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

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.

October 24, 2010

Abstract

We analyze data on abortion rate in Italy with a particular focus on different behaviours in different regions in Italy. The aim is to try to reveal the relationship between the abortion rate and several covariates that describe in some way the modernity of the region and the condition of the women there. The data are mostly underdispersed and the degree of underdispersion also varies with the covariates. To analyze these data recent techniques for flexible modelling of a mean and dispersion function in a double exponential family framework are further developed now in a Generalized Additive Model context for dealing with the multivariate setup. The appealing unified framework and approach even allow to semi-parametric modelling of the covariates without any additional efforts. The methodology is illustrated on ozone level data, and leads to interesting findings in the Italian abortion data.

Keywords and phrases: dispersion function, heteroscedasticity, mean function, non-parametric estimation, overdispersion, P-splines, underdispersion.

1 Introduction

In 1978 the Italian parlement approved a very disputed law which made induced abortion (IA) legal. Also it requested the Italian Health Ministry to provide a yearly report on data collected on induced abortions. This allows the Italian National Statistics Institute (ISTAT) and the Italian National Health System to regularly produce reports on the state of induced abortion in Italy (see e.g. Boccuzzo (2000) and Spinelli *et al.* (2006)) and researchers to study the phenomenon (e.g. Figà-Talamanca *et al.* (1986), Salvini Bettarini and Schifini D'Andrea (1996) or Spinelli and Grandolfo (2001)). To our knowledge, the study of the phenomenon has mostly focussed on a temporal perspective, on the geographical differences and on basic socio-economical characteristics of women undergoing

IA. A brief general overview of the historical changes and the present state of induced abortion in Italy is given in Section 6.

In this work instead we intend to study the relationship between the induced abortion rate and other covariates via regression-type models. For such an analysis we use the very rich ISTAT dataset that publishes, among others, for each year in each Italian province, the induced abortion rate (AR), defined as 1000 times the number of induced abortion over the average resident female population aged between 15 and 49. Also additional quantities are available for different years in each Italian province. For our analysis we focus on some of these variables in the ISTAT data base, and study the relationship between them and the abortion rate, and this for the data for the year 2001. In that year a general families census was held, and we can therefore use also informations coming from those data (also made available by the ISTAT). The data for 2001 are quite complete for most variables, although the whole information on the abortion rate for the region of Campania is missing for that year. Therefore we excluded this region from the analysis.

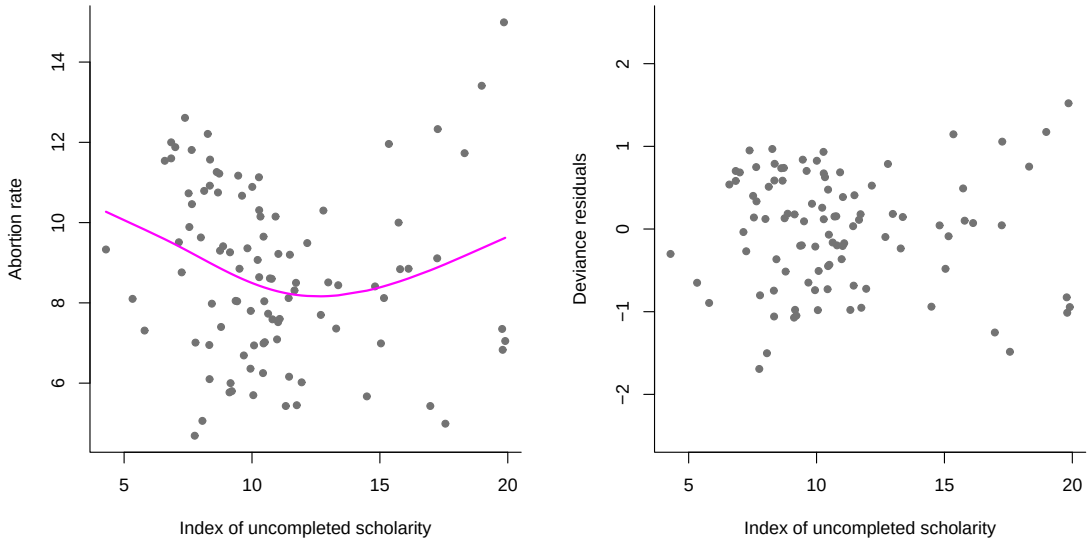

Figure 1.1: *The Italian abortion data. Left panel: the data together with an estimate for the mean function as a function of the index of uncompleted scholarity. Right panel: the deviance residuals from the fitted mean function.*

One of the variables included in the analysis is the Index of not finishing compulsory education for the female population between 15 and 52 (computed as the ratio between women who did not obtained the basic middle school diploma and the total female population). We would like to investigate how this variable influences the abortion rate (AR).

In Figure 1.1 we present a scatterplot of the data: the abortion rate data as response variable (vertical axis) and the Index of uncompleted scholarity as covariate (horizontal axis). Included in Figure 1.1 is also a smooth estimate of the mean regression function depicting the mean influence that the Index of uncompleted scholarity has on the abortion rate. This mean function has been estimated via P-splines (see Section 3), modelling the data as coming from a Poisson distribution. The estimated function has a quadratic shape that is strongly influenced by five data points belonging to the Puglia region which have high levels for the index of uncompleted scholarity and also high abortion rate. Puglia is in fact a region that shows extremely high abortion rates and in general seems to behave differently compared to other Italian regions (see Section 6). The estimated mean function nevertheless indicates that up to a certain point the AR decreases in provinces where women are less educated. For provinces with index of uncompleted scholarity higher than 13 the decrease stops and we see an increase, due mainly to the presence of the data of Puglia.

The right panel of Figure 1.1 shows the deviance residuals of the fitted model. These residuals should vary uniformly around zero, while we can see that the size of the residuals does not seem to be constant but changes for different values of the covariate. The variance we observe from the data in fact does not correspond to the one we would expect to find from the theoretical model (the assumed Poisson model). Therefore we introduce an extra dispersion parameter to model this anomaly in the variance (see Section 5). Once we estimate the dispersion function, we can smoothly estimate the variance function and standardize the deviance residuals by dividing them for this estimated dispersion. The resulting estimated variance function and the standardized deviance residuals are plotted in Figure 1.2. The standardized residuals show a more constant variance than the ones of Figure 1.1 (right panels). The estimated variance function is in a large part of the domain much lower than the one we would expect from the theoretical model (a phenomenon called underdispersion), and only in the right part of the domain the estimated variance is slightly larger than the theoretical one (a phenomenon called overdispersion). The presented estimates have been obtained via P-splines (see Section 5.2 for a brief description or see Gijbels *et al.* (2010) for a detailed study of flexible estimation of the dispersion function in the univariate case). In conclusion, Figure 1.1 illustrates the need for extra modeling efforts for the dispersion, in order to account for the observed over- and underdispersion phenomenon.

So far we only briefly illustrated how one of the covariates influences the response

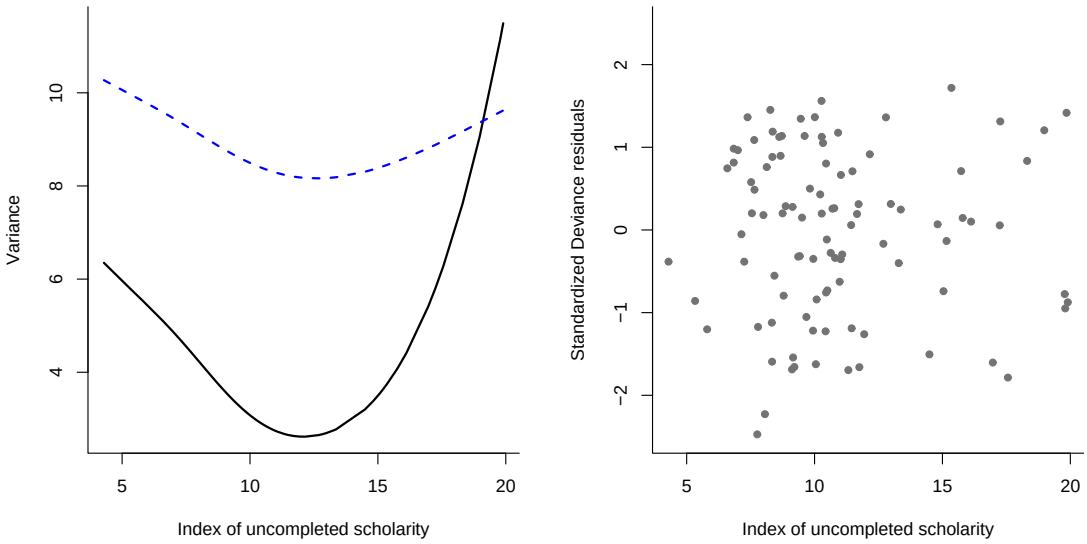

Figure 1.2: *The Italian abortion data. Left panel: estimated variance function in case of a constant dispersion (dashed curve) and estimated variance obtained when also estimating an extra dispersion parameter (solid curve). Right panel: the standardized deviance residuals from the fit with estimated dispersion.*

variable (the AR). It is of interest however to find out how the AR varies as a function of not only one covariate but more than one (Generalized Linear Models, see Section 2). Moreover we would like to be able to obtain flexible estimates, rather than polynomial shapes (Generalized additive models, see Section 4). As is seen from the above example the variance of the analyzed data does not behave as one would expect from the theoretical model (the Poisson model) and therefore in Section 5 we introduce a new class of models which extends the usual GLM models and allow us to also estimate the dispersion as a function of covariates. By applying the GAM methodology in these new models we obtain smooth estimates for the dispersion function. In Section 6 we return to the analysis of the abortion rate in the Italian provinces, obtaining smooth estimates for both the mean and the dispersion as function of covariates. The analysis reveals some very interesting findings. Finally, in Section 7 the perfomance of the proposed methods is further investigated by means of a simulation study.

2 A brief introduction to GLMs and GAMs

Generalized Additive Models (GAMs, Hastie and Tibshirani (1986 and 1990)) could be described as a smooth extension of Generalized Linear Models (GLMs), a generalization themselves of linear models (see McCullagh and Nelder (1989) for a complete discussion of GLMs). Since the seminal book of Hastie and Tibshirani (1990) a lot of work has been done on extending and developping GAMs; see for example the recent monograph of Wood (2006a) and the paper by Marx and Eilers (1998), among others.

In the GLM setting one is interested in studying the relationship between the mean of a response variable Y and a set of covariates $\boldsymbol{X}_d = (X_1, \ldots, X_d)$, assuming the relationship to be linear (or polynomial). Generalized Additive Models (GAM) extend GLM by taking the relationship between the expected value of Y and the covariates to be smooth and unknown rather than polynomial. In both frameworks, one assumes that the response variable Y given $\boldsymbol{X}_d = \boldsymbol{x}_d$, with $\boldsymbol{x}_d = (x_1, \ldots, x_d) \in I\!R^d$, follows a distribution coming from the Exponential Family of Distributions with conditional density function

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

where $b(\cdot)$ and $c(\cdot; \phi)$ are known functions, identifying specific distributions and ϕ is a scale parameter. 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 $\mu(\boldsymbol{x}_d) = \mathrm{E}[Y|\boldsymbol{X}_d = \boldsymbol{x}_d] = b'(\theta(\boldsymbol{x}_d))$ and $\mathrm{Var}[Y|\boldsymbol{X}_d = \boldsymbol{x}_d] = \phi b''(\theta(\boldsymbol{x}_d))$. In GLM and GAM one more generally models $E[Y|\boldsymbol{X}_d = \boldsymbol{x}_d]$ by introducing a link function $g(\cdot)$ which links the expected value of the conditional distribution to $\eta(\boldsymbol{x}_d)$: $\eta(\boldsymbol{x}_d) = g(\mu(\boldsymbol{x}_d))$. 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 the remainder of the paper we will use, unless otherwise stated, canonical link functions. In a GLM setting the function $\eta(\boldsymbol{x}_d)$ is taken to be a linear function of the covariates. In GAM finally $\eta(\boldsymbol{x}_d)$ is modelled as a linear combination of smooth (unknown) functions of the explanatory variables:

$$g(\mu(\boldsymbol{x}_d)) = \eta(\boldsymbol{x}_d) = \alpha_0 + \eta_1(x_1) + \ldots + \eta_d(x_d) = \alpha_0 + \sum_{j=1}^{d} \eta_j(x_j)\ , \tag{2.2}$$

where $\eta_j(x_j)$ is a smooth function which needs to be determined. When taking all the $\eta_j(x_j)$ to be of a parametric linear shape, we fall back into the standard GLM setting. Of particular interest is the situation in which we have some variables entering the model (2.2) in a parametric linear fashion (e.g. as a polynomial), and others entering in a nonparametric fashion (i.e. via a smooth unknown function). The latter situation leads

to a semi-parametric model. Generally, the model in (2.2) is defined up to a constant and is not identifiable, in the sense that we could add and substract the same constant β_0 from two $\eta_i(x_i)$ and $\eta_j(x_j)$ components (with $i \neq j$), without this affecting the final fit. In order to avoid this identifiability issue we introduce a constraint on the expected value of each smooth component: $\mathrm{E}[\eta_j(X_j)] = 0$. Different smoothing methods can be used to estimate the smooth (unknown) functions. In this paper we follow Marx and Eilers (1998) by using a direct modelling via penalized splines to estimate the smooth functions. With this modelling approach Generalized Additive Models are reduced to penalized Generalized Linear Models, with a relatively small number of parameters to be estimated. Moreover, P-splines allows to fit smooth function with a very easy set up and have attractive numerical properties. See Eilers and Marx (1996) or Section 4.1 in Wood (2006a) for more extensive discussion on the advantages of P-splines on other smoothing techniques. In the next section we briefly introduce Penalized splines (P-splines) smoothing techniques and necessary notations.

3 Penalized splines: a brief overview

To briefly introduce Penalized splines we refer to a GLM setting with only one covariate X. The main idea, as introduced in Eilers and Marx (1996), is to extend the traditional GLM allowing $\mathrm{E}[Y|X = x] = \mu(x)$ to be a smooth (unknown) function. Assuming that $(Y|X = x) \sim \mathrm{EF}(\theta(x), \phi)$, we model the linear predictor $\eta(x)$ as a linear combination of B-spline basis functions: $\eta(x) = \alpha_1 B_1(x) + \ldots + \alpha_K B_K(x)$. 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 $K = k + p + 1$. We then can approximate the unknown $\eta(\cdot)$ function in the space of B-spline basis functions

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

where we denote $\boldsymbol{B}(x) = (B_1(x), \ldots, B_K(x))^T$ the B-splines base and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^T$, the unknown vector of parameters. Here the superscript T denotes the transpose of a vector or a matrix. Taking a large set of B-spline functions leads to a better approximation in (3.1) but the resulting fit will also show a large variability. To control this overfitting

a penalty term is introduced in the likelihood. In particular in P-splines regression we use a penalty based on the 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 difference operator: $\Delta\alpha_j = \alpha_j - \alpha_{j-1}$, $\Delta^2\alpha_j = \Delta\Delta\alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$ and so on.

For data points $(\boldsymbol{x}, \boldsymbol{y}) = ((x_1, y_1), \ldots, (x_n, y_n))^T$, an i.i.d. realisation 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 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 the so-called smoothing parameter and $\mathbf{1}_n = (1, 1, \ldots, 1)^T$ denotes the unit vector of length n. By $b(\boldsymbol{B}\boldsymbol{\alpha})$ we mean to apply the function $b(\ )$ to each element of the vector $\boldsymbol{B}\boldsymbol{\alpha}$, so that $b(\boldsymbol{B}\boldsymbol{\alpha}) = (b(\boldsymbol{B}^T(x_1)\boldsymbol{\alpha}), \ldots, b(\boldsymbol{B}^T(x_n)\boldsymbol{\alpha}))^T$. 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 $\sum_{j=m+1}^{K}(\Delta^m\alpha_j)^2$. See Gijbels *et al.* (2010) for more details.

Maximization of (3.2) with respect to $\boldsymbol{\alpha}$ leads to the maximum penalized likelihood estimator of $\boldsymbol{\alpha}$. This estimator is obtained by using iterative procedures, like Fisher scoring. After some algebra (for details see e.g. Eilers and Marx (1996) or Gijbels *et al.* (2010)) we find that, for a given λ and for a current value $\tilde{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$, an updated value for $\boldsymbol{\alpha}$ is obtained from the updating rule

$$\boldsymbol{\alpha} = (\boldsymbol{B}^T\tilde{\boldsymbol{W}}\boldsymbol{B} + \lambda\boldsymbol{D}_m^T\boldsymbol{D}_m)^{-1}\boldsymbol{B}^T\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}, \tag{3.3}$$

with $\tilde{\boldsymbol{z}}$ the current working variable vector

$$\tilde{\boldsymbol{z}} = \boldsymbol{B}\tilde{\boldsymbol{\alpha}} + (\boldsymbol{y} - b'(\boldsymbol{B}\tilde{\boldsymbol{\alpha}}))\frac{1}{b''(\boldsymbol{B}\tilde{\boldsymbol{\alpha}})}, \tag{3.4}$$

and $\tilde{\boldsymbol{W}}$ the current diagonal matrix

$$\tilde{\boldsymbol{W}} = \operatorname{diag}\left(\frac{1}{\phi}b''(\boldsymbol{B}\tilde{\boldsymbol{\alpha}})\right). \tag{3.5}$$

So far, λ was supposed to be given. The choice of λ is rather crucial though: larger values of λ correspond to smoother estimated function; lower values of λ instead, lead to more wiggly estimates (i.e. overfitting).

4 Direct P-Splines Generalized Additive Models

Marx and Eilers (1998) proposed to use P-splines in order to estimate the smooth components of a GAM model. Each smooth unknown component $\eta_j(x_j)$ in (2.2) is modelled as a linear combination of K_j B-splines and overfitting for the component is avoided by adding to the likelihood a penalty term based on the finite differences of adjacent coefficients.

This modelling can be easily extended to allow for parametric modelling for a subset of the covariates. We first illustrate this via an example. In Figure 4.1 (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 (Y) as a quadratic parametric function of the inversion base temperature (X_1) and as a smooth unknown function of the inversion base height (X_2) and the daggett pressure gradient (X_3):

$$g(\mu(x_1, x_2, x_3)) = \eta(x_1, x_2, x_3) = \alpha_0 + \eta_1(x_1) + \eta_2(x_2) + \eta_3(x_3) \ , \tag{4.1}$$

with $\eta_1(x_1) = x_1\alpha_{11} + x_1^2\alpha_{12} = \boldsymbol{B}_1^T(x_1)\boldsymbol{\alpha}_1$, a parametric quadratic function and with nonparametric components $\eta_2(x_2) = \boldsymbol{B}_2^T(x_2)\boldsymbol{\alpha}_2$ and $\eta_3(x_3) = \boldsymbol{B}_3^T(x_3)\boldsymbol{\alpha}_3$, where $\boldsymbol{B}_2(x_2) = (B_{2,1}(x_2), \ldots, B_{2,K_2}(x_2))^T$ (respectively $\boldsymbol{B}_3(x_3)$) is a B-spline basis of dimension K_2 (respectively K_3) and $\boldsymbol{\alpha}_2$ ($\boldsymbol{\alpha}_3$) is a vector of K_2 (K_3) parameters which needs to be estimated. The vectors $\boldsymbol{B}_1(x_1) = (x_1, x_1^2)^T$ and $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12})^T$ are the usual parametric design vector and vector of parameters found in the GLM setting. Taking $\boldsymbol{x}_d = (x_1, x_2, x_3)$, where $d = 3$, defining a 'design' matrix $\boldsymbol{B}(\boldsymbol{x}_d) = [1, \boldsymbol{B}_1^T(x_1), \boldsymbol{B}_2^T(x_2), \boldsymbol{B}_3^T(x_3)]^T$ and a vector of parameters $\boldsymbol{\alpha} = (\alpha_0, \boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \boldsymbol{\alpha}_3^T)^T$ we can finally rewrite (4.1) as

$$g(\mu(\boldsymbol{x}_d)) = \eta(\boldsymbol{x}_d) = \alpha_0 + \eta_1(x_1) + \eta_2(x_2) + \eta_3(x_3) = \boldsymbol{B}^T(\boldsymbol{x}_d)\boldsymbol{\alpha} \ , \tag{4.2}$$

which is of analogous form as (3.1) in the univariate case. Similarly as in Section 3 large sets of knots are used to build the B-splines bases $\boldsymbol{B}_2(x_2)$ and $\boldsymbol{B}_3(x_3)$, and, in order to avoid overfitting, we add two difference penalties of order m_1 and m_2 and introduce smoothing parameters λ_1 and λ_2 for governing the smoothness of each component. We define the block diagonal penalty matrix $\boldsymbol{P} = \text{blockdiag}\ (0, 0, 0, \lambda_1\boldsymbol{D}_{m_1}^T\boldsymbol{D}_{m_1}, \lambda_2\boldsymbol{D}_{m_2}^T\boldsymbol{D}_{m_2})$ where the zeros in the first three elements reflect the fact that we do not need to introduce any penalization for the parametric part of the fit (including the intercept).

For data points $(\boldsymbol{x}, \boldsymbol{y}) = ((x_{11}, x_{21}, x_{31}, y_1), \ldots, (x_{1n}, x_{2n}, x_{3n}, y_n))^T$, an i.i.d. realisation from (X_1, X_2, X_3, Y), we can build the design matrix $\boldsymbol{B}$, in which the ith row consists of $\boldsymbol{B}^T(\boldsymbol{x}_{d,i})$ with $\boldsymbol{x}_{d,i} = (x_{1i}, x_{2i}, x_{3i})$. For a given set of smoothing parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$,

we find similarly as in Section 3, that the updating rule for $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha} = (\boldsymbol{B}^T \tilde{\boldsymbol{W}} \boldsymbol{B} + \boldsymbol{P})^{-1} \boldsymbol{B}^T \tilde{\boldsymbol{W}} \tilde{\boldsymbol{z}}, \tag{4.3}$$

given the current value $\tilde{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$, with $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{W}}$ defined as in (3.4) and (3.5).

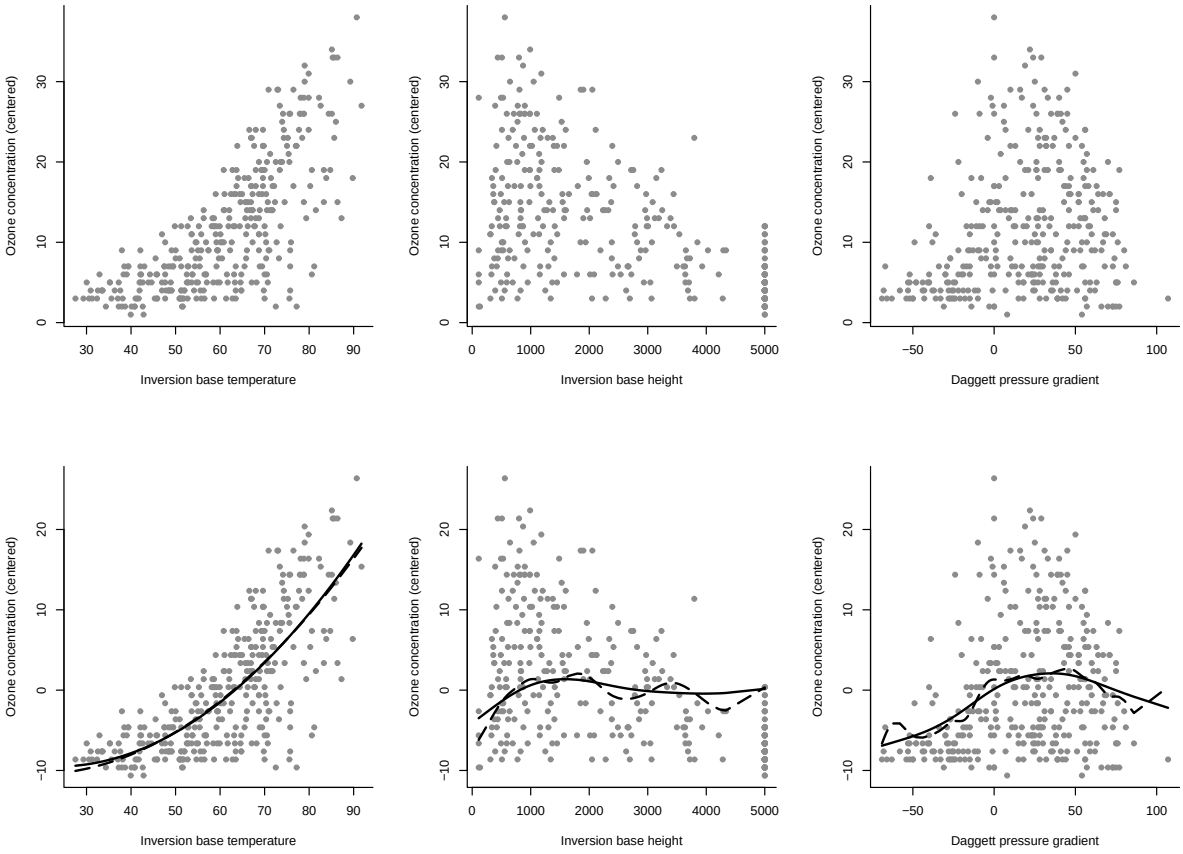

Figure 4.1: *The ozone data. Top panels: scatterplots of the (centered) data; Bottom panels: centered data together with estimates of the components* η_1, η_2 *and* η_3 *using* $\boldsymbol{\lambda} = (0.0475, 0.0025)$ *(dashed curves) or* $\boldsymbol{\lambda} = (4.5, 3.5)$ *(solid curves).*

We model the ozone data as coming from a (conditional) normal distribution, with canonical link the identity function. In Figure 4.1 (lower panels) we depict the estimates for the mean components $\eta_1(\cdot)$, $\eta_2(\cdot)$ and $\eta_3(\cdot)$ in (4.1). The estimates are centered around zero to avoid identifiability issues. As a consequence of our modelling strategy the estimated first component is a quadratic function whereas the two other components are modelled via by the P-splines approach and estimated nonparametrically. The first component shows a strong effect. The second and third component vary less around zero but show interesting shapes. See also Section 3.1 of Buja, Hastie and Tibshirani

(1989) for a discussion on the analysis of these data. In Figure 4.1 two different estimates corresponding to two different choices of the smoothing parameters values $\boldsymbol{\lambda}$ are shown: lower values of $\boldsymbol{\lambda}$ tend to give too wiggly estimates. In Section 5.3 we briefly discuss a data-driven way to choose these parameters.

Similary in general GAMs when the expected value of $Y|\boldsymbol{X}_d$ is modelled as a function of the covariates $\boldsymbol{X}_d = (X_1, \ldots, X_d)$ through the link function $g(\cdot)$ as in (2.2) one can allow both parametric and nonparametric dependencies of the covariates. More specifically assume that $d_{\mathrm{P}} \leq d$ covariates, say $\boldsymbol{X}_d^{\mathrm{P}} = (X_1, \ldots, X_{d_{\mathrm{P}}})$, enter the model parametrically, while $d_{\mathrm{NP}} = d - d_{\mathrm{P}}$ covariates, say $\boldsymbol{X}_d^{\mathrm{NP}} = (X_{d_{\mathrm{P}}+1}, \ldots, X_d)$ are modelled nonparametrically (via approximations with P-splines). Denote $\boldsymbol{x}_d = (x_1, \ldots, x_{d_{\mathrm{P}}}, x_{d_{\mathrm{P}}+1}, \ldots, x_d) = (\boldsymbol{x}_d^{\mathrm{P}}, \boldsymbol{x}_d^{\mathrm{NP}})$, and let $\boldsymbol{B}_j(x_j)$ be the parametric model basis of dimension K_j for modelling the parametric component of x_j for $j = 1, \ldots, d_{\mathrm{P}}$. The global basis for the parametric part is then $\boldsymbol{B}^{\mathrm{P}}(\boldsymbol{x}_d^{\mathrm{P}}) = \left[\boldsymbol{B}_1^T(x_1), \ldots, \boldsymbol{B}_{d_{\mathrm{P}}}^T(x_{d_{\mathrm{P}}})\right]$ with dimension $K_{\mathrm{P}} = \sum_{j=1}^{d_{\mathrm{P}}} K_j$. Similarly, we have d_{NP} sets of B-splines basis functions for the flexible modelling of the $d_{\mathrm{NP}} = d - d_{\mathrm{P}}$ other covariates, denoted by $\boldsymbol{B}_j(x_j)$ of dimension K_j, for $j = d_{\mathrm{P}} + 1, \ldots, d$. Denote by $\boldsymbol{B}^{\mathrm{NP}}(\boldsymbol{x}_d^{\mathrm{NP}}) = \left[\boldsymbol{B}_{d_{\mathrm{P}}+1}^T(x_{d_{\mathrm{P}}+1}) \ldots \boldsymbol{B}_d^T(x_d)\right]$ the global basis for this flexible (nonparametric) modelling part, of dimension $K_{\mathrm{NP}} = \sum_{j=d_{\mathrm{P}}+1}^{d} K_j$. Finally, defining $\boldsymbol{B}(\boldsymbol{x}_d) = \left[1, \boldsymbol{B}^{\mathrm{P}}(\boldsymbol{x}_d^{\mathrm{P}}), \boldsymbol{B}^{\mathrm{NP}}(\boldsymbol{x}_d^{\mathrm{NP}})\right]^T$ we obtain the model basis of dimension $K = 1 + K_{\mathrm{P}} + K_{\mathrm{NP}}$ and can rewrite (2.2) as

$$g(\mu(\boldsymbol{x}_d)) = \eta(\boldsymbol{x}_d) = \alpha_0 + \boldsymbol{B}^{\mathrm{P}}(\boldsymbol{x}_d^{\mathrm{P}})\boldsymbol{\alpha}^{\mathrm{P}} + \boldsymbol{B}^{\mathrm{NP}}(\boldsymbol{x}_d^{\mathrm{NP}})\boldsymbol{\alpha}^{\mathrm{NP}} = \boldsymbol{B}^T(\boldsymbol{x}_d)\boldsymbol{\alpha} \tag{4.4}$$

with $\boldsymbol{\alpha} = (\alpha_0, (\boldsymbol{\alpha}^{\mathrm{P}})^T, (\boldsymbol{\alpha}^{\mathrm{NP}})^T)^T$ the vector of unknown parameters, of dimension K, that need to be estimated. The B-splines bases that form $\boldsymbol{B}^{\mathrm{NP}}(\boldsymbol{x}_d^{\mathrm{NP}})$ are built, once again, using large sets of knots, and in order to avoid overfitting we introduce a vector of smoothing parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d_{\mathrm{NP}}})$ and difference order penalties of order $(m_1, \ldots, m_{\mathrm{NP}})$.

For data points $(\boldsymbol{x}, \boldsymbol{y}) = ((x_{11}, x_{21}, \ldots, x_{d1}, y_1), \ldots, (x_{1n}, x_{2n}, \ldots, x_{dn}, y_n))^T$, an i.i.d. realisation from $(X_1, \ldots, X_d, Y)$ we build the model matrix $\boldsymbol{B} = [\boldsymbol{1}_n \ \boldsymbol{B}^{\mathrm{P}} \ \boldsymbol{B}^{\mathrm{NP}}]$. For a given smoothing parameters vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d_{\mathrm{NP}}})$ the penalty matrix is $\boldsymbol{P} = \text{blockdiag}\ (0, \boldsymbol{0}_{K_{\mathrm{P}}}, \lambda_1 \boldsymbol{D}_{m_1}^T \boldsymbol{D}_{m_1}, \ldots, \lambda_{d_{\mathrm{NP}}} \boldsymbol{D}_{m_{d_{\mathrm{NP}}}}^T \boldsymbol{D}_{m_{d_{\mathrm{NP}}}})$, where the first $1 + K_{\mathrm{P}}$ zero elements reflect the fact that there is no need to penalize the parametric components of the model. The updating rule for $\boldsymbol{\alpha}$ is as in (4.3) with now the more general forms for $\boldsymbol{B}$ and $\boldsymbol{P}$.

5 Flexible modelling of the multivariate mean and dispersion function

When working within the GLM or GAM framework we are assuming that the response variable Y given $\boldsymbol{X}_d = \boldsymbol{x}_d$ comes from a distribution belonging to the one-parameter exponential family of distributions. This implies that the relationship between the mean and the variance is known and is $\text{Var}[Y|\boldsymbol{X}_d = \boldsymbol{x}_d] = \phi b''(\theta(\boldsymbol{x}_d))$ (see Section 2), with ϕ a constant. As illustrated via the example in Section 1 (see Figures 1.1 and 1.2) the form of the variance in the one-parameter exponential family can be too restrictive: data (especially count or proportion data) sometimes show a variance that is larger (respectively smaller) than the one we would expect from the theoretical model. This is referred to as overdispersion (respectively underdispersion). In addition, the amount of overdispersion (underdispersion) may vary as a function of a (set of) covariate(s). Also, continuous normal data can exhibit a variance that is not constant, as we usually assume in the linear model context, but actually depends on a set of covariates (a problem referred to as heteroscedasticity). Indeed, recall that for the normal model $b(\theta) = \theta^2/2$, $b''(\theta) = 1$ and $\phi = \sigma^2$ leading to a constant variance for the theoretical model.

Different methods have been proposed to analyze data whose variance differs from the theoretical one (see for example Hinde and Demétrio (1998) for a good review of some common methods). Overdispersed proportion or count data are often analyzed via beta-binomial or negative binomial distributions. These approaches though can only handle overdispersed data, and no similar techniques exist for handling underdispersed data. Another approach is to extend the original models and assume a general parametric form for the variance function, possibly introducing extra parameters in the model. Examples of this latter approach are the pseudo-likelihood approach (Davidian and Carroll (1987)), the Extended Quasi-Likelihood proposed by Nelder and Pregibon (1987) and the modelling via a Double Exponential family of distributions (Efron (1986)). These different approaches have each their advantages and disavantages: see Nelder and Lee (1992) and Davidian and Carroll (1988), among others, for a comparison of the different methods.

The aim of this paper is to allow for a *flexible modelling* of the dispersion *function* in the multivariate setup. To achieve this we focus on the Double Exponential Family framework, which allows to analyze with a *unique approach* heteroscedastic data and both overdispersed and underdispersed data, or even data showing both overdispersion and underdispersion. We will model the dispersion function, combining the GAM modelling

approach of Section 4 with the Double Exponential Family framework, extending as such the one-dimensional approach presented in Gijbels *et al.* (2010).

Previous work on flexible modelling of the mean and variance function include Chapter 14 of Ruppert *et al.* (2003), in which a mixed model approach for normal heteroscedastic data is proposed. Flexible estimation of the mean and variance function for normal heteroscedastic data is also studied by Yuan and Wahba (2004), who propose a procedure based on penalized likelihood. Nott (2006) also works within the Double Exponential Family of distributions and uses a Bayesian Mixed Model approach to semi-parametric modelling. Finally Rigby and Stasinopoulos' (2005) Generalized Additive Models for Location Scale and Shape (GAMLSS) also tackles the issue of variance function estimation by using hierarchical modelling and Bayesian reasoning. Hierarchical modelling is also the basis of the work of Lee and Nelder (2006). The compactness of the approach allowed by the Double Exponential Family framework is an advantage over an hierarchical modelling approach. As shown in Gijbels *et al.* (2010) the performance of two approaches is quite comparable, with the Double Exponential family having the advantage that it allows the modelling of data which show overdispersion in some areas of the covariates domain and underdispersion in others. It should also be mentioned that the direct penalized approach that is used here has the advantage of being computationally very reasonable. Estimates in fact are obtained via a unique Penalized Iterative Reweighted Least square, with no need for the backfitting algorithm or the numerical approximations that would be needed if taking a Mixed Model approach to smoothing. In this last approach in fact, one assumes that the spline coefficients would come from a specific random distributions. This results in the final likelihood of the problem to be an integral which can not be solved analytically and requires numerical approximation.

Variance (or dispersion) estimation is not only needed from the point of view of correcting models whose assumptions are too restrictive for real data, but in many cases the estimation of the variance function is of interest in itself. Carroll and Ruppert (1988) discuss the importance of variance estimation and provide a nice overview of parametric variance estimation methods in the linear regression context.

5.1 Double Exponential Family of distribution

Efron (1986) introduced the Double Exponential Family of Distributions, which will allow here for a unique compact framework for tackling both over- and under-dispersed data

and heteroscedasticity. See Section 5.2. The Double Exponential Family extends the usual one-parameter Exponential Family by introducing an extra parameter controlling the variance independently from the mean. For simplicity of presentation consider first the non-regression case (i.e. absence of covariates). Given an exponential family as in (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

$$\tilde{f}_Y(y;\theta,\phi,\gamma) = c(\theta,\gamma)\gamma^{-\frac{1}{2}} e_Y(y;\theta,\phi)^{\frac{1}{\gamma}} e_Y(y;\theta_S,\phi)^{1-\frac{1}{\gamma}} , \tag{5.1}$$

where $c(\theta,\gamma)$ is a normalizing constant, such that $\int_{-\infty}^{\infty} \tilde{f}_Y(y;\theta,\phi,\gamma)dy = 1$. This normalizing constant can be approximated (in first order) by 1 (see Efron (1986) and Lee and Nelder (2000) for a discussion on the quality of this approximation). Recalling that the deviance for a one-parameter exponential family is $d(y,\theta) = 2[\log(e_Y(y;\theta_S,\phi)) - \log(e_Y(y;\theta,\phi))]$, the approximation of (5.1) can be written as

$$f_Y(y;\theta,\phi,\gamma) = \gamma^{-\frac{1}{2}} e_Y(y;\theta,\phi)^{\frac{1}{\gamma}} e_Y(y;\theta_S,\phi)^{1-\frac{1}{\gamma}} = \gamma^{-\frac{1}{2}} \left\{ \exp\left[\frac{1}{2} d(y,\theta)\right] \right\}^{-\frac{1}{\gamma}} e_Y(y;\theta_S,\phi). \tag{5.2}$$

We refer to this 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 $\text{Var}[Y] = \gamma\phi b''(\theta)$. With this the interpretation of the γ parameter becomes clear: it is an extra parameter which governs the dispersion. When $\gamma = 1$ we fall back to the original one-parameter Exponential Family in (2.1), while the case of $\gamma > 1$ (respectively $\gamma < 1$) corresponds to overdispersion (underdispersion). In the case when Y is normally distributed γ coincides with the variance parameter (usually noted with σ^2), when taking $\phi = 1$, and the normalizing constant $c(\theta,\gamma)$ has exactly value 1. For other distributions, the actual value of 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. For a given value of θ, the estimation of γ will then lead to a unique estimation of the variance. When talking about variance or dispersion estimation we thus basically refer to the same issue: the estimation of γ. In particular we are interested in estimating γ as a (flexible) function of a set of covariates $(X_1,\ldots,X_d)$. For this we incorporate the GAM methods into the Double Exponential Family approach in the next subsection.

5.2 Flexible modelling of the mean and dispersion function

In Section 4 we have seen how to flexibly estimate the mean function of a dependent variable Y coming from a distribution belonging to the Exponential Family via P-splines in the GAM approach. We now wish also to obtain flexible estimates for the dispersion function. We therefore take the dependent variable Y as coming from the Double Exponential Family and we model both the mean and the dispersion as flexible functions of covariates via an extended GAM approach. We first explain the general framework and the estimation method, and then illustrate the proposed procedure on the ozone data example.

As for the flexible modelling of the mean we allow that part of the covariates $\boldsymbol{X}_d = (X_1, \ldots, X_d)$ enters the modelling of the dispersion function in a parametric fashion, whereas for the remaining part no specific parametric modelization can be justified. Given the set of d covariates $\boldsymbol{X}_d = (X_1, \ldots, X_d)$, we model the mean as a function of a certain set of d_μ covariates, with $d_\mu \leq d$. Also, we wish to have d_{P_μ} covariates entering the mean model in a parametric fashion and $d_{\mathrm{NP}_\mu} = d_\mu - d_{\mathrm{P}_\mu}$ covariates entering the model in a flexible (i.e. nonparametric) way. Similarly the dispersion function, denoted by $\gamma(\cdot)$, can be modelled as a function of a set of d_γ covariates ($d_\gamma \leq d$), possibly a different set than the one used to model the mean. Again d_{P_γ} covariates will enter the model parametrically and $d_{\mathrm{NP}_\gamma} = d_\gamma - d_{\mathrm{P}_\gamma}$ covariates are allowed to influence the mean response in a flexible fashion. We define $\boldsymbol{X}_{d_\mu} = (X_1, \ldots, X_{d_{\mathrm{P}_\mu}}, X_{d_{\mathrm{P}_\mu}+1}, \ldots, X_{d_\mu}) = (\boldsymbol{X}^{\mathrm{P}}_{d_\mu}, \boldsymbol{X}^{\mathrm{NP}}_{d_\mu})$ the set of covariates we use to model the mean function $\mu(\boldsymbol{x}_{d_\mu})$, and $\boldsymbol{X}_{d_\gamma} = (X_1, \ldots, X_{d_{P_\gamma}}, X_{d_{P_\gamma}+1}, \ldots, X_{d_\gamma}) = (\boldsymbol{X}^{\mathrm{P}}_{d_\gamma}, \boldsymbol{X}^{\mathrm{NP}}_{d_\gamma})$ the set of covariates we use to model the dispersion function $\gamma(\boldsymbol{x}_{d_\gamma})$.

Estimation of the dispersion function $\gamma(\boldsymbol{x}_{d_\gamma})$ is done via a P-splines technique and the introduction of a link function. Let $h(\cdot)$ be the link function such that $\gamma(\boldsymbol{x}_{d_\gamma}) = h(\xi(\boldsymbol{x}_{d_\gamma}))$ where $\xi(\boldsymbol{x}_{d_\gamma})$ will be modelled using a P-spline basis setup. Note that the link function $h(\cdot)$ should be chosen in such a way that $\gamma(\boldsymbol{x}_{d_\gamma})$ is always non-negative.

Summarizing, we assume $(Y|\boldsymbol{X}_d = \boldsymbol{x}_d) \sim \mathrm{DEF}(b(\theta(\boldsymbol{x}_{d_\mu})), \phi, \gamma(\boldsymbol{x}_{d_\gamma}))$, where ϕ is assumed to be constant and known, and we model

$$g(\mu(\boldsymbol{x}_{d_\mu})) = \eta(\boldsymbol{x}_{d_\mu}) = \alpha_{\mu 0} + \eta_1(x_1) + \ldots + \eta_{d_\mu}(x_{d_\mu}) \,, \tag{5.3}$$

and

$$h^{-1}(\gamma(\boldsymbol{x}_{d_\gamma})) = \xi(\boldsymbol{x}_{d_\gamma}) = \alpha_{\gamma 0} + \xi_1(x_1) + \ldots + \xi_{d_\gamma}(x_{d_\gamma}), \tag{5.4}$$

with $\alpha_{\mu 0}$ and $\alpha_{\gamma 0}$ the two intercept parameters for the mean and the dispersion function. The components $\eta_1(x_1), \ldots, \eta_{d_\mu}(x_{d_\mu})$ and $\xi_1(x_1), \ldots, \xi_{d_\gamma}(x_{d_\gamma})$ for the mean and the disper-

sion function, can then be either of a parametric or a flexible (i.e. nonparametric) type. The components allowed to have a unknown smooth shape are modelled via P-splines, so that we also need to introduce smoothing parameters $\boldsymbol{\lambda}^{\mu} = (\lambda_1^{\mu}, \ldots, \lambda_{d_{\mathrm{NP}_\mu}}^{\mu})$ and $\boldsymbol{\lambda}^{\gamma} = (\lambda_1^{\gamma}, \ldots, \lambda_{d_{\mathrm{NP}_\gamma}}^{\gamma})$ and difference order penalties of order $m_1, \ldots, m_{d_{\mathrm{NP}_\mu}}$ and $\ell_1, \ldots, \ell_{d_{\mathrm{NP}_\gamma}}$ to governing the smoothness of each flexible component of the mean and the dispersion function. Just as in Section 4 we rewrote (2.2) as (4.4), we can rewrite (5.3) and (5.4) by defining $\boldsymbol{B}_{\mu}(\boldsymbol{x}_{d_\mu}) = \left[1, \boldsymbol{B}_{\mu}^{\mathrm{P}}(\boldsymbol{x}_{d_\mu}^{\mathrm{P}}), \boldsymbol{B}_{\mu}^{\mathrm{NP}}(\boldsymbol{x}_{d_\mu}^{\mathrm{NP}})\right]^T$ and $\boldsymbol{B}_{\gamma}(\boldsymbol{x}_{d_\gamma}) = \left[1, \boldsymbol{B}_{\gamma}^{\mathrm{P}}(\boldsymbol{x}_{d_\gamma}^{\mathrm{P}}), \boldsymbol{B}_{\gamma}^{\mathrm{NP}}(\boldsymbol{x}_{d_\gamma}^{\mathrm{NP}})\right]^T$ so that we have

$$g(\mu(\boldsymbol{x}_{d_\mu})) = \eta(\boldsymbol{x}_{d_\mu}) = \alpha_0 + \boldsymbol{B}_{\mu}^{\mathrm{P}}(\boldsymbol{x}_{d_\mu}^{\mathrm{P}})\boldsymbol{\alpha}_{\mu}^{\mathrm{P}} + \boldsymbol{B}_{\mu}^{\mathrm{NP}}(\boldsymbol{x}_{d_\mu}^{\mathrm{NP}})\boldsymbol{\alpha}_{\mu}^{\mathrm{NP}} = \boldsymbol{B}_{\mu}^{T}(\boldsymbol{x}_{d_\mu})\boldsymbol{\alpha}_{\mu}\ , \tag{5.5}$$

and

$$h^{-1}(\gamma(\boldsymbol{x}_{d_\gamma})) = \xi(\boldsymbol{x}_{d_\gamma}) = \alpha_{\gamma 0} + \boldsymbol{B}_{\gamma}^{\mathrm{P}}(\boldsymbol{x}_{d_\gamma}^{\mathrm{P}})\boldsymbol{\alpha}_{\gamma}^{\mathrm{P}} + \boldsymbol{B}_{\gamma}^{\mathrm{NP}}(\boldsymbol{x}_{d_\gamma}^{\mathrm{NP}})\boldsymbol{\alpha}_{\gamma}^{\mathrm{NP}} = \boldsymbol{B}_{\gamma}^{T}(\boldsymbol{x}_{d_\gamma})\boldsymbol{\alpha}_{\gamma}\ , \tag{5.6}$$

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

For a given sample of n observations $(\boldsymbol{x}, \boldsymbol{y})$, an i.i.d. realisation from $(\boldsymbol{X}_d, Y)$, we can extract from $\boldsymbol{x}$ the $\boldsymbol{x}_{\mu}$ vector in which we consider only the observed values of the $\boldsymbol{X}_{d_\mu}$ covariates, and the $\boldsymbol{x}_{\gamma}$ vector in which we consider only the observed values of the $\boldsymbol{X}_{d_\gamma}$ covariates. We then build the 'design' matrices $\boldsymbol{B}_{\mu} = [\mathbf{1}_n\ \boldsymbol{B}_{\mu}^{\mathrm{P}}\ \boldsymbol{B}_{\mu}^{\mathrm{NP}}]$ and $\boldsymbol{B}_{\gamma} = [\mathbf{1}_n\ \boldsymbol{B}_{\gamma}^{\mathrm{P}}\ \boldsymbol{B}_{\gamma}^{\mathrm{NP}}]$ similarly as in Section 4. Also, for given $\boldsymbol{\lambda}^{\mu}$ and $\boldsymbol{\lambda}^{\gamma}$, we can build the two penalty matrices $\boldsymbol{P}_{\mu} = \text{blockdiag}\ (0, \mathbf{0}_{K_{\mathrm{P}_\mu}}, \lambda_1^{\mu}\boldsymbol{D}_{m_1}^T\boldsymbol{D}_{m_1}, \ldots, \lambda_{d_{\mathrm{NP}_\mu}}^{\mu}\boldsymbol{D}_{m_{d_{\mathrm{NP}_\mu}}}^T\boldsymbol{D}_{m_{d_{\mathrm{NP}_\mu}}})$ and $\boldsymbol{P}_{\gamma} = \text{blockdiag}\ (0, \mathbf{0}_{K_{\mathrm{P}_\gamma}}, \lambda_1^{\gamma}\boldsymbol{D}_{\ell_1}^T\boldsymbol{D}_{\ell_1}, \ldots, \lambda_{d_{\mathrm{NP}_\gamma}}^{\gamma}\boldsymbol{D}_{\ell_{d_{\mathrm{NP}_\gamma}}}^T\boldsymbol{D}_{\ell_{d_{\mathrm{NP}_\gamma}}})$. Estimates of $\boldsymbol{\alpha}_{\mu}$ and $\boldsymbol{\alpha}_{\gamma}$ for given smoothing parameter vectors $\boldsymbol{\lambda}^{\mu}$ and $\boldsymbol{\lambda}^{\gamma}$, are obtained by maximizing the penalized log-likelihood:

$$\begin{aligned} l(\boldsymbol{\alpha}_{\mu}, \boldsymbol{\alpha}_{\gamma}; \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}^{\mu}, \boldsymbol{\lambda}^{\gamma}, \phi) = &-\frac{1}{2}\mathbf{1}_n^T\left\{\log(h(\boldsymbol{B}_{\gamma}\boldsymbol{\alpha}_{\gamma})) + \frac{1}{h(\boldsymbol{B}_{\gamma}\boldsymbol{\alpha}_{\gamma})}d(\boldsymbol{y}, \boldsymbol{B}_{\mu}\boldsymbol{\alpha}_{\mu})\right\} \\ &- \frac{1}{2}\boldsymbol{\alpha}_{\mu}^T\boldsymbol{P}_{\mu}\boldsymbol{\alpha}_{\mu} - \frac{1}{2}\boldsymbol{\alpha}_{\gamma}^T\boldsymbol{P}_{\gamma}\boldsymbol{\alpha}_{\gamma}. \end{aligned} \tag{5.7}$$

Maximization of (5.7) is done via a two-steps iterative procedure: first we 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. For starting the iterative procedures one needs some starting values denoted by $\boldsymbol{\alpha}_{\mu}^{(0)}$ and $\boldsymbol{\alpha}_{\gamma}^{(0)}$. In our implementation we take constant initial values such that $\hat{\mu}^{(0)} = n^{-1}\sum_{i=1}^{n} y_i = \overline{y}$ and $\hat{\gamma}^{(0)} = \phi_n$, where ϕ_n is the constant value we would expect in the one-parameter exponential family.

Once initial values are chosen the two steps procedure is started and alternates between the estimation of $\boldsymbol{\alpha}_\mu$ (Step (a)) and the estimation of $\boldsymbol{\alpha}_\gamma$ (Step (b)), until both the mean and the dispersion estimates converge. Denote the starting values at the ith iteration step by $\hat{\mu}^{(i-1)}(\boldsymbol{x}_\mu)$ and $\hat{\gamma}^{(i-1)}(\boldsymbol{x}_\gamma)$. Each ith iteration then alternates between the following steps:

- STEP (a): estimation of $\boldsymbol{\alpha}_\mu$.

In order to obtain a maximum penalized log-likelihood estimation of $\boldsymbol{\alpha}_\mu$ we rewrite (5.7) as a function of $\boldsymbol{\alpha}_\mu$ only, taking $\gamma(\boldsymbol{x}_\gamma) = \hat{\gamma}^{(i-1)}(\boldsymbol{x}_\gamma)$. After some algebra (see Gijbels *et al.* (2010)) it is found that the updating rule for $\boldsymbol{\alpha}_\mu$ is:

$$\boldsymbol{\alpha}_\mu = (\boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu \boldsymbol{B}_\mu + \boldsymbol{P}_\mu)^{-1} \boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu \tilde{\boldsymbol{z}_\mu} \,, \tag{5.8}$$

with $\tilde{\boldsymbol{z}}_\mu = \boldsymbol{B}_\mu \tilde{\boldsymbol{\alpha}}_\mu + (\boldsymbol{y} - b'(\boldsymbol{B}_\mu \tilde{\boldsymbol{\alpha}}_\mu))/b''(\boldsymbol{B}_\mu \tilde{\boldsymbol{\alpha}}_\mu)$ the vector of working variables as in (4.3) and $\tilde{\boldsymbol{W}}_\mu$ the current diagonal matrix $\tilde{\boldsymbol{W}}_\mu = \text{diag}\left(\frac{b''(\boldsymbol{B}_\mu \tilde{\boldsymbol{\alpha}}_\mu)}{\phi\gamma(\boldsymbol{x}_\gamma)}\right)$ in which the (estimated) gamma values appear in the denominator.

- STEP (b): estimation of $\boldsymbol{\alpha}_\gamma$.

Similarly to Step (a), maximum penalized log-likelihood estimate for $\boldsymbol{\alpha}_\gamma$ are obtained by rewriting (5.7) as a function of $\boldsymbol{\alpha}_\gamma$ only, taking $\theta(\boldsymbol{x}_\mu) = \hat{\theta}^{(i)}(\boldsymbol{x}_\mu)$. The updating rule for $\boldsymbol{\alpha}_\gamma$ can be found to be (Gijbels *et al.* (2010))

$$\boldsymbol{\alpha}_\gamma = (\boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma \boldsymbol{B}_\gamma + \boldsymbol{P}_\gamma)^{-1} \boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma \tilde{\boldsymbol{z}}_\gamma, \tag{5.9}$$

with $\tilde{\boldsymbol{z}}_\gamma$ the working variable vector

$$\tilde{\boldsymbol{z}}_\gamma = \boldsymbol{B}_\gamma \tilde{\boldsymbol{\alpha}}_\gamma + (d(\boldsymbol{y}, \theta(\boldsymbol{x}_\mu)) - h(\boldsymbol{B}_\gamma \tilde{\boldsymbol{\alpha}}_\gamma)) \frac{1}{h'(\boldsymbol{B}_\gamma \tilde{\boldsymbol{\alpha}}_\gamma)} \tag{5.10}$$

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

$$\tilde{\boldsymbol{W}}_\gamma = \frac{1}{2} \text{diag} \left(\frac{h'(\boldsymbol{B}\gamma \tilde{\boldsymbol{\alpha}}_\gamma)}{h(\boldsymbol{B}_\gamma \tilde{\boldsymbol{\alpha}}_\gamma)} \right)^2 . \tag{5.11}$$

At the final convergence we have that $\hat{\eta}(\boldsymbol{x}_\mu) = \boldsymbol{B}_\mu \hat{\boldsymbol{\alpha}}_\mu = \boldsymbol{H}_\mu \hat{\boldsymbol{z}}_\mu$ and $\hat{\xi}(\boldsymbol{x}_\gamma) = \boldsymbol{B}_\gamma \hat{\boldsymbol{\alpha}}_\gamma = \boldsymbol{H}_\gamma \hat{\boldsymbol{z}}_\gamma$, with $\boldsymbol{H}_\mu = \boldsymbol{B}_\mu (\boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu \boldsymbol{B}_\mu + \boldsymbol{P}_\mu)^{-1} \boldsymbol{B}_\mu^T \tilde{\boldsymbol{W}}_\mu$ and $\boldsymbol{H}_\gamma = \boldsymbol{B}_\gamma (\boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma \boldsymbol{B}_\gamma + \boldsymbol{P}_\gamma)^{-1} \boldsymbol{B}_\gamma^T \tilde{\boldsymbol{W}}_\gamma$ the hat matrices respectively of the mean and the dispersion estimation.

The algorithm has been presented until now taking $\boldsymbol{\lambda}^\mu$ and $\boldsymbol{\lambda}^\gamma$ to be known and fixed. Nevertheless, different choices of $\boldsymbol{\lambda}^\mu$ and $\boldsymbol{\lambda}^\gamma$ will lead to different estimates and it would

be desirable to be able to choose optimally the smoothing parameters values. In Section 5.3 we discuss a data-driven choice for these parameters.

We now illustrate the above procedure on the ozone data example. The need for estimating the variance (dispersion) is clear from Figure 4.1. For example the variability of the ozone concentration is much higher for higher inversion base temperature values. Our new point of interest is then to estimate the variance function as a function of the covariates. We are modelling the ozone data assuming that the ozone concentration values (Y) are normally distributed with mean μ and variance γ ($Y \sim \mathrm{N}(\mu, \gamma)$), and we assume that both μ and γ vary as a function of the covariates: $Y|\boldsymbol{X}_d = \boldsymbol{x}_d \sim \mathrm{N}(\mu(\boldsymbol{x}_d), \gamma(\boldsymbol{x}_d))$. As in (4.1) the model for the mean will be:

$$g(\mu(x_1, x_2, x_3)) = \eta(\boldsymbol{x}_d) = \alpha_{\mu 0} + \eta_1(x_1) + \eta_2(x_2) + \eta_3(x_3) \ , \tag{5.12}$$

with $g(\cdot)$ the identity function, while for modelling $\gamma(\boldsymbol{x}_d)$ we take

$$h^{-1}(\gamma(x_1, x_2, x_3)) = \xi(\boldsymbol{x}_d) = \alpha_{\gamma 0} + \xi_1(x_1) + \xi_2(x_2) + \xi_3(x_3) \tag{5.13}$$

where $h^{-1}(\cdot)$ is a link function such that the estimated $\gamma(x_1, x_2, x_3)$ is positive. In this and the other examples we took $h^{-1}(\gamma(x_1, x_2, x_3)) = \log(\gamma(x_1, x_2, x_3))$. As for estimation of the mean function, we take $\xi_1(x_1)$ to be of a quadratic form, i.e. $\xi_1(x_1) = x_1\alpha_{11} + x_1^2\alpha_{12}$, while both $\xi_2(x_2)$ and $\xi_3(x_3)$ are modelled flexibly via B-splines. Similarly to what we have done in Section 4 we take $\boldsymbol{B}_{\gamma 1}(x_1) = (x_1 \ x_1^2)^T$, while $\boldsymbol{B}_{\gamma 2}(x_2)$ and $\boldsymbol{B}_{\gamma 3}(x_3)$ are B-splines bases of dimension $K_{\gamma 2}$ and $K_{\gamma 3}$. We then take $K_{\gamma 2}$ and $K_{\gamma 3}$ to be large and we avoid overfitting for $\xi_2(x_2)$ and $\xi_3(x_3)$ by adding difference order penalties of order ℓ_1 and ℓ_2 and introducing smoothing parameters λ_1^γ and λ_2^γ. Taking $\boldsymbol{B}_\gamma(\boldsymbol{x}_d) = [\mathbf{1}, \ \boldsymbol{B}_{\gamma 1}^T(x_1), \ \boldsymbol{B}_2^T(\gamma x_2), \ \boldsymbol{B}_{\gamma 3}^T(x_3)]^T$ we rewrite (5.13) as:

$$h^{-1}(\gamma(x_1, x_2, x_3)) = \xi(\boldsymbol{x}_d) = \alpha_{\gamma 0} + \xi_1(x_1) + \xi_2(x_2) + \xi_3(x_3) = \boldsymbol{B}_\gamma^T(\boldsymbol{x}_d)\boldsymbol{\alpha}_\gamma \ , \tag{5.14}$$

with $\boldsymbol{\alpha}_\gamma$ the vector of parameters which needs to be estimated.

In Figure 5.1 we see the estimated mean and variance function components for the ozone data. The points plotted in the lower panels of the figure are $\log d(\boldsymbol{y}, \hat{\theta}(\boldsymbol{x}_d))$. The smoothing parameters value are $\boldsymbol{\lambda}_\mu = (6, 20.43)$ and $\boldsymbol{\lambda}_\gamma = (9216.37, 18000)$. These values were selected via GCV (see Section 5.3). The quadratic component $\xi_1(x_1)$ seems to have the strongest effect on the variance estimation. This partially reflects the fact that the first component has also a strong effect on the mean estimation. In Figure 5.1 we also

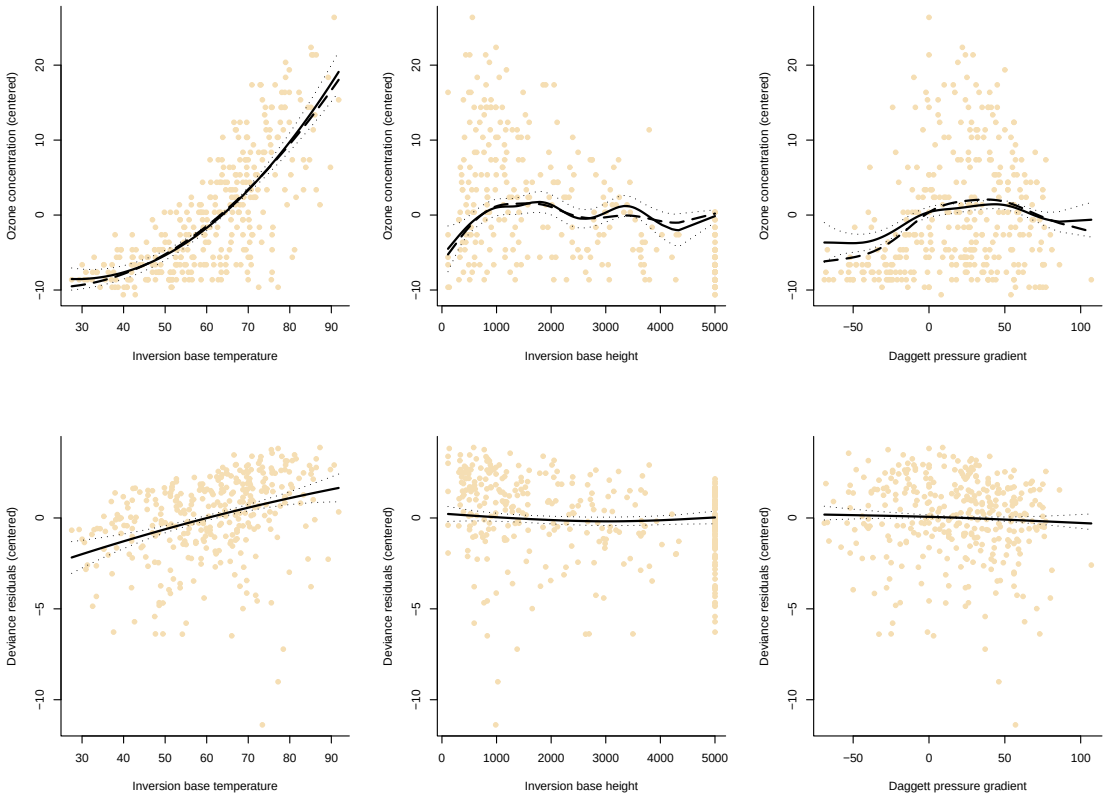

Figure 5.1: *The ozone data: mean and variance estimation. Top panels: (centered) data together with the estimated mean components η_1, η_2 and η_3. Lower panels: (centered) residuals are plotted with the ξ_1, ξ_2 and ξ_3 estimates. 95% confidence bands for each component are also drawn (dotted lines). The dashed lines in the above panels correspond to the estimation of the mean components when taking the variance to be constant.*

plot (as dashed curves) the mean component estimates we obtain when considering the variance to be constant with smoothing parameters $\boldsymbol{\lambda}_\mu = (6, 7.51)$. We see that estimating the variance also has an impact on the mean estimation, although the general contribution of each component to the final estimate does not change drastically. Theconfidence bands (dotted lines) displayed in Figure 5.1 are calculated following Wood (2006b). See also Gijbels *et al.* (2010) for details in a univariate context.

5.3 Choosing the smoothing parameters

The smoothing parameters $\boldsymbol{\lambda}^\mu$ and $\boldsymbol{\lambda}^\gamma$ can have a strong impact on the final estimates. It is therefore desirable to be able to choose these parameters in a somehow optimal way based on the observed sample $(\boldsymbol{x}, \boldsymbol{y})$. In the context of mean estimation different techniques and

methods have been proposed to choose the smoothing parameters optimally: see Eilers and Marx (1996), Gu and Xiang (2001), Wood (2006a) or Wood (2008) for some discussions. We propose to use the Generalized Cross Validation (GCV) technique for choosing the smoothing parameter for the mean function. However, the usual GCV score needs to be modified to account for the presence of the extra dispersion in the DEF. For a certain set of $\boldsymbol{\lambda}^{\mu}$ values we find a specific hat matrix $\boldsymbol{H}_{\mu}(\boldsymbol{\lambda}^{\mu})$, from which we can estimate the mean values via $\hat{\theta}_{\boldsymbol{\lambda}^{\mu}}(\boldsymbol{x}_{\mu}) = \boldsymbol{H}_{\mu}(\boldsymbol{\lambda}^{\mu})\boldsymbol{z}_{\mu}$. Also we take $\mathrm{df}(\boldsymbol{\lambda}^{\mu}) = \mathrm{tr}(\boldsymbol{H}_{\mu}(\boldsymbol{\lambda}^{\mu}))$ to be the equivalent degrees of freedom for the fit. We then select $\boldsymbol{\lambda}^{\mu}$ by minimizing:

$$\mathrm{GCV}(\boldsymbol{\lambda}^{\mu}) = \frac{n\ \mathbf{1}_n^T \left[d(\boldsymbol{y}, \hat{\theta}_{\boldsymbol{\lambda}^{\mu}}(\boldsymbol{x}_{\mu}))/\gamma(\boldsymbol{x}_{\gamma}) \right]}{(n - \mathrm{df}(\boldsymbol{\lambda}^{\mu}))^2}, \tag{5.15}$$

This quantity $\mathrm{GCV}(\boldsymbol{\lambda}^{\mu})$ is a commonly used criterion to choose the smoothing parameters when estimating a mean function. Less work is present in the literature about the choice of the smoothing parameter for the dispersion function estimation, and we mimic here the ideas behind the construction of $\mathrm{GCV}(\boldsymbol{\lambda}^{\mu})$. For a certain set of smoothing parameters $\boldsymbol{\lambda}^{\gamma}$ we take $\hat{\gamma}_{\boldsymbol{\lambda}^{\gamma}}(\boldsymbol{x}_{\gamma}) = h(\boldsymbol{H}_{\gamma}(\boldsymbol{\lambda}^{\gamma})\boldsymbol{z}_{\gamma})$ to be the estimated dispersion values and $\mathrm{df}(\boldsymbol{\lambda}^{\gamma}) = \mathrm{tr}(\boldsymbol{H}_{\gamma}(\boldsymbol{\lambda}^{\gamma}))$ to be the approximation for the effective degrees of freedom of the fit. Also, we need to define the deviance residuals for the variance estimation $d_{\gamma}(\gamma_S, \gamma) = 2[\log(f_Y(\gamma_S; \theta, y)) - \log(f_Y(\gamma; \theta, y))]$. Since $\gamma_S = d(y, \theta)$ we can write:

$$d_{\gamma}(d(y,\theta), \gamma) = \left\{ \log \frac{\gamma}{d(y,\theta)} + \frac{d(y,\theta)}{\gamma} - 1 \right\} .$$

An optimal choice of $\boldsymbol{\lambda}^{\gamma}$ is then given by minimizing:

$$\mathrm{GCV}(\boldsymbol{\lambda}^{\gamma}) = \frac{n\ \mathbf{1}_n^T d_{\gamma}(d(\boldsymbol{y}, \theta(\boldsymbol{x}_{\mu})), \hat{\gamma}_{\boldsymbol{\lambda}^{\gamma}}(\boldsymbol{x}_{\gamma}))}{(n - \mathrm{df}(\boldsymbol{\lambda}^{\gamma}))^2} . \tag{5.16}$$

6 The Italian abortion data

6.1 An overview

Induced abortion in Italy was made legal in 1978 with a very debated law. In 1981 a national referendum rejected the repeal of the law with a large majority (80%), but abortion is still a controversial topic in the country. With the legalization of abortion, the ISTAT provided a form which must be filled in for each carried out induced abortion. These forms are collected by the regions and transfered to the ISTAT and the Ministry of

Health. Thanks to these data we can try to have a look at how the phenomenon evolved, and what was the situation in 2001.

The first years after the legalization of induced abortion, the abortion rate showed a general increase, probably due to the fact that clandestine abortions were diminishing in favor of legal abortions, carried out in public or authorized private facilities. This is also indicated by the substantial decrease of natural miscarriages, which were actually partly the result of complications due to badly performed clandestine abortions. Moreover, some of the regions (typically the southern regions) took longer time to provide the needed services, so the data of the first years do not actually refer to the whole country, but only to some regions. With time women had the possibility of performing a legal abortion in all the national territory. In 1982 the highest national AR was registered and since then we can observe a decline, resulting from a better information on contraceptive methods. In the last years the abortion rate seems to be quite stable, although it should be noted that a greater contribution to the number of IA comes from immigrated women, who tend to have a higher abortivity rate than the women of Italian citizenship. Since the presence of immigrated women is very different in the Italian territory, this higher abortivity rate for immigrated women has different effects on the different Italian provinces.

The diversity of the Italian territory plays indeed a great role in the study of induced abortion in Italy. Typically the southern regions (with the notable exception of Puglia) took longer to create the necessary conditions to make legal abortion possible and still in these regions the accessability to the national health system and in particular to legal abortion services is lower. The Italian health system is in fact organized via the regions: the same services are provided throughout the country, but the practical organization of the health care is managed by the regions. Provinces are part of a specific region, on which they depend for the health care organization. The actual offer of health services is very different from region to region, with the southern regions typically showing less accessability to health care, including legal abortion services. This partially explains why the AR is generally lower in the southern regions, where the demand of legal abortion is not met by the regional health institutions, and women might either fall back upon illegal methods, or travel to other regions to perform a legal abortion or give up the idea of abortion and give birth to an unwanted child. A notable difference in this landscape is the case of Puglia, a southern region which, since the introduction of the 1978 law, made an effort to offer its citizens the possibility to undergo legal abortion and has, since the beginning, registered high abortion rates. Southern regions are also characterized by

generally lower socio-economic conditions and by more traditional behaviours, which also contribute to have lower AR.

A study of any social behaviour can therefore not ignore the differences among the Italian regions. In fact in the late 90s we can recognize two main patterns in the abortivity of Italian women (Boccuzzo (2000)). Married older women, possibly with children, might use abortion as a final method to control the family size once the desired size of the family has been reached and contraceptive methods have failed in avoiding the pregnancy. In fact among married women the abortivity rate increases with the number of already present children: abortion is seen as an extreme method to keep the family size stable. This behaviour is more present in the southern regions where the expected family size is larger than in the north: therefore this behaviour is typical for older women who have reached the desired family size. This explains why, among married women, the southern regions have the highest abortivity rates. Also among southern women between 35 and 39 married women exhibit much higher AR than unmarried women. The other pattern which can be identified is more frequent in the more modern central and northern regions: the abortion is an extemporaneous event through which unmarried and may be younger women avoid unexpected accidental pregnancy. In fact in these regions AR for unmarried women are higher than those of married women.

6.2 The data analysis

Among the rich ISTAT dataset we consider the following variables (next to the AR variable): the average age at first marriage for women, the Index of not finishing compulsory education for the female population between 15 and 52 and the percentage of families consisting of only one person (i.e. uniperson families). These variables might help in characterizing the differences in provinces related to modernity and conditions for women.

We analyzed the data on the Abortion Rate as coming from a double Poisson model (Efron 1986). As already seen in Section 1, the data show in fact a very strong underdispersion, with a variance that is much lower than what we would expect to find in Poisson data (the sample mean and variance are respectively 8.75 and 4.61 to give a rough indication). A plot of the data can be seen in Figure 6.1, where we can already notice some underdispersion. We can recognize in the data some differences in the Italian macro areas (South, Center and North) and it is striking to see how the Abortion rate for the provinces in Puglia is extremely high, much higher than the rest of the southern regions, with Bari

and Foggia having the highest AR of all Italy. This might be due to the good level of health care provided in the region, specially with regard to the possibility of having legal abortion, but still the data for this region seem to be somehow out of the path of the other data on Italian provinces. Therefore we will show the results of analysis done with and without the data concerning Puglia. Other notable observations are the ones of the provinces of Bolzano (North), Agrigento (South) and Sondrio (North): the three lowest registered AR of Italy. Sondrio and Bolzano are both wealthy area, and Bolzano in particular is a special status province, with very good social services: the low AR rates are possibly also due to the fact that families have the means to substain a child even if they have already reached the desired family size. The low AR of Agrigento instead is probably the result of the more traditional social environment and the poor health services of the area. In Figure 6.2 we see the estimates for the mean and variance components obtained

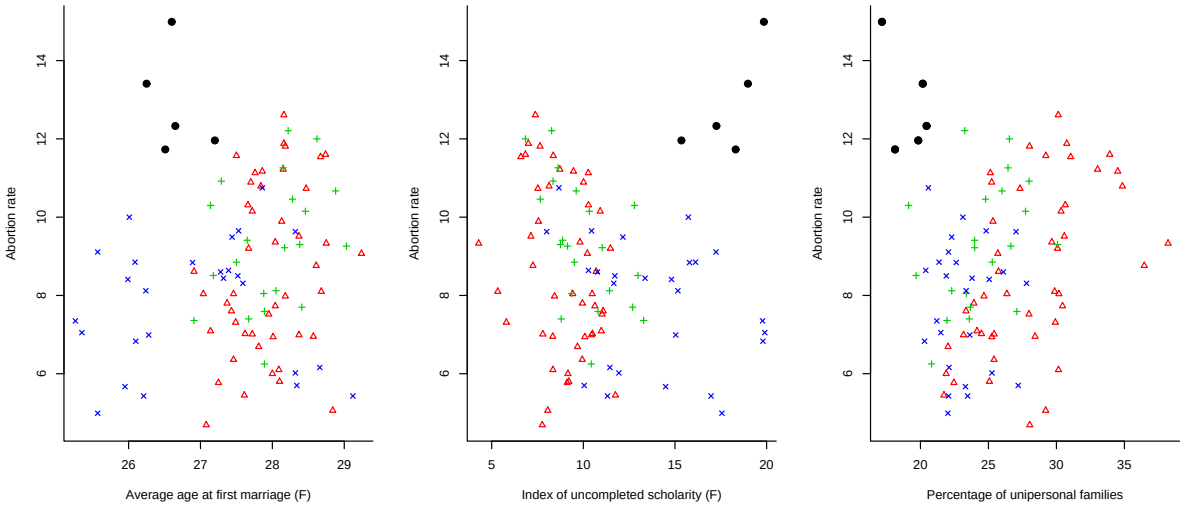

Figure 6.1: *The Italian abortion data (raw data). A* $\triangle$ *indicates observations from northern regions, a* $+$ *central regions and a* $\times$ *southern regions. The solid bullet* $\bullet$ *indicates observations from Puglia.*

by choosing optimal smoothing parameters as exposed in Section 5.3. In the top panels we plot the centered logarithm of the original data (the canonical link function for a Poisson is a logarithm: $g(\mu) = \log(\mu)$), together with the estimated mean components. Similarly, taking the link function for the dispersion to be a logarithm as well ($h^{-1}(\gamma) = \log(\gamma)$), we plot in the lower panel the centered logarithm of the deviance residuals and the estimated dispersion components. The plotted deviance residuals are obtained when estimating the model using all the available observations, including the data of Puglia.

In Figure 6.2 we can see that including or not the data of Puglia can have an effect

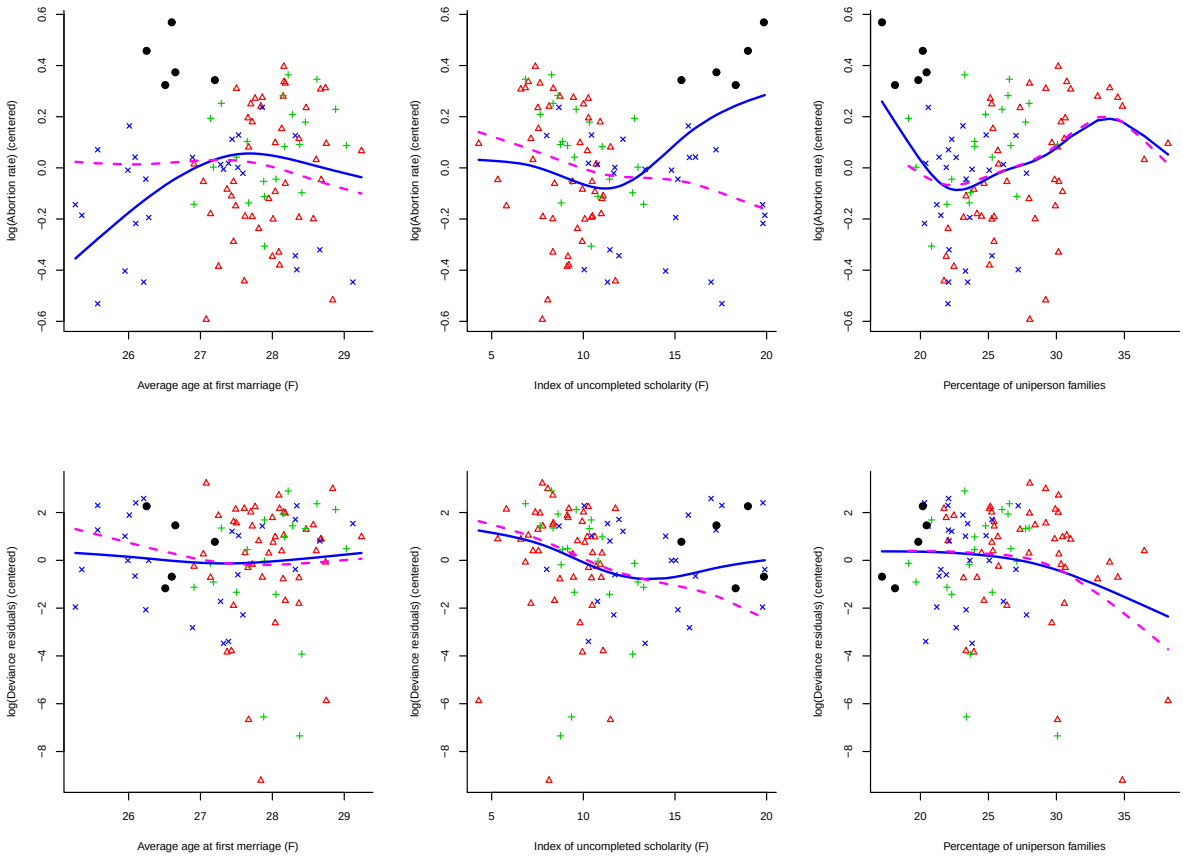

Figure 6.2: *The Italian abortion data. Mean and dispersion function estimates (top and bottom panels respectively). The solid lines correspond to the estimated mean and dispersion components when using all the available data, using respectively 11.56 and 7.41 degrees of freedom. The dashed lines correspond to the estimated mean and dispersion components when not including data of Puglia, using respectively 10.62 and 7.82 degrees of freedom.*

on the estimates both for the mean and the dispersion, possibly leading to different interpretations of the relationship between the covariates and the abortion rate. For example, in the most left upper panel we see that, after a fast increase at the beginning, the AR seems to slowly decrease in provinces where women marry after the age of 27. The relationship between the abortion rate and average age at first marriages for women is though very different depending whether we keep or not keep the data regarding Puglia in our analysis. In the second upper panel we see how the AR changes as function of the index of uncompleted scholarity for women: including the observation from Puglia changes the function from being a decreasing function to a more quadratic-shaped function. For sure

lower levels of uneducated women bring to higher abortion rate. Therefore in provinces in which women in the last years could study, where women became more independent, AR levels are higher. The third variable, the percentage of uniperson families is an indicator of how modern the social environment of the province is: higher percentages indicate provinces where the traditional family model is less present and people are less devoted to building a family. Indeed we see that after a initial decay (of a different size according to whether we keep or not the data from Puglia), the estimated function increases and in areas where more families are of the uniperson type we find higher AR. The extremely high percentages of uniperson families in the Province of Trieste and Savona though correspond to average values of AR, and this modifies the final estimate for the component. Again, these are northern quite wealthy provinces, that exhibit a modern behaviour but where women choose less for abortion compared to other provinces having high percentages of unipersonal families.

The differences in the mean estimates obtained when including or not the data from Puglia have an impact in the computation of the deviance residuals and also the estimates for the components of the dispersion will result in having different shapes. See the lower panels of Figure 6.2 and in particular the lower left panel, where we see how the dispersion estimate changes as a function of the average age at first marriages for women. In provinces where women on average marry after the age of 27 the dispersion has a mild increase. The relationship between the dispersion of the data and the Index of uncompleted scholarity is decreasing up to a certain part, and has different behaviour whether we include or not the data from Puglia. Finally the contribution of the last variable to the dispersion is constant in the beginning, but then the dispersion diminishes with higher percentages of uniperson families.

7 Simulation studies

To have a better understanding of how the proposed methods perform we present some simulation studies which investigate different aspects of the estimation procedure. We focus our attention to the Normal model with two covariates and simulate 1000 samples with two different underlying true models: Model A and Model B. We take the true underlying mean function for Model A (μ_A) and Model B (μ_B) to be:

$$\mu_A(x_1, x_2) = \eta_A(x_1, x_2) = \eta_0 + \eta_1(x_1) + \eta_2(x_2)$$
$$\mu_B(x_1, x_2) = \eta_B(x_1, x_2) = \eta_0 + \eta_2(x_1) + \eta_1(x_2)$$

where we take the identity to be the link function for the mean ($\eta(x_1, x_2) = \mu(x_1, x_2)$). The difference between the two mean structure is that we exchange the η_1 and η_2 components. The two components are shown as solid lines in the upper panels of Figure 7.3: the $\eta_1(\cdot)$ component is much smoother than $\eta_2(\cdot)$. Contrary to what we did for the mean we took the structure for the variance of the two models to be same. Taking the logarithm as the link function for the variance ($\xi(x_1, x_2) = \log(\gamma(x_1, x_2))$) we have

$$\xi_A(x_1, x_2) = \xi_B(x_1, x_2) = \log(\gamma(x_1, x_2)) = \xi_0 + \xi_1(x_1) + \xi_2(x_2)$$

with $\xi_1(\cdot)$ and $\xi_2(\cdot)$ to be the functions depicted as solid lines in the lower panels of Figure 7.3: again, one of the two function ($\xi_1(\cdot)$) is smoother than the other.

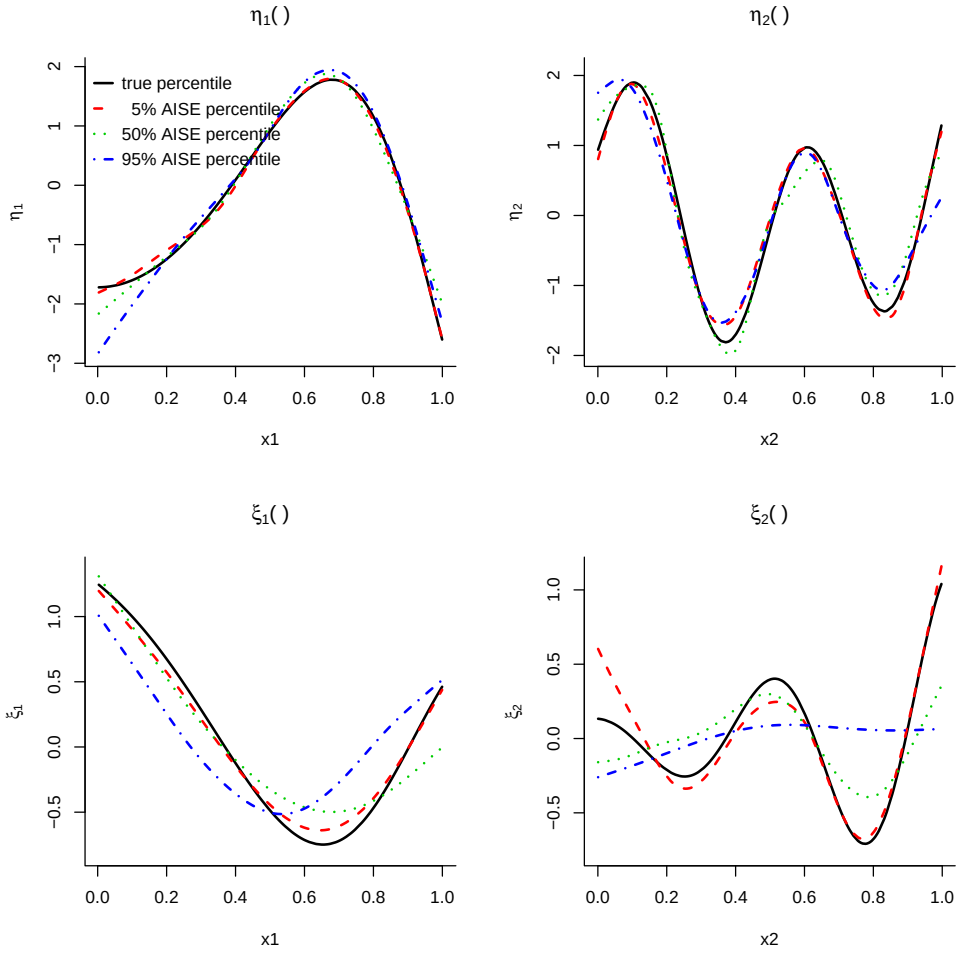

Figure 7.3: *Model A: true mean and variance components with representative estimates.*

In Figure 7.3, beside the true functions we also see some representative estimates obtained for Model A. For each simulated dataset in fact we estimated both the mean and the variance function components with the extended GAM approach presented in Section

5. In order to evaluate the quality of the obtained fits, we computed an approximate integrated squared error (AISE)

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

for each of the simulations (indexed by s). Herein x_{grid} is an appropriate grid of values, $f_{\mathrm{true}}(\cdot)$ and $\hat{f}(\cdot)$ are, respectively, the true and the estimated function (either the global mean and variance functions or the mean and variance components).

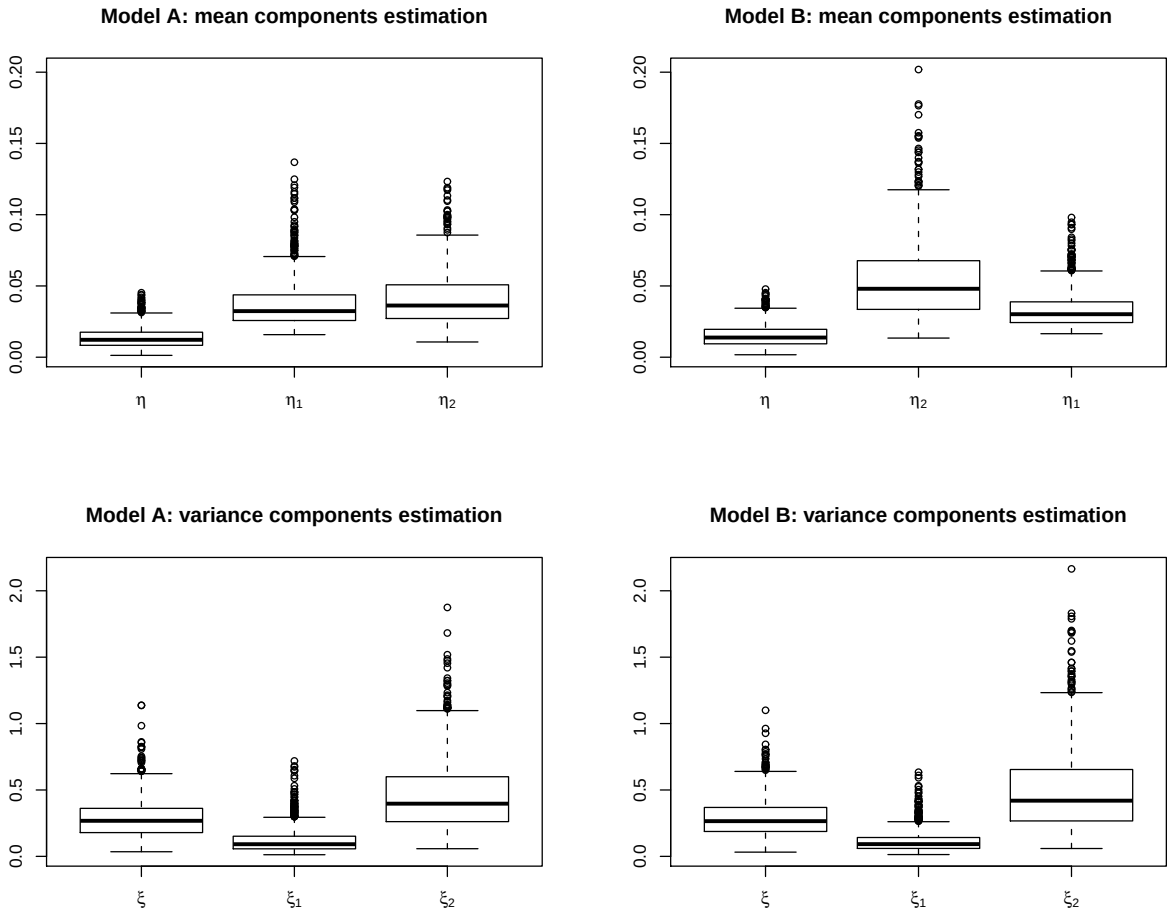

Figure 7.4: *Model A and Model B: boxplots of the AISE values for the mean and the variance function components in the two different models.*

In Figure 7.3 we present the true and the estimated components corresponding to the 5th, the 50th and the 95th quantile of the AISE values for Model A. The estimation procedure in general catches quite well the shape of both the mean and the variance

components, although the estimation of the last ones is, not surprisingly, slightly less precise. Results for Model B are not presented here but do not substantially differ from what can be seen in Figure 7.3.

The simulation framework of two different mean structures composed by swapping the two mean components was intended to investigate whether the smoothness of the mean component for one covariate has an effect on the estimation of the dispersion component for the same covariate, and vice-versa, whether the complexity of the dispersion component affect the estimation of the mean component. In Figure 7.4 we show boxplots of the AISE values for the mean and the variance components in both Model A and Model B. The performance of the estimation in the two models for the general components $\eta(x_1, x_2)$ and $\xi(x_1, x_2)$ is quite comparable, and the median value of the AISE for each component do not differ dramatically in the two model. What we can notice is that the AISE for $\eta_1(\cdot)$ shows larger variability in Model A than in Model B. Similarly, the variability of the AISE for $\eta_2(\cdot)$ is larger in Model B. It seems then, that the estimation of a mean component as a function of a covariate for which the dispersion component is the smooth function $\xi_1(\cdot)$, is somehow more variable. The performance of the estimation of the variance components for the two models is less different, although we observe that the AISE values for the variance components are generally higher than the one for the mean components.

Finally, to have a better understanding of what is the gain obtained by estimating the variance function via extended GAM, for each simulated dataset we also fitted a standard GAM in which the variance is estimated as a constant. In Figure 7.5 we see boxplots of the AISE values for the estimation of the mean and variance functions when the variance is estimated either as a function or as a constant. Results are presented for both Model A and Model B and we can see that estimating the variance has a positive impact on the quality of the mean estimation. Not surprisingly, estimating the variance function, when the variance is indeed changing as a function of the covariates, gives much lower AISE values than taking the variance to be constant.

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

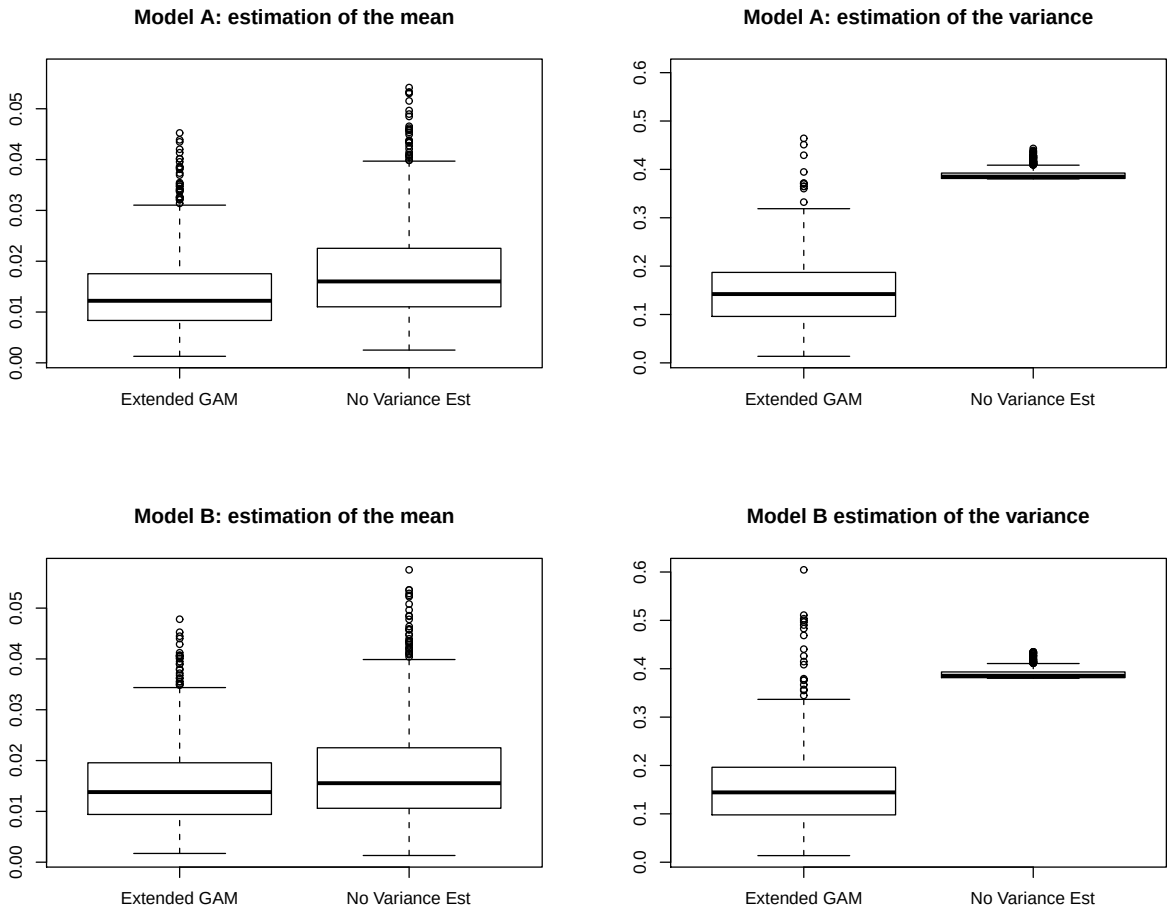

Figure 7.5: *Model A and Model B: boxplots of the AISE values for the mean and the variance function for the two different models.*

Boccuzzo, G. (editor) (2000). *L'abortività volontaria in Italia dalla legalizzazione ad oggi*. ISTAT (Informazioni 3/2000). (in Italian).

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

Buja, A., Hastie, T. J. and Tibshirani, R. (1989). Linear Smoothers and Additive Models. An overview of linear smoothing technology, including a proof of the convergence of the backfitting algorithm. *Annals of Statistics*, **17**, 453–555.

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall: New York.

Davidian, M. and Carroll, R.J. (1987). Variance function Estimation, *Journal of the American Statistical Association*, **82**, 1079–1091.

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

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

Efron, B. (1986). Double Exponential Families and their use in generalized linear regression. *Journal of the American Statistical Association*, **81**, 809–721.

Figà-Talamanca, I., Grandolfo, M. E. and Spinelli, A. (1986). Epidemiology of legal abortion in Italy. *International Journal of Epidemiology*, **15**, 343–351

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

Gu, C. and Xiang, D. (2001). Cross-validating non-Gaussian Data: generalized approximate cross-validation revisited *Journal of Computational and Graphical Statistics*, **10**, 581–591

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.

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

Lee, Y. and Nelder, J. A. (2006). Double hierarchical generalized linear models. *Applied Statistics*, **55**,139–185.

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*, 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, Ser. B*, **54**, 273–284.

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

Nott, D. (2006). Semiparametric estimation of mean and variance functions for non-Gaussian data. *Computational Statistics*, **21** 603–620.

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

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

Salvini Bettarini, S. and Schifini D'Andrea, S. (1996). Induced Abortion in Italy: levels, trends and characteristics, *Family Planning Perspectives*, **28**, 267–277

Spinelli, A. Forcella E., Di Rollo S. and Grandolfo M. (2006). L'interruzione volontaria di gravidanza tra le donne straniere in Italia, *Rapporti ISTISAN*, **17**.

Spinelli, A. and Grandolfo, M. E. (2001). Abortion in Italy. *Bollettino epidemiologico nazionale*, **14**, available at `http://www.epicentro.iss.it/ben/precedenti/aprile/1_en.htm`

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, Ser. B*, **70**, 495–518.

Yuan, M., Wahba, G. (2004). Doubly penalized likelihood estimator in heteroscedastic regression. *Statistics & Probability Letters*, **69**, 11–20.