis functions leads to a better approximation in (3.1), but will also lead to a large variability of the fit. This overfitting is controlled by introducing a penalty term in the log-likelihood. In P-splines regression this penalty is often taken to be based on finite differences of adjacent coefficients α_j, namely a penalty term $\sum_{j=m+1}^{K} (\Delta^m \alpha_j)^2$, where m is the order of the differencing operator. Examples are: with $m = 1$, $\Delta\alpha_j = \alpha_j - \alpha_{j-1}$; and with $m = 2$, $\Delta^2\alpha_j = \Delta\Delta\alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$.

From i.i.d. observations $(\boldsymbol{x}, \boldsymbol{y}) = ((x_1, y_1), \ldots, (x_n, y_n))^T$ from (X, Y), we obtain $\boldsymbol{B}(x_i)$, for all $i = 1, \ldots n$, and build from this the B-splines bases matrix $\boldsymbol{B}$ of dimension $n \times K$ in which the ith row is given by $\boldsymbol{B}^T(x_i) = (B_1(x_i), \ldots, B_K(x_i))$. The penalized log-likelihood, using (2.1), is defined as

$$l(\boldsymbol{\alpha}; \boldsymbol{x}, \boldsymbol{y}, \phi, \lambda) = \frac{\boldsymbol{y}^T \boldsymbol{B}\boldsymbol{\alpha} - \mathbf{1}_n^T b(\boldsymbol{B}\boldsymbol{\alpha})}{\phi} - \frac{1}{2}\lambda \boldsymbol{\alpha}^T \boldsymbol{D}_m^T \boldsymbol{D}_m \boldsymbol{\alpha}, \tag{3.2}$$

where $\lambda > 0$ is a smoothing parameter and $\mathbf{1}_n = (1, 1, \ldots, 1)^T$ denotes the unit vector of length n. The notation $b(\boldsymbol{B}\boldsymbol{\alpha})$ means that we apply the function $b(\cdot)$ to each element of the vector $\boldsymbol{B}\boldsymbol{\alpha}$, and obtain as such $b(\boldsymbol{B}\boldsymbol{\alpha}) = (b(\boldsymbol{B}^T(x_1)\boldsymbol{\alpha}), \ldots, b(\boldsymbol{B}^T(x_n)\boldsymbol{\alpha}))^T$. The same notation holds for other functions applied to a vector of values. The quantity $\boldsymbol{\alpha}^T \boldsymbol{D}_m^T \boldsymbol{D}_m \boldsymbol{\alpha}$ is the matrix representation of $\sum_{j=m+1}^{K} (\Delta^m \alpha_j)^2$. Maximization of (3.2) with respect to $\boldsymbol{\alpha}$ leads to the maximum penalized log-likelihood estimator of $\boldsymbol{\alpha}$. This estimator is obtained by using iterative procedures, like Fisher scoring.

We now turn to the general set up: the case of multivariate covariates (i.e. general d), the case of estimation of the mean and the dispersion function (see (2.3) and (2.6)), and the case of allowing for a subset of the covariates to be modeled via parametric functions.

We start by discussing the estimation of the mean function in (2.3) using the P-splines technique.

The expected value of $(Y|\boldsymbol{X}_d)$ is modeled as a function of the covariates $\boldsymbol{X}_d = (X_1, \ldots, X_d)$ through the link function $g(\cdot)$ in (2.3). Assume now that $d_{\text{P}} \leq d$ covariates, say $\boldsymbol{X}_d^{\text{P}} = (X_1, \ldots, X_{d_{\text{P}}})$, enter the model parametrically, while $d_{\text{NP}} = d - d_{\text{P}}$ covariates, say $\boldsymbol{X}_d^{\text{NP}} = (X_{d_{\text{P}}+1}, \ldots, X_d)$ are modeled nonparametrically via approximations with P-splines. Writing $\boldsymbol{x}_d = (x_1, \ldots, x_{d_{\text{P}}}, x_{d_{\text{P}}+1}, \ldots, x_d) = (\boldsymbol{x}_d^{\text{P}}, \boldsymbol{x}_d^{\text{NP}})$, and denoting by $\boldsymbol{B}_j(x_j)$ the parametric model basis of dimension K_j for modeling the parametric component of x_j for $j = 1, \ldots, d_{\text{P}}$. The whole parametric part can then be modeled via the basis $\boldsymbol{B}^{\text{P}}(\boldsymbol{x}_d^{\text{P}}) = \left[\boldsymbol{B}_1^T(x_1), \ldots, \boldsymbol{B}_{d_{\text{P}}}^T(x_{d_{\text{P}}})\right]$ of dimension $K_{\text{P}} = \sum_{j=1}^{d_{\text{P}}} K_j$. Consider for example the case $d = 3$, in which the effect of X_1 is modeled linearly (i.e. via $\alpha_1^{\text{P}} x_1$) and the effect of X_2 is modeled via a cubic function (i.e. via $\alpha_2^{\text{P}} x_2 + \alpha_3^{\text{P}} x_2^2 + \alpha_4^{\text{P}} x_2^3$). Then $K_1 = 1$ and $K_2 = 3$ and $\boldsymbol{B}_1^T(x_1) = x_1$ and $\boldsymbol{B}_2^T(x_2) = (x_2 \; x_2^2 \; x_2^3)$.

For the covariates entering the model in a nonparametric way we have d_{NP} sets of B-splines basis functions for the flexible modeling of these $d_{\text{NP}} = d - d_{\text{P}}$ covariates, denoted by $\boldsymbol{B}_j(x_j)$ of dimension K_j, for $j = d_{\text{P}} + 1, \ldots, d$. We denote by $\boldsymbol{B}^{\text{NP}}(\boldsymbol{x}_d^{\text{NP}}) = \left[\boldsymbol{B}_{d_{\text{P}}+1}^T(x_{d_{\text{P}}+1}) \ldots \boldsymbol{B}_d^T(x_d)\right]$ the global basis for this nonparametric modeling part, of dimension $K_{\text{NP}} = \sum_{j=d_{\text{P}}+1}^{d} K_j$. Finally, define $\boldsymbol{B}(\boldsymbol{x}_d) = [1, \boldsymbol{B}^{\text{P}}(\boldsymbol{x}_d^{\text{P}}), \boldsymbol{B}^{\text{NP}}(\boldsymbol{x}_d^{\text{NP}})]^T$ to obtain the model basis of dimension $K = 1 + K_{\text{P}} + K_{\text{NP}}$. Expression (2.3) can then be rewritten as

$$g(\mu(\boldsymbol{x}_d)) = \eta(\boldsymbol{x}_d) = \alpha_{\mu 0} + \boldsymbol{B}^{\text{P}}(\boldsymbol{x}_d^{\text{P}})\boldsymbol{\alpha}^{\text{P}} + \boldsymbol{B}^{\text{NP}}(\boldsymbol{x}_d^{\text{NP}})\boldsymbol{\alpha}^{\text{NP}} = \boldsymbol{B}^T(\boldsymbol{x}_d)\boldsymbol{\alpha} \; , \tag{3.3}$$

with $\boldsymbol{\alpha} = (\alpha_0, (\boldsymbol{\alpha}^{\text{P}})^T, (\boldsymbol{\alpha}^{\text{NP}})^T)^T$ the vector of unknown parameters of dimension K, to be estimated. Using large sets of knots the B-splines bases $\boldsymbol{B}^{\text{NP}}(\boldsymbol{x}_d^{\text{NP}})$ are built, and overfitting is avoided by introducing a vector of smoothing parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d_{\text{NP}}})$ and differencing type of penalties of order $(m_1, \ldots, m_{\text{NP}})$.

Based on i.i.d. observations $(\boldsymbol{x}, \boldsymbol{y}) = ((x_{11}, x_{21}, \ldots, x_{d1}, y_1), \ldots, (x_{1n}, x_{2n}, \ldots, x_{dn}, y_n))^T$ from $(X_1, \ldots, X_d, Y)$ one builds the model matrix $\boldsymbol{B} = [\boldsymbol{1}_n \; \boldsymbol{B}^{\text{P}} \; \boldsymbol{B}^{\text{NP}}]$. For a given smoothing parameter vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d_{\text{NP}}})$ the penalty matrix is $\boldsymbol{P} = \text{blockdiag}\,(0, \boldsymbol{0}_{K_{\text{P}}}, \lambda_1 \boldsymbol{D}_{m_1}^T \boldsymbol{D}_{m_1}, \ldots, \lambda_{d_{\text{NP}}} \boldsymbol{D}_{m_{d_{\text{NP}}}}^T \boldsymbol{D}_{m_{d_{\text{NP}}}})$, where the first $1 + K_{\text{P}}$ zero elements come from the fact that there is no need to penalize the parametric components of the model. The maximum penalized log-likelihood estimator for $\boldsymbol{\alpha}$ is obtained via iterative methods of the Fisher scoring type.

We now turn to the full generality of estimating both mean and dispersion function, with some covariates possibly entering in a parametric way. Of course the set of covariates

entering in a parametric way can be different for the mean and the dispersion function estimation. We thus need to expose this in the general description of the estimation procedure. Given the set of d covariates $\boldsymbol{X}_d = (X_1, \ldots, X_d)$, the mean is modeled as a function of a certain set of d_μ covariates, with $d_\mu \leq d$. Of these d_μ covariates a subset of d_{P_μ} covariates enter the mean model in a parametric fashion and the remaining $d_{\mathrm{NP}_\mu} = d_\mu - d_{\mathrm{P}_\mu}$ covariates enter the model in nonparametric way. The dispersion function $\gamma(\cdot)$ is modeled as a function varying with a set of d_γ covariates ($d_\gamma \leq d$), possibly different from the set of covariates used to model the mean function. Here d_{P_γ} covariates enter the model parametrically and the relation between the remaining $d_{\mathrm{NP}_\gamma} = d_\gamma - d_{\mathrm{P}_\gamma}$ covariates and the dispersion function is modeled in a nonparametric way. Denote by $\boldsymbol{X}_{d_\mu} = (X_1, \ldots, X_{d_{\mathrm{P}_\mu}}, X_{d_{\mathrm{P}_\mu}+1}, \ldots, X_{d_\mu}) = (\boldsymbol{X}^{\mathrm{P}}_{d_\mu}, \boldsymbol{X}^{\mathrm{NP}}_{d_\mu})$ the set of covariates that is used to model the mean function $\mu(\boldsymbol{x}_{d_\mu})$, and by $\boldsymbol{X}_{d_\gamma} = (X_1, \ldots, X_{d_{P_\gamma}}, X_{d_{P_\gamma}+1}, \ldots, X_{d_\gamma}) = (\boldsymbol{X}^{\mathrm{P}}_{d_\gamma}, \boldsymbol{X}^{\mathrm{NP}}_{d_\gamma})$ the set of covariates for modeling the dispersion function $\gamma(\boldsymbol{x}_{d_\gamma})$.

The components in the nonparametric part are modeled via P-splines, for which smoothing parameters $\boldsymbol{\lambda}^\mu = (\lambda^\mu_1, \ldots, \lambda^\mu_{d_{\mathrm{NP}_\mu}})$ and $\boldsymbol{\lambda}^\gamma = (\lambda^\gamma_1, \ldots, \lambda^\gamma_{d_{\mathrm{NP}_\gamma}})$ and differencing type of penalties of order $m_1, \ldots, m_{d_{\mathrm{NP}_\mu}}$ and $\ell_1, \ldots, \ell_{d_{\mathrm{NP}_\gamma}}$ are introduced. Similar as in the previous paragraph when we rewrote (2.3) into (3.3), we now re-express (2.3) and (2.6) by defining $\boldsymbol{B}_\mu(\boldsymbol{x}_{d_\mu}) = \left[1, \boldsymbol{B}^{\mathrm{P}}_\mu(\boldsymbol{x}^{\mathrm{P}}_{d_\mu}), \boldsymbol{B}^{\mathrm{NP}}_\mu(\boldsymbol{x}^{\mathrm{NP}}_{d_\mu})\right]^T$ and $\boldsymbol{B}_\gamma(\boldsymbol{x}_{d_\gamma}) = \left[1, \boldsymbol{B}^{\mathrm{P}}_\gamma(\boldsymbol{x}^{\mathrm{P}}_{d_\gamma}), \boldsymbol{B}^{\mathrm{NP}}_\gamma(\boldsymbol{x}^{\mathrm{NP}}_{d_\gamma})\right]^T$ so that

$$g(\mu(\boldsymbol{x}_{d_\mu})) = \eta(\boldsymbol{x}_{d_\mu}) = \alpha_{\mu 0} + \boldsymbol{B}^{\mathrm{P}}_\mu(\boldsymbol{x}^{\mathrm{P}}_{d_\mu})\boldsymbol{\alpha}^{\mathrm{P}}_\mu + \boldsymbol{B}^{\mathrm{NP}}_\mu(\boldsymbol{x}^{\mathrm{NP}}_{d_\mu})\boldsymbol{\alpha}^{\mathrm{NP}}_\mu = \boldsymbol{B}^T_\mu(\boldsymbol{x}_{d_\mu})\boldsymbol{\alpha}_\mu \,, \tag{3.4}$$

and

$$h^{-1}(\gamma(\boldsymbol{x}_{d_\gamma})) = \xi(\boldsymbol{x}_{d_\gamma}) = \alpha_{\gamma 0} + \boldsymbol{B}^{\mathrm{P}}_\gamma(\boldsymbol{x}^{\mathrm{P}}_{d_\gamma})\boldsymbol{\alpha}^{\mathrm{P}}_\gamma + \boldsymbol{B}^{\mathrm{NP}}_\gamma(\boldsymbol{x}^{\mathrm{NP}}_{d_\gamma})\boldsymbol{\alpha}^{\mathrm{NP}}_\gamma = \boldsymbol{B}^T_\gamma(\boldsymbol{x}_{d_\gamma})\boldsymbol{\alpha}_\gamma \,, \tag{3.5}$$

with $\boldsymbol{\alpha}_\mu = (\alpha_{\mu 0}, (\boldsymbol{\alpha}^{\mathrm{P}}_\mu)^T, (\boldsymbol{\alpha}^{\mathrm{NP}}_\mu)^T)^T$ and $\boldsymbol{\alpha}_\gamma = (\alpha_{\gamma 0}, (\boldsymbol{\alpha}^{\mathrm{P}}_\gamma)^T, (\boldsymbol{\alpha}^{\mathrm{NP}}_\gamma)^T)^T$ the vectors of unknown parameters to be estimated.

For a given sample of n i.i.d. observations $(\boldsymbol{x}, \boldsymbol{y})$ from $(\boldsymbol{X}_d, Y)$, we extract from $\boldsymbol{x}$ the $\boldsymbol{x}_\mu$ vector and the $\boldsymbol{x}_\gamma$ vector in which we consider only the observed values of respectively the $\boldsymbol{X}_{d_\mu}$ covariates and the $\boldsymbol{X}_{d_\gamma}$ covariates. From the resulting set of observations we then build the 'design' matrices $\boldsymbol{B}_\mu = [\boldsymbol{1}_n \ \boldsymbol{B}^{\mathrm{P}}_\mu \ \boldsymbol{B}^{\mathrm{NP}}_\mu]$ and $\boldsymbol{B}_\gamma = [\boldsymbol{1}_n \ \boldsymbol{B}^{\mathrm{P}}_\gamma \ \boldsymbol{B}^{\mathrm{NP}}_\gamma]$ as described above. Further, for given $\boldsymbol{\lambda}^\mu$ and $\boldsymbol{\lambda}^\gamma$, we build the two penalty matrices $\boldsymbol{P}_\mu = \text{blockdiag}\,(0, \boldsymbol{0}_{K_{\mathrm{P}_\mu}}, \lambda^\mu_1 \boldsymbol{D}^T_{m_1}\boldsymbol{D}_{m_1}, \ldots, \lambda^\mu_{d_{\mathrm{NP}_\mu}} \boldsymbol{D}^T_{m_{d_{\mathrm{NP}_\mu}}}\boldsymbol{D}_{m_{d_{\mathrm{NP}_\mu}}})$ and $\boldsymbol{P}_\gamma = \text{blockdiag}\,(0, \boldsymbol{0}_{K_{\mathrm{P}_\gamma}}, \lambda^\gamma_1 \boldsymbol{D}^T_{\ell_1}\boldsymbol{D}_{\ell_1}, \ldots, \lambda^\gamma_{d_{\mathrm{NP}_\gamma}} \boldsymbol{D}^T_{\ell_{d_{\mathrm{NP}_\gamma}}}\boldsymbol{D}_{\ell_{d_{\mathrm{NP}_\gamma}}})$. For given smoothing parameter vectors $\boldsymbol{\lambda}^\mu$ and $\boldsymbol{\lambda}^\gamma$,

the estimates for $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_\gamma$ are obtained by maximizing the penalized log-likelihood:

$$l(\boldsymbol{\alpha}_\mu, \boldsymbol{\alpha}_\gamma; \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}^\mu, \boldsymbol{\lambda}^\gamma, \phi) = -\frac{1}{2}\mathbf{1}_n^T \left\{ \log(h(\boldsymbol{B}_\gamma \boldsymbol{\alpha}_\gamma)) + \frac{1}{h(\boldsymbol{B}_\gamma \boldsymbol{\alpha}_\gamma)} d(\boldsymbol{y}, \boldsymbol{B}_\mu \boldsymbol{\alpha}_\mu) \right\} - \frac{1}{2}\boldsymbol{\alpha}_\mu^T \boldsymbol{P}_\mu \boldsymbol{\alpha}_\mu - \frac{1}{2}\boldsymbol{\alpha}_\gamma^T \boldsymbol{P}_\gamma \boldsymbol{\alpha}_\gamma \ , \tag{3.6}$$

where the log-likelihood function follows from (2.5). Maximization of (3.6) is done via a two-steps iterative procedure: we first maximize with respect to $\boldsymbol{\alpha}_\mu$ and then with respect to $\boldsymbol{\alpha}_\gamma$ and iterate between the two steps until convergence. Each of the two maximization steps is done via Fisher scoring. More precisely, the estimation algorithm iterates between the following two steps:

- Step (a): estimation of $\boldsymbol{\alpha}_\mu$. Taking $\boldsymbol{\alpha}_\gamma$ to be fixed, an optimal value for $\boldsymbol{\lambda}^\mu$ is selected, and estimates of $\boldsymbol{\alpha}_\mu$ are obtained via the following updating rule

$$\boldsymbol{\alpha}_\mu = (\boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu \boldsymbol{B}_\mu + \boldsymbol{P}_\mu)^{-1} \boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu \tilde{\boldsymbol{z}_\mu} \ , \tag{3.7}$$

  with $\tilde{\boldsymbol{\alpha}}_\mu$ the current estimated value of $\boldsymbol{\alpha}_\mu$, $\tilde{\boldsymbol{z}}_\mu = \boldsymbol{B}_\mu \tilde{\boldsymbol{\alpha}}_\mu + (\boldsymbol{y} - b'(\boldsymbol{B}_\mu \tilde{\boldsymbol{\alpha}}_\mu))/b''(\boldsymbol{B}_\mu \tilde{\boldsymbol{\alpha}}_\mu)$ the vector of working variables and with $\tilde{\boldsymbol{W}}_\mu$ the current diagonal matrix $\tilde{\boldsymbol{W}}_\mu = \text{diag}\left(\frac{b''(\boldsymbol{B}_\mu \tilde{\boldsymbol{\alpha}}_\mu)}{\phi\gamma(\boldsymbol{x}_\gamma)}\right)$. Once convergence is reached, values for $\mu(\boldsymbol{x})$ and $d(\boldsymbol{y}, \boldsymbol{B}_\mu \boldsymbol{\alpha}_\mu)$ are computed.

- Step (b): estimation of $\boldsymbol{\alpha}_\gamma$. Taking $\boldsymbol{\alpha}_\mu$ to be fixed, an optimal value for $\boldsymbol{\lambda}^\gamma$ is selected, and estimates of $\boldsymbol{\alpha}_\gamma$ are obtained via the updating rule

$$\boldsymbol{\alpha}_\gamma = (\boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma \boldsymbol{B}_\gamma + \boldsymbol{P}_\gamma)^{-1} \boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma \tilde{\boldsymbol{z}}_\gamma, \tag{3.8}$$

  with $\tilde{\boldsymbol{\alpha}}_\gamma$ the current value of $\boldsymbol{\alpha}_\gamma$, $\tilde{\boldsymbol{z}}_\gamma$ the working variable vector

$$\tilde{\boldsymbol{z}}_\gamma = \boldsymbol{B}_\gamma \tilde{\boldsymbol{\alpha}}_\gamma + (d(\boldsymbol{y}, \theta(\boldsymbol{x}_\mu)) - h(\boldsymbol{B}_\gamma \tilde{\boldsymbol{\alpha}}_\gamma)) \frac{1}{h'(\boldsymbol{B}_\gamma \tilde{\boldsymbol{\alpha}}_\gamma)} \ ,$$

  and with $\tilde{\boldsymbol{W}}_\gamma$ the current diagonal matrix of weights

$$\tilde{\boldsymbol{W}}_\gamma = \frac{1}{2}\text{diag}\left(\frac{h'(\boldsymbol{B}_\gamma \tilde{\boldsymbol{\alpha}}_\gamma)}{h(\boldsymbol{B}_\gamma \tilde{\boldsymbol{\alpha}}_\gamma)}\right)^2 .$$

The smoothing parameters $\boldsymbol{\lambda}^\mu$ and $\boldsymbol{\lambda}^\gamma$ are chosen by generalized cross validation criteria. See Eilers and Marx (1996), Gu and Xiang (2001), Wood (2006a, 2008) and Gijbels *et al.* (2010). For more details on the estimation algorithm and practical implementations see Gijbels and Prosdocimi (2011).

Asymptotic properties of penalized spline estimators of a mean regression function in a univariate setting are studied in a general context in Claeskens *et al.* (2009). Antoniadis *et*

al. (2011) provide rates of convergence in P-spline estimation of a univariate mean regression function for general classes of penalty functions. Consistency of P-spline estimation of a mean regression function in a multivariate additive modeling setting is established in Antoniadis *et al.* (2011), relying on the univariate consistency results and consistency of a backfitting procedure in additive modeling (see e.g. Horowitz *et al.* (2006)). The method in this paper involves P-spline estimation of both the (multivariate) mean and dispersion function. Although no theoretical results are available yet, it is to be expected that the estimation of mean and variance/dispersion remains consistent as long as the estimation of the latter function is based on appropriate residuals of the consistent mean estimation. See for example Hall and Carroll (1989) and Fan and Yao (1998), among others, for theoretical results on variance estimation in a regression context.

4 Simulation study

The aim of this section is to investigate the finite-sample performance of the estimation of the parametric and nonparametric components in (3.4) and (3.5) via the maximum penalized log-likelihood estimation method in (3.6). We also provide, in Section 4.2, comparisons with: (i) the standard GAM method in which the variance is constant; and (ii) a hierarchical modeling approach in case of over-dispersion.

4.1 Finite-sample performance of the method

In this simulation study we consider the $d = 2$ covariate case for different type of models: a normal model, in which we are thus estimating mean and variance functions; and a Poisson model, in which we estimate mean and dispersion functions. In all simulation models the two covariates are generated as independent $\mathrm{U}(0, 1)$ random variables.

For each modeling type, the normal and the Poisson model, we consider two different settings, referred to as Models A and B hereafter.

Normal model settings:

$$\mu(x_1, x_2) = \eta(x_1, x_2) = \alpha_{\mu 0} + \eta_1(x_1) + \eta_2(x_2)$$
$$\log(\gamma_A(x_1, x_2)) = \xi_A(x_1, x_2) = \alpha_{\gamma 0} + \xi_1(x_1) + \xi_2(x_2)$$
$$\log(\gamma_B(x_1, x_2)) = \xi_B(x_1, x_2) = \alpha_{\gamma 0} + \xi_2(x_1) + \xi_1(x_2)$$

where

$$\alpha_{\mu 0} = 45 \qquad \eta_1(x) = -18x + 7x^2 + 6x^3 \qquad \eta_2(x) = 7.2 \sin(90 + 6x) \cos((x + 0.5)^{1.5})$$
$$\alpha_{\gamma 0} = 0.5 \qquad \xi_1(x) = 1.5x - 0.5x^2 + 0.15x^3 \qquad \xi_2(x) = 0.4 \cos(4(x - 0.2)) \exp(0.6x).$$

The functions $\eta_1(\cdot)$ and $\xi_1(\cdot)$ are entering the model in a parametric way. The modeling of the parametric part as a third degree polynomial is motivated by the fact that such a function often can capture global influences (of linear, quadratic or cubic type) of covariates. See also Section 5.

Poisson model settings:

$$\log(\mu(x_1, x_2)) = \eta(x_1, x_2) = \alpha_{\mu 0} + \eta_1(x_1) + \eta_2(x_2)$$
$$\log(\gamma_A(x_1, x_2)) = \xi_A(x_1, x_2) = \alpha_{\gamma 0} + \xi_1(x_1) + \xi_2(x_2)$$
$$\log(\gamma_B(x_1, x_2)) = \xi_B(x_1, x_2) = \alpha_{\gamma 0} + \xi_2(x_1) + \xi_1(x_2)$$

where

$\alpha_{\mu 0} = 4$ $\quad \eta_1(x) = 0.85\cos(\pi x)\sin(7.2(x - 0.5))$ $\quad \eta_2(x) = 1.8\sin(x)\cos(x)$

$\alpha_{\gamma 0} = 0.5$ $\quad \xi_1(x) = 1.5\exp(0.5x^2)$ $\quad \xi_2(x) = 0.6(\sin(90 + 6x) + 4\cos(x^{1.5}))$.

Here all covariates enter the model in a nonparametric way.

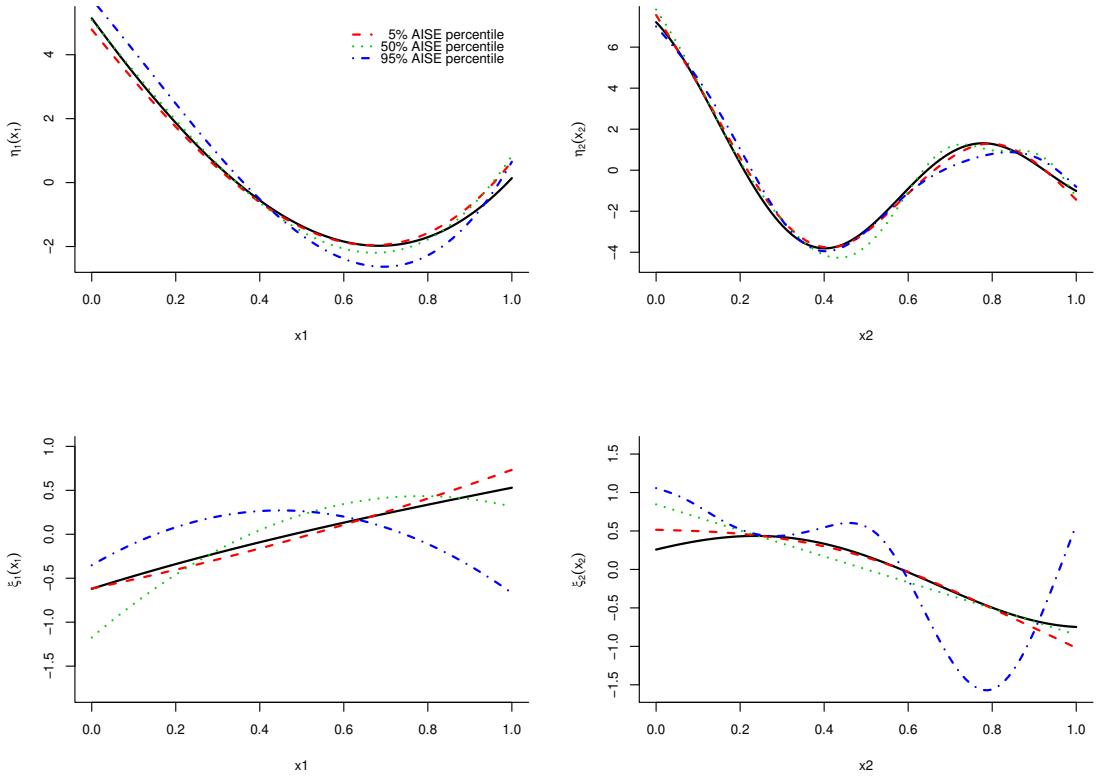

Figure 4.1: *Normal model setting A: true mean and dispersion components with representative estimates.*

Note that the difference between the two models (Models A and B) in each of the settings is simply obtained by swapping the roles played by the covariates. By doing so we aim at a better understanding of how the estimation of the mean and the disper-

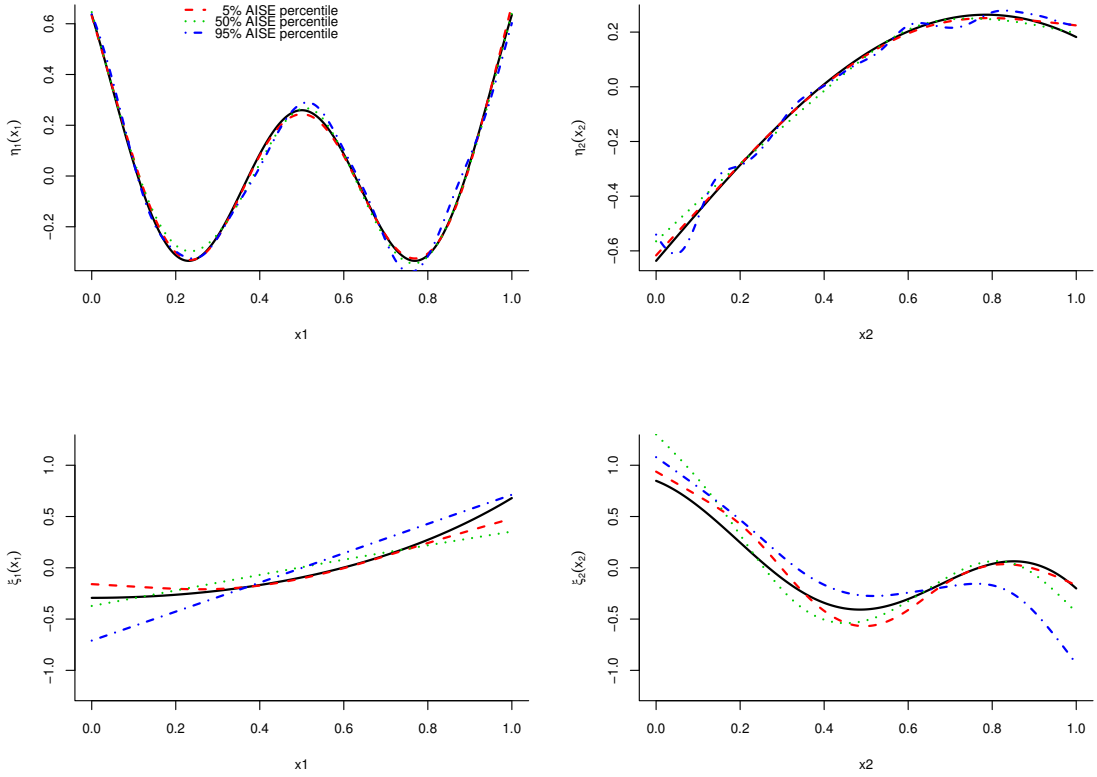

Figure 4.2: *Poisson model setting A, sample size $n = 300$: true mean and dispersion components with representative estimates.*

sion interact with each other and influence each other. The true mean and dispersion components are displayed as bold solid lines in Figures 4.1 and 4.2.

For the normal models we took samples of size $n = 100$, while for the Poisson models we use two different sample sizes: $n = 70$ and $n = 300$. For all the simulation settings we simulate 1000 data sets and evaluate the quality of the obtained fits via an approximate integrated squared error (AISE)

$$\text{AISE}^{(s)} = \frac{\sum_{x_{\text{grid}}} \left(\hat{f}^{(s)}(x_{\text{grid}}) - f_{\text{true}} x_{\text{grid}}) \right)^2}{\sum_{x_{\text{grid}}} \left(f_{\text{true}}(x_{\text{grid}}) \right)^2}, \quad \text{for } s = 1, \ldots, 1000,$$

for each of the simulations (indexed by s).

In Figures 4.1 and 4.2 we present, for the model settings A, the true component functions with some representative estimates taken as the estimated curves associated to the 5th, the 50th and the 95th percentile of the ordered AISE-values across the 1000 simulation results. As is seen from Figures 4.1 and 4.2 the quality of the estimation is quite good.

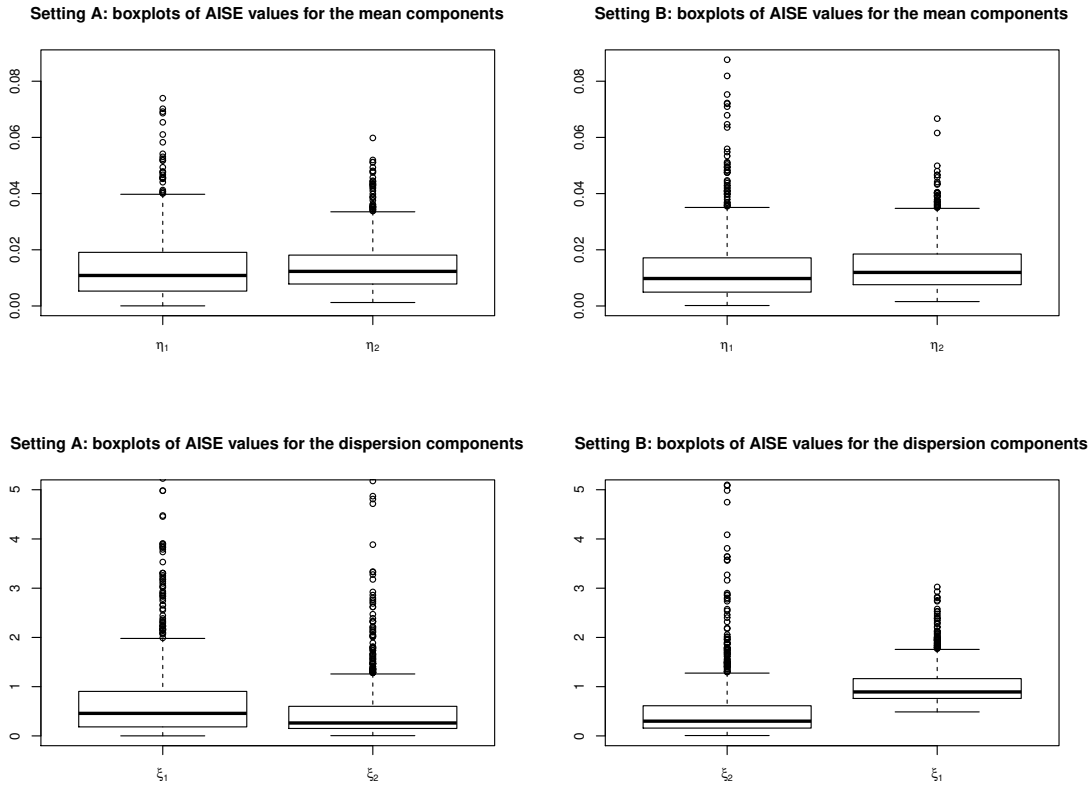

Figure 4.3: *Normal model settings. Boxplots of the AISE-values for the mean and the dispersion function for model setting A (left panels) and model setting B (right panels).*

In Figure 4.3 we show boxplots of the AISE-values for the mean and the dispersion components in both settings for the normal data. The performance of the estimation in the two settings for the mean component is quite comparable, while we see a considerable difference in the quality of the estimation of $\xi_1(\cdot)$ (the parametric component) in the dispersion function. Somehow the estimation of this parametric component in the dispersion function appears more difficult than the estimation of the nonparametric component.

Figures 4.4 and 4.5 present boxplots of the AISE-values for the mean and the dispersion components for the Poisson data settings, for sample sizes $n = 70$ and $n = 300$ respectively. Note that the estimation of the component $\eta_1(\cdot)$ in the mean function is remarkably better when the underlying component in the dispersion function is not so smooth (model setting B). This difference is already noticeable for the sample size $n = 70$. Overall the estimation of the smoother dispersion component $\xi_1(\cdot)$ is less biased but shows a larger variability.

Since in the normal model settings one covariate enters the model parametrically and the other one nonparametrically it is of particular interest to investigate this modeling

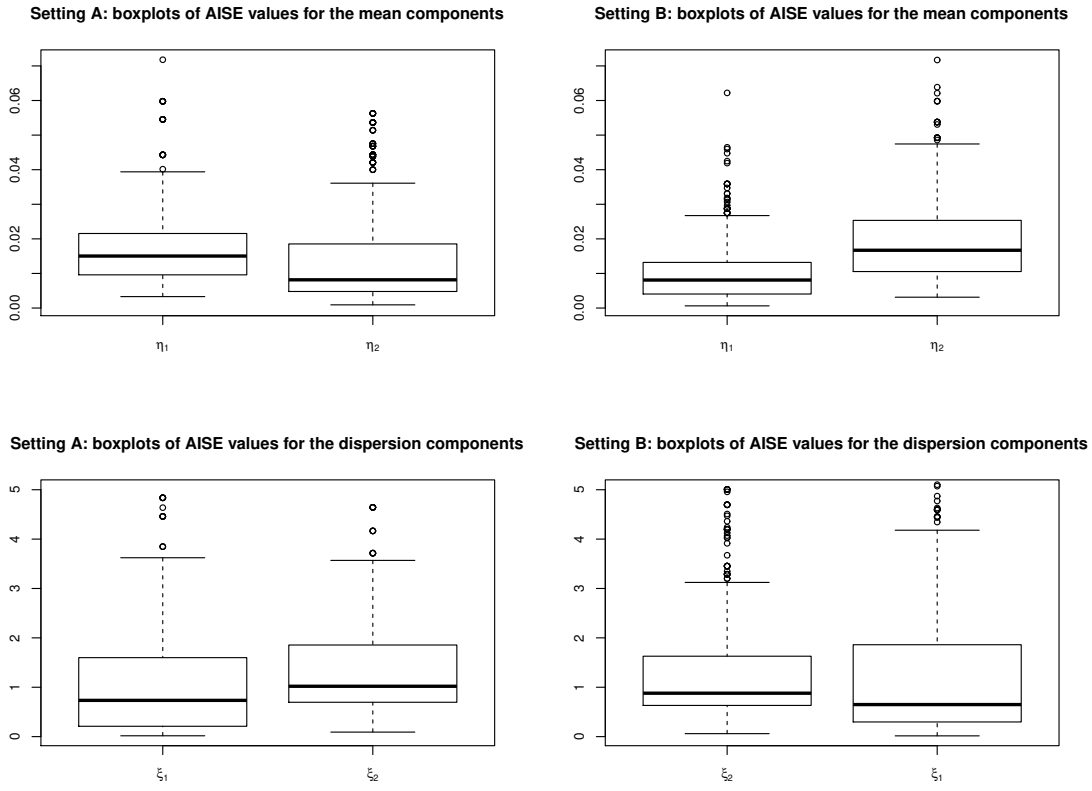

Figure 4.4: *Poisson data, sample size* $n = 70$. *Boxplots of the AISE-values for the mean and the dispersion function for model setting A (left panels) and model setting B (right panels).*

strategy aspect. In the normal model setting A, the true parameter values are

$$\boldsymbol{\alpha}_{\mu}^{\mathrm{P}} = (-18, 7, 6) \qquad \text{and} \qquad \boldsymbol{\alpha}_{\gamma}^{\mathrm{P}} = (1.5, -0.5, 0.15)\ .$$

In Figures 4.6 and 4.7 we present boxplots of the estimated parameters for the parametric component estimation of respectively the mean and the variance function. Not surprisingly, the estimation of the parameters in the parametric dispersion component function is of a lesser quality than these in the parametric mean component function. For comparison purpose we also applied the estimation method when BOTH covariates enter the model in a nonparametric way. The performance of the (nonparametrically) estimated parametric components is summarized in Figure 4.8 which should be compared with Figure 4.1. From a comparison of these two figures, it is clear that nonparametric estimation of a parametric component increases slightly the estimation variability (most noticeable in the estimation of the mean components) but that the quality remains good.

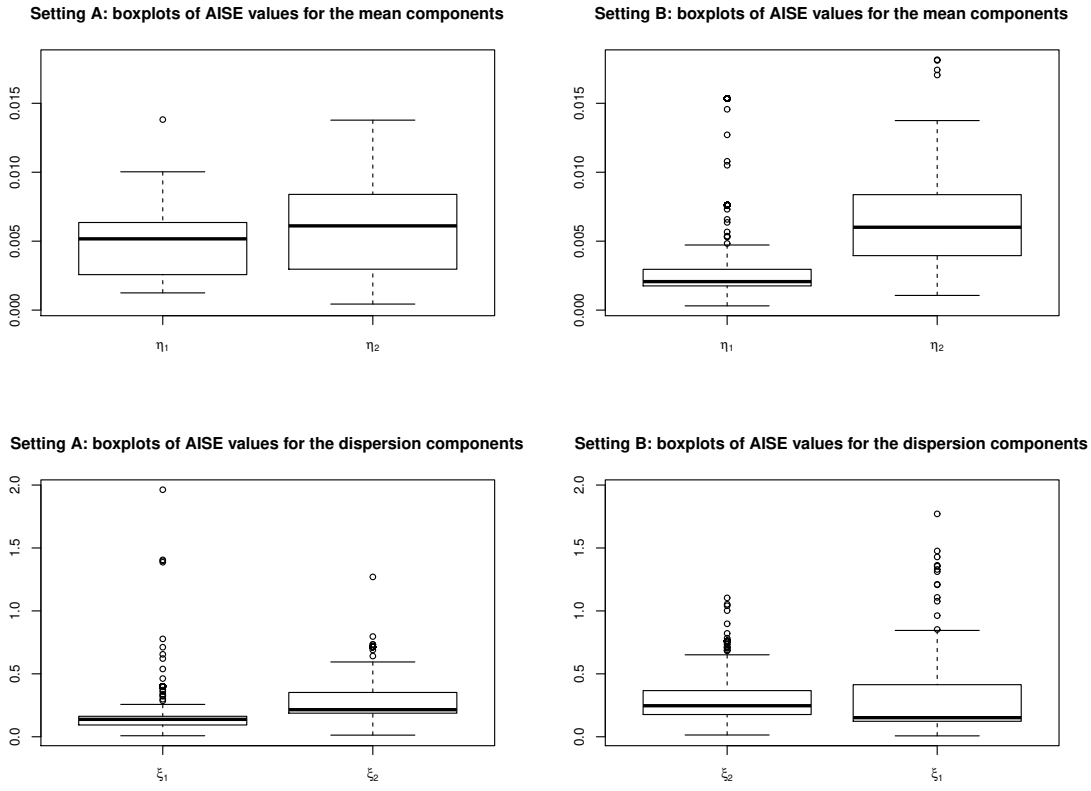

Figure 4.5: *Poisson data, sample size* $n = 300$. *Boxplots of the AISE-values for the mean and the dispersion function for the model setting A (left panels) and the model setting B (right panels).*

4.2 Comparisons with other methods

One might wonder what happens if one ignores the fact that the dispersion function changes with the values of the covariates. In Figures 4.9 and 4.10 we show the performances of the global mean (i.e. $\eta(x_1, x_2)$) and dispersion estimation (i.e. $\xi(x_1, x_2)$) when (i) we estimate the dispersion function as a constant; and (ii) we estimate the dispersion effectively as a function of the two covariates. Since in our simulation models the dispersion function is varying with the values of the covariates, considering this dispersion as a constant parameter (as is done in standard GAM modeling) is expected to deliver poorer results also in terms of mean estimation. The simulation results summarized in Figures 4.9 and 4.10 confirm this. Here the abbreviation 'DoubleGAM' refers to the method discussed in this paper and the abbreviation 'GAM' refers to the results for a standard GAM fitting. The poor results for the standard GAM fitting are mostly visible in the case of the Poisson data because of the larger sample size: not surprisingly, in order to obtain a good estimate for the dispersion function, which can be considerably beneficial for the mean function estimation, larger sample sizes are desirable.

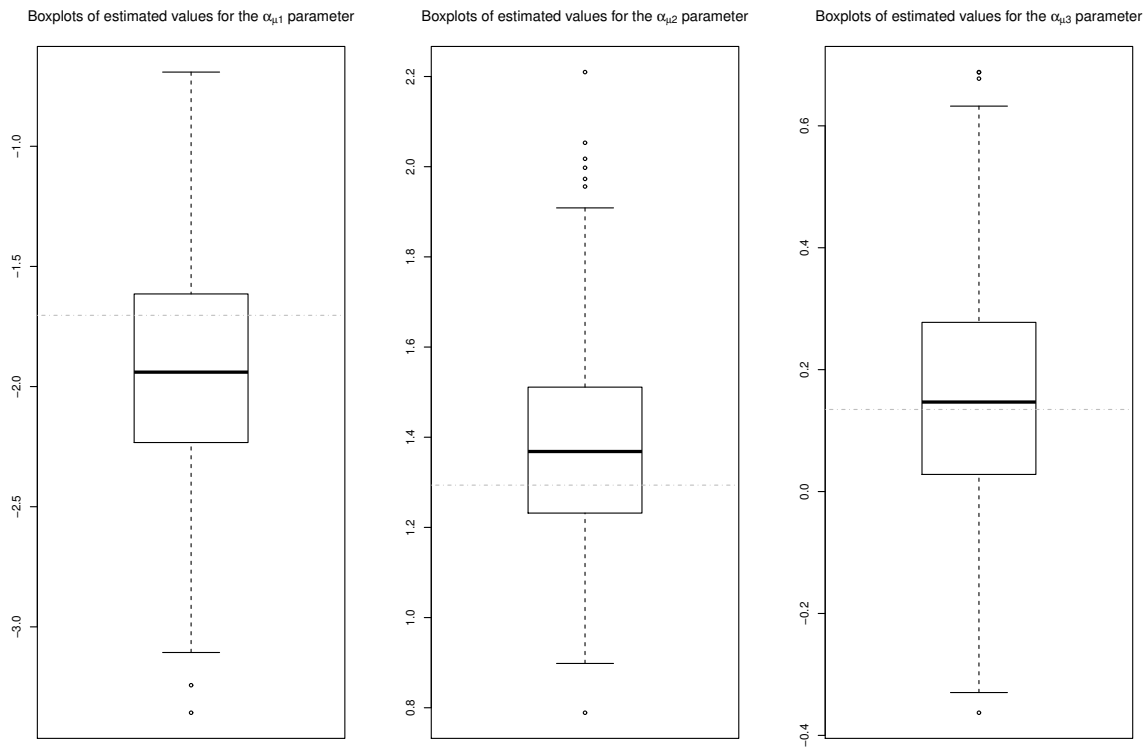

Figure 4.6: *Normal data: mean function estimation. Boxplots of the estimated parameters of the parametric mean component. The dashed lines indicates the true values* $\boldsymbol{\alpha}_{\mu}^{P} = (-18, 7, 6)$.

Rigby and Stasinopoulos (2005) also propose methods to flexibly estimate the mean and dispersion functions of over-dispersed data, although in their approach one needs to assume that the data come from a specific distribution which can handle over-dispersed data. It is common practice to assume that over-dispersed count data come from a Negative Binomial distribution, which is an extension of a Poisson distribution using some hierarchical modeling. One of the major drawbacks of this approach is that, unlike the double exponential family, it does not allow the data to be under-dispersed.

In order to compare the finite-sample performance of the proposed method with those found in Rigby and Stasinopoulos (2005) we perform a simulation study, using count data with the same mean structure used in Section 4.1. We can only compare the two methods in the situation of over-dispersion. We therefore changed some of the parameters for the dispersion function in the Poisson model setting of Section 4.1 to make sure that $\gamma(\boldsymbol{x})$ would always be larger than 1. Moreover, since the two methods assume that the data come from two different distributions, we analyze the data with both methods and generate samples using both distributions as the underlying data-generation process. It can be shown that for a r.v. $(Y|\boldsymbol{X}_d = \boldsymbol{x}_d) \sim$ Negative Binomial$(\mu(\boldsymbol{x}_d), \sigma(\boldsymbol{x}_d))$ we have:

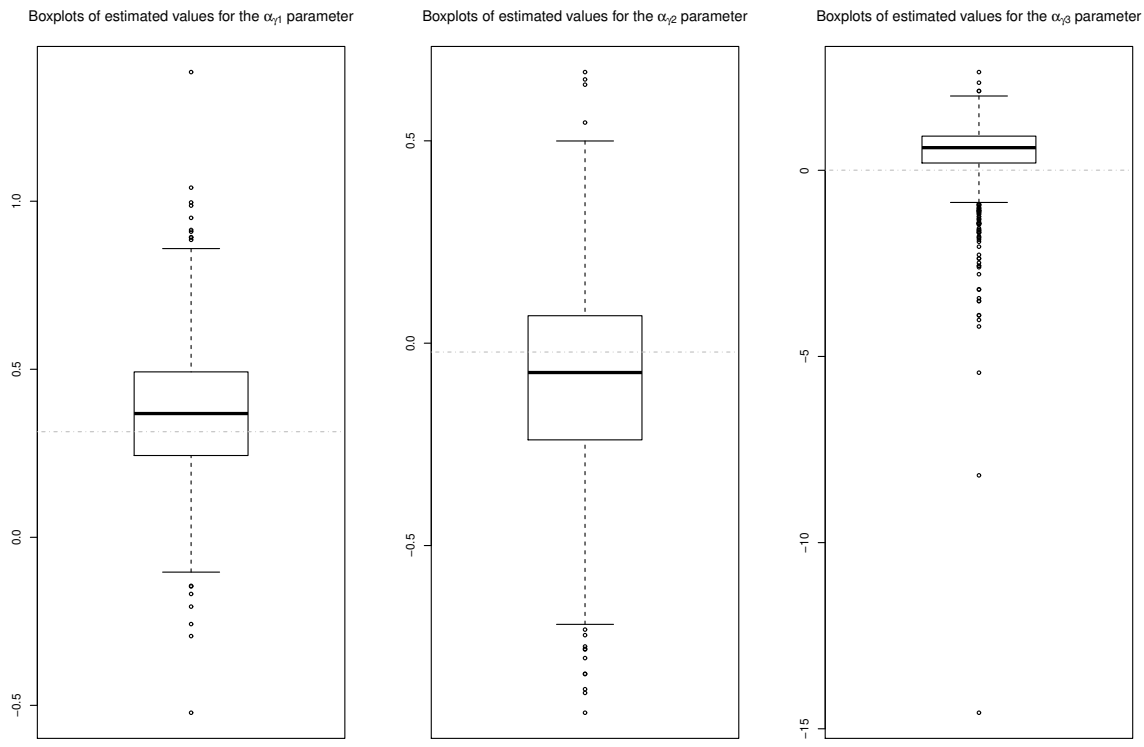

Figure 4.7: *Normal data: dispersion function estimation. Boxplots of the estimated parameters of the parametric dispersion component. The dashed lines indicates the true values* $\boldsymbol{\alpha}_\gamma^P = (1.5, -0.5, 0.15)$.

$E[Y|\boldsymbol{X}_d = \boldsymbol{x}_d] = \mu(\boldsymbol{x}_d)$ and $\text{Var}[Y|\boldsymbol{X}_d = \boldsymbol{x}_d] = (1 + \sigma(\boldsymbol{x}_d))\mu(\boldsymbol{x}_d)$, with $\sigma(\boldsymbol{x}_d) > 0$. From this it is easy to see the relationship between the $\sigma(\boldsymbol{x}_d)$ function for the Negative Binomial and the $\gamma(\boldsymbol{x}_d)$ in the double exponential family. The data simulated for the double exponential family have a dispersion function $\gamma^*(x_1, x_2) = \exp\left\{\alpha^*_{\gamma 0} + \xi^*_1(x_1) + \xi^*_2(x_2)\right\}$ where

$$\alpha^*_{\gamma 0} = 0.3 \qquad \xi^*_1(x) = 0.7\exp(0.5x^2) \qquad \text{and} \qquad \xi^*_2(x) = 0.2\left(\sin(90 + 6x) + 4\cos(x^{1.5})\right) ,$$

while for the data generated using the Negative Binomial we take the dispersion function $\sigma^*(x_1, x_2) = \gamma^*(x_1, x_2) - 1$. We generate 1000 samples of size $n = 300$ data points for each distribution and then analyze the data using both the estimation techniques. In Figure 4.11 we show the results regarding count data generated from a distribution belonging to the double exponential family and modeled using either the proposed Double GAM method or the GAMLSS technique of Rigby and Stasinopoulos (2005). Figures 4.11 depicts the boxplots of the AISE-values for both methods of estimation of γ^* and σ^*. In Figure 4.12 instead we compare the finite-sample performance of the two methods when the data are generated via a Negative Binomial. For the mean estimation both methods

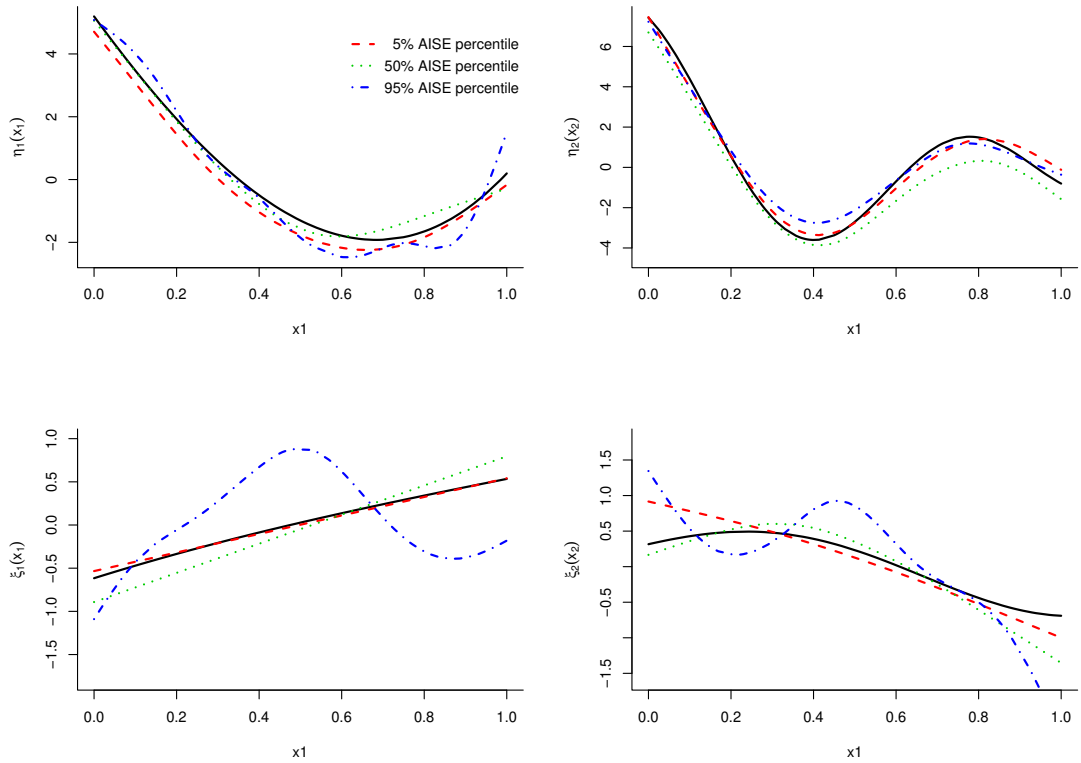

Figure 4.8: *Normal model setting A: true mean and dispersion components with representative estimates when the parametric functions are estimated via P-splines.*

perform comparable, but for the dispersion/variance estimation the Double GAM method performs slightly better.

Apart from estimation quality also the computational efforts should be taken into account. For both data-generation processes, the Double GAM estimation method took on average 63 seconds, whereas the GAMLSS estimation needed on average 19 seconds. Even if this difference in computing time is quite relevant, both methods are computationally very feasible. The proposed method has the considerable advantage of allowing for various departures of a constant dispersion modeling (e.g. under-dispersion as well as over-dispersion, or combinations of both).

5 Real data example: the Boston housing data

In this section we illustrate the discussed flexible estimation method on the Boston housing data. These are data, presented in Harrison and Rubinfeld (1978), regarding the prices of houses in Boston in 1970. The interest is in the median price of owner-occupied homes in

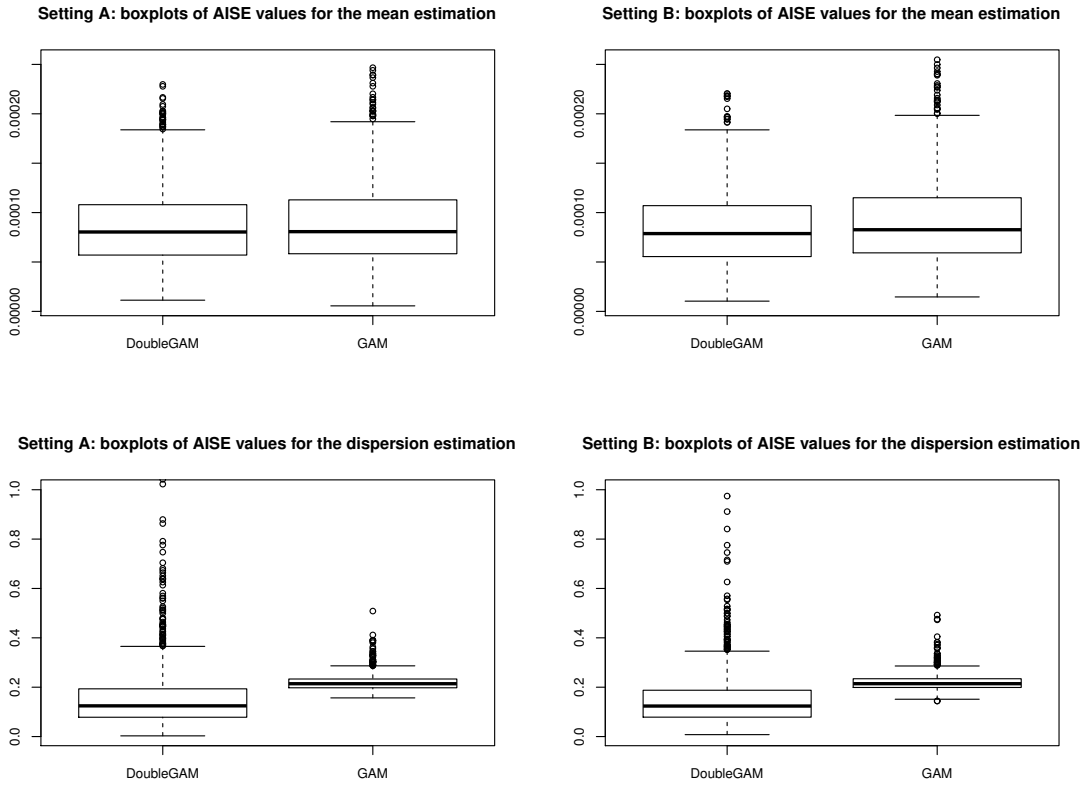

Figure 4.9: *Normal model settings. Boxplots of the AISE-values for the mean and the dispersion function for the model setting A (left panels) and the model setting B (right panels).*

the Boston area expressed in $1000 \times$ US Dollars. The proposed method allows for modeling heteroscedasticity in the covariates, either in a parametric or a nonparametric way. In addition, different covariates can enter in the mean and the dispersion function modeling. In the upper panels of Figure 5.1 we see the estimated influence of three covariates, a first covariate being the weighted mean of distances to five Boston employment centers (`dis`), a second covariate (denoted by `black`) and defined as $1000(\text{Bk} - 0.63)^2$ where Bk is the proportion of blacks by town, and a third covariate defined as the percentage of the population which is in the lower status (denoted as `lstat`). The covariates `dis` and `black` enter the model parametrically as polynomial functions of degree 3, while the covariate `lstat` is modeled nonparametrically via P-splines. The lower panels of Figure 5.1 depict the estimated components of the dispersion function. Again, `lstat` enters the model nonparametrically, while the relationship between the dispersion function and `dis` is modeled as a polynomial of degree 1. We did not include `black` in the dispersion function model as that component was not significant. From our analysis it is clear that the covariate `lstat` has the most complex influence on the response variable, and requires

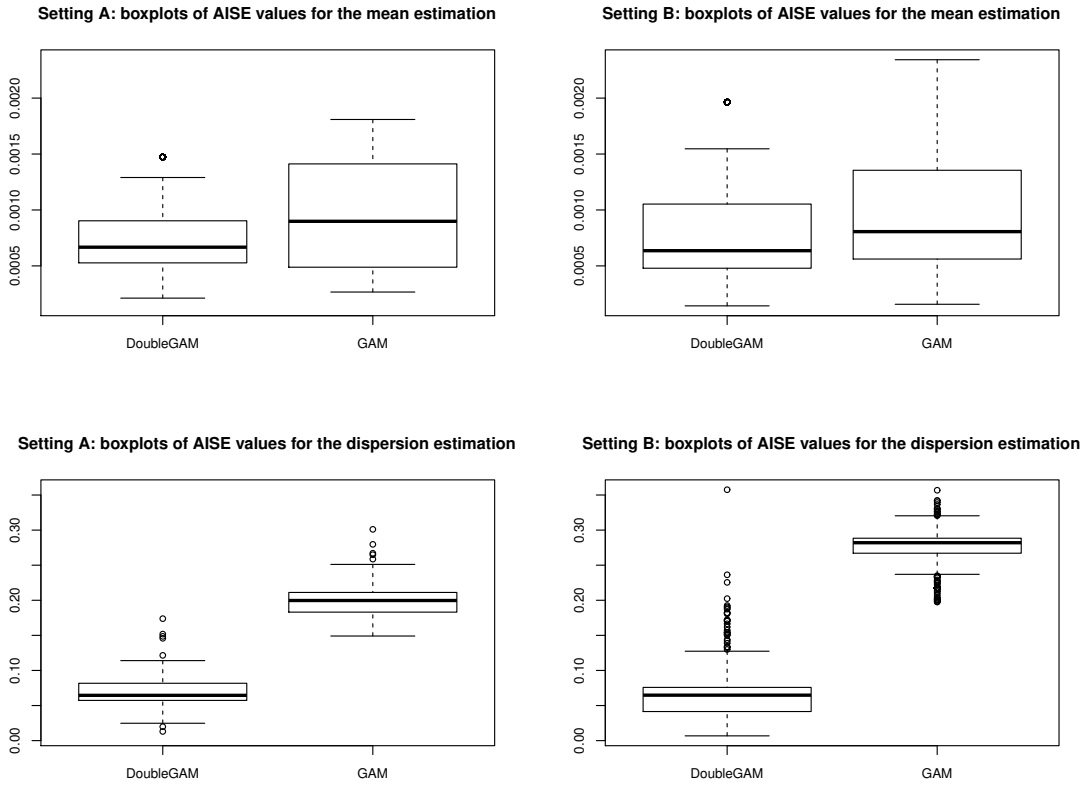

Figure 4.10: *Poisson model settings, sample size* $n = 300$. *Boxplots of the AISE-values for the mean and the dispersion function for the model setting A (left panels) and the model setting B (right panels).*

nonparametric estimation techniques.

In Figure 5.1 we also show 95% confidence bands for each of the estimated components. These are based on the approximate distributions for $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_\gamma$

$$\hat{\boldsymbol{\alpha}}_\mu \sim \mathrm{N}(\boldsymbol{\alpha}_\mu, (\boldsymbol{B}_\mu^T \boldsymbol{W}_\mu \boldsymbol{B}_\mu + \boldsymbol{P}_\mu)^{-1} \boldsymbol{B}_\mu^T \boldsymbol{W}_\mu \boldsymbol{B}_\mu (\boldsymbol{B}_\mu^T \boldsymbol{W}_\mu \boldsymbol{B}_\mu + \boldsymbol{P}_\mu)$$

and

$$\hat{\boldsymbol{\alpha}}_\gamma \sim \mathrm{N}(\boldsymbol{\alpha}_\gamma, (\boldsymbol{B}_\gamma^T \boldsymbol{W}_\gamma \boldsymbol{B}_\gamma + \boldsymbol{P}_\gamma)^{-1} \boldsymbol{B}_\gamma^T \boldsymbol{W}_\gamma \boldsymbol{B}_\gamma (\boldsymbol{B}_\gamma^T \boldsymbol{W}_\gamma \boldsymbol{B}_\gamma + \boldsymbol{P}_\gamma) \ ,$$

with $\boldsymbol{W}_\mu$ and $\boldsymbol{W}_\gamma$ the weight matrices as in (3.7) and (3.8) respectively, associated with the values of the estimated parameter vectors after convergence has been reached. These approximations are a generalization of the approximations used to do inference for generalized linear models. See Wood (2006b) for a more complete discussion on ways to build confidence bands for GAMs.

Note that the confidence bands are wider in areas with lesser data points, as to be expected.

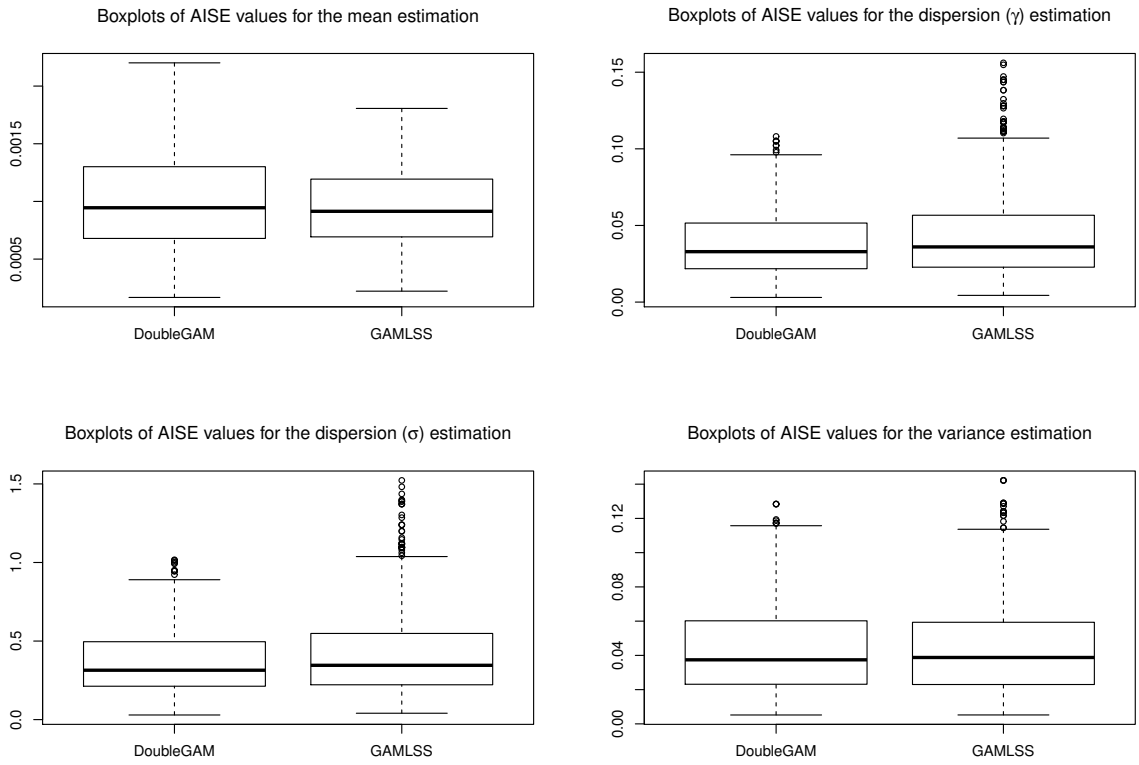

Figure 4.11: *Data generation: double exponential family. Boxplots of the AISE-values for all estimated functions using the Double GAM method or GAMLSS method.*

Acknowledgements

The authors thank the Associate Editor and the reviewers for their valuable comments. Financial 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, and from the FWO-project G.0328.08N of the Flemish Science Foundation.

References

Antoniadis, A., Gijbels, I. and Nikolova, M. (2011). Penalized likelihood regression for generalized linear models with nonquadratic penalties. *The Annals of the Institute of Statistical Mathematics*, **63**, 585–615.

Antoniadis, A., Gijbels, I. and Verhasselt, A. (2011). Variable selection in additive models using P-splines. *Under revision.*

Croux, C., Gijbels, I. and Prosdocimi, I. (2011). Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics*, to appear. DOI: 10.1111/j.1541-0420.2011.01630.x

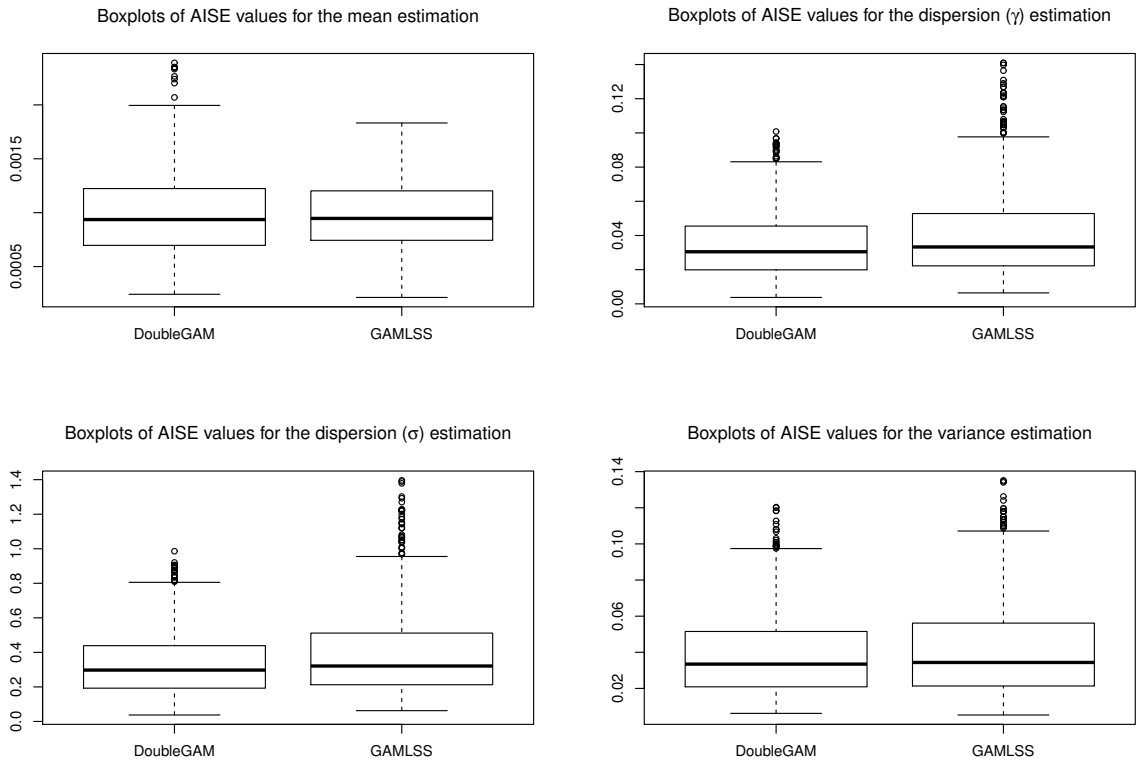

Figure 4.12: *Data generation: Negative Binomial. Boxplots of the AISE-values for all estimated functions using the Double GAM method or GAMLSS method.*

Claeskens, G., Krivobokova, T. & Opsomer, J.D. (2009). "Asymptotic properties of penalized spline estimators". *Biometrika*, **96**, 529–544.

Davidian, M. and Carroll, R.J. (1988). A note on extended Quasi-likelihood. *Journal of the Royal Statistical Society, Series B*, **50**, 74–82.

Efron, B. (1986). Double Exponential Families and their Use in Generalized Linear Regression. *Journal of the American Statistical Association*, **81**, 809–721.

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

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**, 645–660.

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

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*, **38**, 2391–2411.

Gu, C. and Xiang, D. (2001). Cross-validating non-Gaussian Data: generalized approximate cross-validation revisited. *Journal of Computational and Graphical Statistics*,

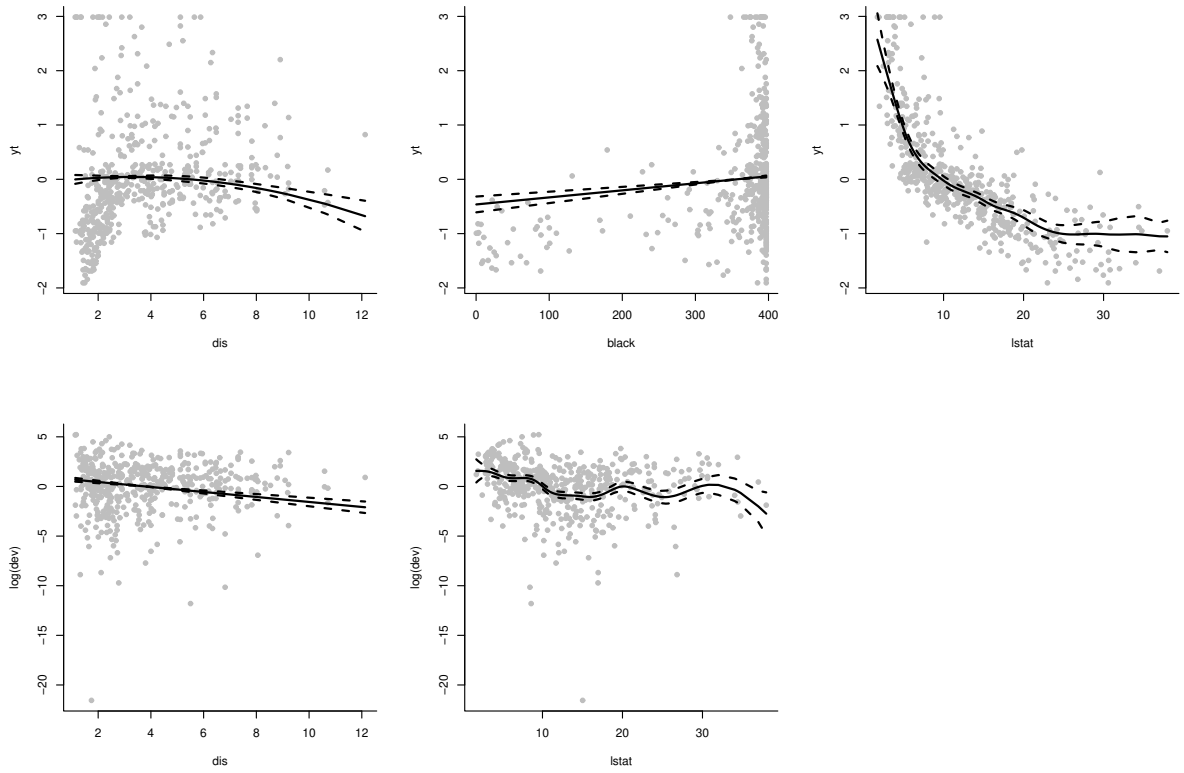

Figure 5.1: *The Boston housing data: mean and dispersion estimation. Top panels: (centered) data together with the estimated mean components η_1, η_2 and η_3. Lower panels: (centered) residuals are plotted with the ξ_1 and ξ_2 estimates. The dotted lines are the approximate 95% confidence bands.*

10, 581–591.

Hall, P. and Carroll, R.J. (1989). Variance function estimation in regression: the effect of estimating the mean. *Journal of the Royal Statistical Society, Series B*, **51**, 3–14.

Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, **5**, 81–102.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, **1**, 297–310.

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

Horowitz, J., Klemelä, J. & Mammen, E. (2006). "Optimal estimation in additive regression models". *Bernoulli*, **12**, 271–298.

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

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

Nelder, J.A. and Lee, Y. (1992). Likelihood, Quasi-likelihood and Pseudolikelihood: some comparisons. *Journal of the Royal Statistical Society, Series B*, **54**, 273–284.

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

Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, Issue 7.

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

Wood, S.N. (2006b). On confidence intervals for Generalized Additive Models based on penalized regression splines. *Australian & New Zealand Journal of Statistic*s, **48**, 445–464.

Wood, S.N. (2008). Fast stable direct fitting and smoothness selection for Generalized Additive Models. *Journal of the Royal Statistical Society, Series B*, **70**, 495–518.