On local estimating equations in additive multiparameter models

On local estimating equations in additive multiparameter models

Gerda Claeskens[1] [2] **and Marc Aerts**[3]

[1] *Department of Statistics, Eindhoven University of Technology, Den Dolech 2, P.O. Box 513, NL-5600 MB Eindhoven, The Netherlands*

[3] *Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium*

Abstract

Estimating all parameters in a multiparameter response model as smooth functions of an explanatory variable is very similar to estimating the different components of an additive model for the response mean. It is shown that, in a general estimating framework, local polynomial backfitting estimators in an additive one-parameter model do not work optimally. For a multiparameter model, however, a backfitting algorithm can be defined that leads to local polynomial estimators that do have optimal properties.

Keywords: Additive models, Backfitting, Estimating equations, Local polynomial estimators, Multiparameter models.

1 Introduction

In lots of statistical problems it is desirable to obtain nonparametric estimators of two or more curves at the same time. Here, we consider the general regression setting in which unknown parameters of an m dimensional response vector $\boldsymbol{Y}$ are modeled as a function of a D dimensional vector of covariates $\boldsymbol{X}$. We also focus on local polynomial estimation as smoothing technique, combined with the backfitting scheme as fitting algorithm. In this

[1]Corresponding author. Tel.:+31-40-2475583; fax:+31-40-2465995; e-mail: G.A.M.Claeskens@tue.nl

[2]Research supported by grant "Research Assistant of the Fund for Scientific Research Flanders – Belgium (F.W.O.)" at Center for Statistics, Limburgs Universitair Centrum.

context, simultaneous estimation of several curves can essentially occur in two different situations.

First, these curves can correspond to the different components of an additive model for one single "natural" parameter (typically the mean). We refer to this case as CASE I. This type of multicurve estimation has been studied extensively in the context of "classical" regression (response = mean + error, $m = 1$) by Hastie and Tibshirani (1990). They formulate the backfitting algorithm for general linear smoothers and show how it can be generalized to one-parameter likelihood models. The specific application of local polynomial smoothing to "classical" additive models is studied in detail by Opsomer and Ruppert (1997). Its generalization to general (quasi-)likelihood is one of our objectives.

Suppose next that there is only one single covariate of interest ($D = 1$). A completely different situation in which more curves have to be fit is when using a multiparameter response model (CASE II). Examples are the mean and variance function in a Gaussian regression model, or the probability of success and the correlation in a beta-binomial model for correlated binary data. Local polynomial smoothing in this case has been studied by Aerts and Claeskens (1997) and Carroll, Ruppert and Welsh (1998), but they only study in detail the case where one single bandwidth parameter is used for all curves. This is clearly not flexible enough, since one might expect a different degree of smoothness for each parameter. Therefore, each component should be estimated using a different bandwidth parameter. A call for such an estimation scheme is also expressed in Davison and Ramesh (1998) and Carroll, Ruppert and Welsh (1998). An interesting question here is whether local polynomial backfitting estimators can be used to achieve this goal and how optimal their properties are.

Finally, there is the general case where each curve in a multiparameter model may be a function of more than one covariate (CASE III). More explicit examples of all situations are given in the next section.

A desirable property is that each curve can be estimated in the most optimal, efficient way, that is, the estimator of one curve should have the same asymptotic properties as if all other components were known and the estimation scheme is essentially one-dimensional.

In Section 3 we obtain that, for case I, local polynomial backfitting estimators do not satisfy this optimality property, even not for independent covariates. An important result

in Section 4 is that a system of backfitting equations can be defined for the multiparameter case II, which lead to estimators with optimal statistical properties.

All results are formulated in a general estimating equations framework containing least squares, quasi-likelihood, etc. as special cases.

2 Examples

In this section we give some examples of parametric response models. Suppose we want to estimate some parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_\kappa)^T$, based on a sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, and suppose we use the following system of equations

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{Y}_i; \boldsymbol{\theta}) = 0.$$

When estimating $\boldsymbol{\theta}$ as a smooth function of $\boldsymbol{X}$ by local polynomials, each score contribution $\boldsymbol{\psi}(\boldsymbol{Y}_i; \boldsymbol{\theta})$ has to be weighted by a kernel function in combination with the backfitting algorithm.

Example 1: Classical regression

Let $m = 1, \kappa = 1$ and $\theta = E(Y_1)$. Then the least-squares estimating equations are $\psi(Y_i; \theta) = (Y_i - \theta)$. When there are covariates $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{Di})$ associated with Y_i, an additive model assumes that $\theta(X_{1i}, \ldots, X_{Di}) = \theta_1(X_{1i}) + \ldots + \theta_D(X_{Di})$. This is a case I example and most literature focuses on this particular setting.

Next, assume that Y_i is normally distributed and that, next to the mean $\theta_1 = E(Y_1)$, also the variance $\theta_2 = \mathrm{Var}(Y_1)$ is a parameter of interest. When estimating $\boldsymbol{\theta} = (\theta_1, \theta_2)$ ($\kappa = 2$) as a function of one covariate X, one has two options: estimating both parameters separately by moment type estimators or simultaneously by maximizing the (local) likelihood function as a function of $\boldsymbol{\theta}$. Both solutions are asymptotically equivalent. For this case II example, $\psi_k(Y_i; \boldsymbol{\theta}) = \partial / \partial \theta_k \ln f(Y_i; \theta_1, \theta_2)$, $k = 1, 2$, where f is the normal density. If $\boldsymbol{\theta}$ is estimated by the (local) likelihood method as a function of more covariates, we get a case III example.

Example 2: Categorical data models

An extension of the first case I - example to one-parameter exponential family additive models (like the logistic or Poisson regression model) is discussed in Hastie and Tibshirani (1990). Often overdispersion is present, due to the clustered nature of the data (a

cluster might be a litter in a toxicological experiment or a household in a survey) and different methods should be applied. One possibility is to use local quasi-likelihood as in Fan, Heckman and Wand (1995). Here $\kappa = 1$ and $\psi(Y_i; \boldsymbol{\theta}) = \partial/\partial\theta \ln f(Y_i; \theta)$ is the appropriate (quasi)likelihood function. The use of local polynomial backfitting estimators as presented in Section 3 has, to our knowledge, not been studied before.

Another more sophisticated way to handle overdispersion is to model it directly by compound distributions like the negative-binomial (count data) or the beta-binomial (binary data), see, e.g., Morgan (1992). For clustered binary data there are several other multiparameter models like the Bahadur model (Bahadur 1961), the conditional model of Molenberghs and Ryan (1999) etc. Next to the success probability θ_1, these models all include one or more parameters $\theta_2, \theta_3, \ldots$ to describe the association between outcomes (so here $\kappa \geq 2$). Several examples where smoothing is involved are discussed in Claeskens (1999). For our case II, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_\kappa)$ are modeled as a function of one covariate ($D = 1$) and for case III as an additive function of several covariates ($D > 1$). Here, the score functions ψ_k are the derivatives of the log(quasi-)likelihood with respect to θ_k.

Example 3: Other complex data

There are still many other data settings and models leading to multiparameter models. For clustered data and longitudinal data one can also apply the generalized estimating equations (see, e.g., Liang and Zeger, 1986) or pseudo-likelihood (see, e.g., Geys, Molenberghs and Ryan, 1999). A step further is dealing with multivariate response data $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$, like, e.g., a multivariate generalized linear model where the mean of each response component is assumed to be additive in the covariates. More and more parameters have to be estimated and estimating equations get more complex. Finally, a complete different type of data are extremes, where location, scale and shape vary according to smooth functions over time (Davison and Ramesh, 1998).

Although it might get complicated to implement certain multiparameter models with local polynomial backfitting estimators, theory as presented in next sections covers all given examples. To keep presentation simple, we state all results without regularity conditions. These conditions are essentially a mixture of conditions as given by Opsomer and Ruppert (1997) and Aerts and Claeskens (1997).

3 Additive models

In this section we consider the one-parameter multiple covariate case (case I). The data we observe are $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ where $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{Di})$. In an additive model with $D \geq 1$ covariates, the unknown parameter of interest has the following form

$$\theta(x_1, \ldots, x_D) = \theta_1(x_1) + \ldots + \theta_D(x_D), \tag{1}$$

where, for example in a generalized linear model, $\theta(\boldsymbol{x}) = g(E[Y|\boldsymbol{X} = \boldsymbol{x}])$ for a known link function g (see, e.g., Hastie and Tibshirani, 1990).

If there is more than one covariate, that is, if $D > 1$, there is a problem of identifiability, which is typical for this kind of additive models. For a random design, one usually includes an intercept term α in the model and assumes that each of the expected values $E[\theta_d(X_d)]$ is zero. This is also the approach that we will take. The intercept term α is estimated by solving $\sum_{i=1}^{n} \psi(Y_i, \alpha) = 0$. Local polynomial estimators of degree p_k of the curves $\theta_k(\cdot)$ at $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, and of their derivatives up to order p_k, for $(k = 1, \ldots, D)$, can be obtained by solving the following set of kernel weighted estimating equations:

$$\boldsymbol{\psi}\{\boldsymbol{\beta}(\boldsymbol{X}_1)\} = 0, \ldots, \boldsymbol{\psi}\{\boldsymbol{\beta}(\boldsymbol{X}_n)\} = 0 \tag{2}$$

where $\boldsymbol{\psi}\{\boldsymbol{\beta}(\boldsymbol{x})\} =$

$$\begin{cases} \sum_{i=1}^{n} \psi\{Y_i; \alpha + \sum_{j=0}^{p_d} \beta_{dj}(x_d)(X_{di} - x_d)^j + \sum_{k \neq d} \beta_{k0}(X_{ki})\} K_{h_d}(X_{di} - x_d) \\ \quad \vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{for } d = 1, \ldots D) \\ \sum_{i=1}^{n} \psi\{Y_i; \alpha + \sum_{j=0}^{p_d} \beta_{dj}(x_d)(X_{di} - x_d)^j + \sum_{k \neq d} \beta_{k0}(X_{ki})\} (X_{di} - x_d)^{p_d} K_{h_d}(X_{di} - x_d) \ , \end{cases} \tag{3}$$

$K_h(\cdot) = K(\cdot/h)/h$ and $\boldsymbol{\beta}(\boldsymbol{x}) = (\boldsymbol{\beta}_{1x}^T, \ldots, \boldsymbol{\beta}_{Dx}^T)^T$ with $\boldsymbol{\beta}_{dx} = (\beta_{d0}(x_d), \ldots, \beta_{dp_d}(x_d))^T$.

First we need some more notation. The jth moment of the kernel function K is defined by $\nu_j(x, h_k) = \int_{\mathcal{R}_{x,h_k}} u^j K(u) du$, where the integration region is given by, $\mathcal{R}_{x,h_k} = \{t : (x + h_k t) \in \text{supp}(f_k)\} \cap \text{supp}(K)$, and f_k is the marginal probability density function of covariate X_k. This region will indicate the difference between interior and boundary points, the former are identified when $\mathcal{R}_{x,h_k} = \text{supp}(K)$. Let $R_0(x, h_k) = \int_{\mathcal{R}_{x,h_k}} K^2(u) du$. The notational dependence of ν and $\mathcal{R}$ on x and h_k will be omitted for interior points.

Solving the set of equations (2) can be done in practice via the iteratively reweighted backfitting algorithm. For likelihood models, Hastie and Tibshirani (1990, p. 149) moti-

vate the use of an iterative backfitting scheme, where the "smoother" matrices now also depend on the unknown coefficients via the matrix of partial derivatives of the estimating equations, as does the pseudo-response vector or adjusted dependent variable (see, e.g., Hastie and Tibshirani, 1990, p. 139), the latter defined for general estimating equations, not necessarily originating from likelihood equations. This dependence on unknowns requires the additional iterative local scoring procedure where these matrices are "updated" after which the backfitting step is repeated using these updated matrices.

A novel aspect of the next result is that the asymptotic properties of estimators are obtained outside the framework of the classical, homoscedastic regression model (see also Section 2). In order to obtain our results, we rely on the important piece of work as provided by Opsomer and Ruppert (1997). To get the asymptotic conditional bias and variance of the estimators $\hat{\beta}_{kj}(X_{ki})$ it suffices to concentrate on a first order approximation of the equations (2) *if* second partial derivatives of the estimating equations (e.g. third partial derivatives of the log likelihood) are uniformly bounded.

Assuming the sufficient conditions for existence of the estimators to hold, we focus attention to asymptotic bias and variance expressions of the estimators. In parametric estimating equations models, conditions implying existence, consistency and asymptotic normality of the estimators are formulated by Yuan and Jennrich (1998), and for local likelihood equations, see Aerts and Claeskens (1997).

For the bivariate case ($D = 2$) we obtain the following result for the estimators of the curves $\hat{\theta}_1(X_{1i}) = \hat{\beta}_{10}(X_{1i})$ and $\hat{\theta}_2(X_{2i}) = \hat{\beta}_{20}(X_{2i})$. Extensions to the estimators $\hat{\beta}_{dj}(X_{di})$ for $2 < d \leq D$ and $1 \leq j \leq p_d$ are straightforward.

Theorem 1 *For p_1 and p_2 both odd, under the appropriate regularity conditions,*

$$E[\hat{\theta}_1(X_{1i}) - \theta_1(X_{1i})|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] = \frac{h_1^{p_1+1}}{(p_1+1)!}\nu_{p_1+1}(X_{1i}, h_1)\theta_1^{(p_1+1)}(X_{1i})$$
$$+\frac{h_1^{p_1+1}}{(p_1+1)!}\nu_{p_1+1}C_1 - \frac{h_2^{p_2+1}}{(p_2+1)!}\nu_{p_2+1}C_2 + O_P(\frac{1}{\sqrt{n}}) + o_P(h_1^{p_1+1} + h_2^{p_2+1}),$$

a similar expression is obtained for $E[\hat{\theta}_2(X_{2i}) - \theta_2(X_{2i})|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]$.

$$Var(\hat{\theta}_1(X_{1i})|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \frac{1}{nh_1}R_0(K_{p_1}, X_{1i})f_1^{-1}(X_{1i})C_3 + o_P(\frac{1}{nh_1})$$

and similarly for $Var(\hat{\theta}_2(X_{2i})|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. The values C_1, C_2 and C_3 depend in a complicated way on the "Fisher" information $E[\frac{\partial\psi}{\partial\theta}(Y; \theta_1(X_{1i}) + \theta_2(X_{2i}))|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]$ and

the joint density function of the covariates. C_2 *depends also on* $\theta_2^{(p_2+1)}(X_{2i})$, C_1 *on* $\left(\theta_1^{(p_1+1)}(X_{11}), \ldots, \theta_1^{(p_1+1)}(X_{1n})\right)$, *and* C_3 *on* $E[\psi^2(Y; \theta_1(X_{1i}) + \theta_2(X_{2i}))|X_{1i}, X_{2i}]$.

The proof can be obtained along the same lines as the proof of Theorem 4.1 in Opsomer and Ruppert (1997), details are therefore omitted (see Claeskens 1999). All results obtained above reduce to those of the classical additive model under least squares equations. This theorem clearly demonstrates the undesirable property that the bias of the estimators depends on all curves in the model. An important difference with the classical additive models, though, is that in general there is no simplification when X_{1i} and X_{2i} are independent, because of the Fisher information matrix. There is no obvious way to improve these results, i.e. to adapt the estimation scheme to obtain a bias expression depending on one component function only.

A possible solution could be to define a multi-step procedure where for estimation of, say, $\theta_1(x)$, a too small bandwidth is used for estimation of the other components, such that their contribution to the bias will be, at least asymptotically, negligible. In the next steps, the same procedure is repeated for each other component of the additive model. It is yet unknown how this procedure will perform in practical situations.

This so-called "non oracle" efficiency of backfitting estimators has also been noticed by Linton and Nielsen (1995), who proposed an alternative estimation scheme, using integration approaches, properties of which are also studied by Fan, Härdle and Mammen (1998). This, however, is also not fully efficient in mean squared error sense. There are indications that in less general settings, an estimation method combining backfitting and marginal integration is fully efficient, see, e.g., Linton (1997, 1998).

The main conclusion is that, when using local polynomials, backfitting estimators do not work optimally for the purpose they are defined: estimating components of additive models. It should be stressed that this is not only true for additive models having structure (1), but for *any* additive structure, e.g., $\theta(x_1, \ldots, x_D) = \theta_1(x_1, \ldots, x_k) + \theta_2(x_{k+1}, \ldots, x_D)$. In the next section we will show that a similar set of estimating equations does "work" in case the curves are separated, i.e. in a multiparameter context.

4 Multiparameter models

We here deal with a somewhat different setting. While the estimation method for additive models in Section 3 is defined for one-parameter models, we now extend the scope of application to generalized regression models with more than one parameter, which are all modeled as a function of some covariate(s). All "parameter" curves will be estimated simultaneously.

To explain the method, we first restrict attention to the relatively simple case of multiparameter models where each "parameter" is a function of the same univariate covariate (case II). Later on, we extend this to multiparameter models where each parameter is an additive function of several covariates (case III). The last category contains the one-parameter additive models of Section 3 as a special case.

4.1 A univariate covariate

We define two possible sets of local polynomial estimating equations for multiparameter single covariate models. To each parameter θ_k $(k = 1, \ldots, \kappa)$ corresponds an estimating equation ψ_k, e.g. the partial derivative of a local log likelihood function with respect to this parameter. Since the parameters can be structurally very different, e.g., one describing location and the others scale, or shape, there is the need to allow each of the functions to have its own bandwidth parameter. This is the main contribution of this section.

Local polynomial estimators are obtained by solving the following set of $\sum_{j=1}^{\kappa}(p_j + 1)$ equations where for each estimating function ψ_k a smoothing parameter h_k is used. To avoid unnecessary complicated notation, from here on, we take $\kappa = 2$, the two-parameter case. For example, for $k = 1$

$$\begin{aligned}
&\sum_{i=1}^{n} \psi_1\{\boldsymbol{Y}_i; \sum_{j=0}^{p_1} \beta_{1j}(x)(X_i - x)^j, \beta_{20}(X_i)\} K_{h_1}(X_i - x) = 0 \\
&\qquad \vdots \\
&\sum_{i=1}^{n} \psi_1\{\boldsymbol{Y}_i; \sum_{j=0}^{p_1} \beta_{1j}(x)(X_i - x)^j, \beta_{20}(X_i)\} (X_i - x)^{p_1} K_{h_1}(X_i - x) = 0.
\end{aligned} \tag{4}$$

Similar equations are defined for $k = 2$, where now only the second curve is locally approximated by a polynomial of degree p_2. In order to obtain the estimators at the data values, we need to solve a total of $n \times (p_1 + p_2 + 2)$ equations, by choosing x to be one of

these values X_i, $i = 1, \ldots, n$.

Note that the structure of these equations is very different from that of equations (3) where there is only one global parameter $\theta(\cdot)$, and hence a one-dimensional estimating function $\psi(\cdot)$, while in the set of equations (4) there are κ global parameters $\theta_1(\cdot), \ldots, \theta_\kappa(\cdot)$ and also κ estimating functions $\psi_1(\cdot), \ldots, \psi_\kappa(\cdot)$. This also implies that the set of estimating equations (4) does not have identifiability problems, since global parameters are separated.

Some further notation, for $\boldsymbol{N}_p$ be the $(p+1) \times (p+1)$ matrix with (k, ℓ)th entry equal to $\nu_{k+\ell-2}$, the matrix $\boldsymbol{M}_{tp_d}(z)$ is obtained by replacing in $\boldsymbol{N}_{p_d}$ the $(t+1)$th $(t = 0, \ldots, p_d)$ column by $(1,\ z,\ \ldots\ , z^{p_d})^T$, and for $|\boldsymbol{N}_{p_d}| \neq 0$, define $K_{tp_d}(z) = K(z)|\boldsymbol{M}_{tp_d}(z)|/|\boldsymbol{N}_{p_d}|$. Also define $J_{rs}(\boldsymbol{\theta}(x)) = E[(\partial\psi_r/\partial\theta_s)(\boldsymbol{Y}; \boldsymbol{\theta}(X))|X = x]$ and $\boldsymbol{J}(\boldsymbol{\theta}(x)) = [J_{rs}(\boldsymbol{\theta}(x))]_{r,s}$, $\boldsymbol{K}(\boldsymbol{\theta}(x)) = [K_{rs}(\boldsymbol{\theta}(x))]_{r,s}$ where $K_{rs}(\boldsymbol{\theta}(x)) = E[\psi_r\{\boldsymbol{Y}; \boldsymbol{\theta}(X)\}\psi_s\{\boldsymbol{Y}; \boldsymbol{\theta}(X)\}|X = x]$.

Theorem 2 *Under the appropriate set of regularity conditions, for $p_1 = p_2 = p$ odd,*

$$E[\hat{\theta}_k(X_i) - \theta_k(X_i)|X_1, \ldots, X_n] = h_k^{p+1} \frac{\theta_k^{(p+1)}(X_i)}{(p+1)!} \int_{\mathcal{R}_{X_i,h_k}} z^{p+1} K_{0p}(z)dz + O_P(h_k^{p+2}),$$

and the conditional variance by

$$Var(\hat{\theta}_k(X_i)) = \frac{f_X^{-1}(X_i)}{nh_k} [\boldsymbol{J}^{-1}(\boldsymbol{\theta}(X_i))\boldsymbol{K}(\boldsymbol{\theta}(X_i))\boldsymbol{J}^{-1}(\boldsymbol{\theta}(X_i))]_{kk} \int_{\mathcal{R}_{X_i,h_k}} K_{0p}^2(z)dz + o_P(\frac{1}{nh_k}).$$

Proof. The simple formulae in the above theorem follow from the fact that, for each k, partial derivatives of the estimating equation ψ_k with respect to each of the intercept terms $\beta_{j0}(x_i)$ for $j \neq k$ and for all $i = 1, \ldots, n$ is $O[1/(nh_k)]$, and hence can be ignored asymptotically since this is of lower order than the leading bias terms. □

Remark 1. We here focus on the case of local polynomial approximations of the same degree, motivated by Aerts and Claeskens (1997) where it is shown that the leading term of the asymptotic bias is determined by the polynomial of the lowest degree. Extensions to properties of the estimators $\hat{\beta}_{kj}$ and to $\kappa > 2$ are straightforward.

The asymptotic bias and variance expressions in Theorem 2 have an important consequence on the selection of the optimal bandwidth. Since asymptotic mean squared error (AMSE) for the multiparameter model is simply a sum of the AMSE for each component, Theorem 2 implies that any of the bandwidth selectors for one-parameter models can be applied in this multiparameter setting. Alternatively, one can consider a multi-stage

method which assumes the bandwidths for all but one curve to be fixed, and proceeds by selecting this one bandwidth, after which the procedure is repeated for each other component.

There is another set of equations which might be used in case II, a very natural extension of the multiparameter estimating equations with $h_1 = h_2 = h$ (Aerts and Claeskens, 1997). For $k = 1, 2$,

$$\begin{aligned}
&\sum_{i=1}^{n} \psi_k\{\boldsymbol{Y}_i; \sum_{j=0}^{p_1} \beta_{1j}(x)(X_i - x)^j, \sum_{j=0}^{p_2} \beta_{2j}(x)(X_i - x)^j\} K_{h_k}(X_i - x) = 0 \\
&\qquad \vdots \\
&\sum_{i=1}^{n} \psi_k\{\boldsymbol{Y}_i; \sum_{j=0}^{p_1} \beta_{1j}(x)(X_i - x)^j, \sum_{j=0}^{p_2} \beta_{2j}(x)(X_i - x)^j\} (X_i - x)^{p_k} K_{h_k}(X_i - x) = 0
\end{aligned} \tag{5}$$

An advantage of this set of equations is its computational simplicity. A major disadvantage of these formulae is that, in general, for each choice of k the bias expression depends on the complete vector $\boldsymbol{\theta}(x) = (\theta_1(x), \theta_2(x))$, and on both bandwidths h_1 and h_2. For $p_1 = p_2 = p$ odd, the conditional bias of $\hat{\theta}_k(x)$ can be approximated by

$$\begin{aligned}
&E[\hat{\theta}_k(x) - \theta_k(x) | X_1, \ldots, X_n] \\
&= \int_{\mathcal{R}_x} z^{p+1} K_{0p}(z) dz \sum_{r=1}^{2} \sum_{\ell=1}^{2} h_\ell^{p+1} \sqrt{\frac{h_\ell}{h_k}} \frac{\theta_r^{(p+1)}(x)}{(p+1)!} J_{\ell r}(\boldsymbol{\theta}(x)) [\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))]_{k\ell} \\
&\quad + O(\sum_{r=1}^{2} h_r^{p+2} \sqrt{\frac{h_r}{h_k}}).
\end{aligned} \tag{6}$$

The occurrence of the functions $\theta_r(\cdot)$, $r \neq k$, can be explained by the set of local estimating equations (5). By construction, in the set of equations for estimating $\theta_k(x)$ and its derivatives up to order p_k, there is not only a "local polynomial" for this component, but also for each other component. In practical situations, one expects the estimators obtained by (5) to be subject to more bias than the estimators obtained by solving (4). In an extreme case of very different curvature, the influence of one curve on the other should be more pronounced in equations (5).

From (6) it is clearly seen that to avoid problems, one might want to take both bandwidths of the same order, such that the ratios h_r/h_k are $O(1)$ for all r and k. In this case, for $h_k = c_k n^\delta$ (for some δ depending on p), the selection of optimal constants c_1 and c_2 is difficult because both curves $\theta_1(\cdot)$ and $\theta_2(\cdot)$ appear in bias expression (6).

Also the conditional variance of the estimators depends on all parameter curves $\theta_k(\cdot)$

(not shown here). These results in fact discourage the use of different bandwidths for different components.

Remark 2. For some specific sets of estimating functions ψ_k, simplifications can occur. If the "Fisher" information matrix $\boldsymbol{J}(\boldsymbol{\theta}(x))$ is a diagonal matrix, i.e., all off-diagonal elements are zero, the above bias expression simplifies to:

$$E[\hat{\theta}_k(x) - \theta_k(x)|X_1, \ldots, X_n] = h_k^{p+1} \frac{\theta_k^{(p+1)}(x)}{(p+1)!} \int_{\mathcal{R}_x} z^{p+1} K_{0p}(z) dz + O(h_k^{p+2}).$$

One such example is the set of log-likelihood equations for a Gaussian regression model.

Remark 3. Results also simplify in multi-stage equations where, for example, estimating function ψ_k is a function of $\theta_1, \ldots, \theta_k$, but not of the other θ_ℓ's ($\ell > k$). This implies that $\boldsymbol{J}(\boldsymbol{\theta}(x))$ is a lower triangular matrix, i.e., $J_{rs}(\boldsymbol{\theta}(x)) = 0$ for all $r < s$. Because of this structure, only terms containing quotients h_s/h_r with $r < s$ will occur in the leading terms of the conditional bias expressions. More explicitly, for a two parameter model,

$$\begin{aligned} & E[\hat{\theta}_1(x) - \theta_1(x)|X_1, \ldots, X_n] \\ &= \left(h_1^{p+1} \frac{\theta_1^{(p+1)}(x)}{(p+1)!} + h_2^{p+1} \sqrt{\frac{h_2}{h_1}} \left\{ \frac{\theta_1^{(p+1)}(x)}{(p+1)!} J_{21}(\boldsymbol{\theta}(x)) + \frac{\theta_2^{(p+1)}(x)}{(p+1)!} J_{22}(\boldsymbol{\theta}(x)) \right\} \right. \\ &\quad \left. \times [\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))]_{12} \right) \int_{\mathcal{R}} z^{p+1} K_{0p}(z) dz + O(h_1^{p+2} + h_2^{p+2} \sqrt{\frac{h_2}{h_1}}), \end{aligned}$$

$$\begin{aligned} & E[\hat{\theta}_2(x) - \theta_2(x)|X_1, \ldots, X_n] \\ &= h_2^{p+1} \left(\frac{\theta_2^{(p+1)}(x)}{(p+1)!} + \frac{\theta_1^{(p+1)}(x)}{(p+1)!} J_{21}(\boldsymbol{\theta}(x)) [\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))]_{22} \right) \int_{\mathcal{R}} z^{p+1} K_{0p}(z) dz + O(h_2^{p+2}). \end{aligned}$$

The function $\theta_1(\cdot)$ appears in both estimating functions ψ_1 and ψ_2 but $\theta_2(\cdot)$ only in ψ_2, a fact which is clearly reflected in the conditional bias expression. Also here, it will be safe to take $h_2/h_1 = O(1)$.

An example of two-stage equations are the GEE2 equations (see, e.g., Zhao and Prentice, 1990) where the first estimating equation yields an estimator for the mean response and where the second estimating equation is used to obtain an estimator of the correlation structure of the multivariate response vector, given an estimator of the mean.

4.2 A multivariate covariate

From the previous discussion it should be clear that also for multiparameter additive models, problems are to be expected when the backfitting estimation scheme is used to estimate the additive components by local polynomials.

For simplicity of notation, assume that we have a two-parameter model $\kappa = 2$, where each parameter is an additive function of the covariates:

$$\theta_k(\boldsymbol{x}_k) = \theta_{k1}(x_{k1}) + \ldots + \theta_{kd_k}(x_{kD_k})$$

where $\boldsymbol{x}_k = (x_{k1}, \ldots, x_{kD_k}),\ k = 1, \ldots, \kappa$. Assume that the expectation of each of these component functions $\theta_{kd}(X_{kd})$ equals zero, and include an intercept term, such that, for the ith observation, the contribution to the global, unweighted estimating function for the kth parameter at the true parameter values is given by

$$\psi_k\{\boldsymbol{Y}_i; \alpha_1 + \theta_{11}(X_{11i}) + \ldots + \theta_{1D_1}(X_{1D_1i}), \alpha_2 + \theta_{21}(X_{21i}) + \ldots + \theta_{2D_2}(X_{2D_2i})\}.$$

A set of local estimating equations is now defined as:

$$\boldsymbol{\psi}_1(\boldsymbol{\beta}(\boldsymbol{X}_{11})) = \ldots = \boldsymbol{\psi}_1(\boldsymbol{\beta}(\boldsymbol{X}_{1n})) = \mathbf{0} = \boldsymbol{\psi}_2(\boldsymbol{\beta}(\boldsymbol{X}_{21})) = \ldots = \boldsymbol{\psi}_2(\boldsymbol{\beta}(\boldsymbol{X}_{2n}))$$

where, $\boldsymbol{\psi}_1(\boldsymbol{x}_1) =$

$$\begin{cases} \sum_{i=1}^{n} \psi_1\{\boldsymbol{Y}_i; \alpha_1 + \sum_{j=0}^{p_{11}} \beta_{11j}(x_{11})(X_{11i} - x_{11})^j + \beta_{120}(X_{12i}) + \ldots + \beta_{1D_10}(X_{1D_1i}), \\ \qquad \alpha_2 + \beta_{210}(X_{21i}) + \ldots + \beta_{2D_20}(X_{2D_2i})\} K_{h_1}(X_{11i} - x_{11}) \left(1, \cdots, (X_{11i} - x_{11})^{p_{11}}\right)^T \\ \qquad\qquad \vdots \\ \sum_{i=1}^{n} \psi_1\{\boldsymbol{Y}_i; \alpha_1 + \beta_{110}(X_{11i}) + \ldots + \beta_{1,D_1-1,0}(X_{1,D_1-1,i}) \\ \qquad + \sum_{j=0}^{p_{1D_1}} \beta_{1D_1j}(x_{1D_1})(X_{1D_1i} - x_{1D_1})^j, \alpha_2 + \beta_{210}(X_{21i}) + \ldots + \beta_{2D_20}(X_{2D_2i})\} \\ \qquad \times K_{h_1}(X_{1D_1i} - x_{1D_1}) \left(1, \cdots, (X_{1D_1i} - x_{1D_1})^{p_{1D_1}}\right)^T, \end{cases}$$

and similarly for $\boldsymbol{\psi}_2(\boldsymbol{x}_2)$.

It should be stressed that the optimality property of the "separated" components (e.g. the marginal means in previous example) remains to hold. Moreover, this property also holds if we don't assume an additive structure but use multivariate smoothing techniques instead of estimating each of these separated curves. For example, for a bivariate

linear model with two covariates, we can estimate $\theta_1(x_1, x_2)$ and $\theta_2(x_1, x_2)$ each with its own optimal bandwidth matrix by solving a set of local estimating equations similar to (4).

5 Discussion

From these results, it is clear that the way estimating equations are defined is important. A wrong choice of estimating equations will yield estimators with undesirable statistical properties. In additive models we observe, from the conditional bias in Theorem 1, that the backfitting approach does not achieve the "optimal" asymptotic bias. In classical regression models (as studied by Opsomer and Ruppert, 1997), this dependence of the bias on all components of the additive model disappears when the covariates are mutually independent. In this case, the asymptotic bias of the estimator for, say, $\theta_1(\cdot)$ does not depend on additive components other than $\theta_1(\cdot)$. In general models, this property does not hold.

The main message of this paper is that for a multiparameter model, there is a set of estimating equations yielding optimal smoothers. It allows full flexibility in selecting the different optimal smoothing parameters for the different curves. An interesting topic for further research is to investigate the performance of data-driven bandwidth selectors.

Acknowledgements

We like to thank a referee for constructive remarks on improving the presentation.

References

Aerts, M. and Claeskens, G. (1997), Local polynomial estimators in multiparameter likelihood models. *J. Am. Stat. Assoc.* 92, 1536–1545.

Bahadur, R.R. (1961), A representation of the joint distribution of responses of n dichotomous items, in *Studies in item analysis and prediction*, H. Solomon (ed.), Stanford Mathematical Studies in the Social Sciences VI. Stanford, California, Stanford University Press.

Carroll, R.J., Ruppert, D. and Welsh, A.H. (1998), Local estimating equations, *J. Am. Stat. Assoc.* 93, 214–227.

Claeskens, G. (1999), Smoothing techniques and bootstrap methods for multiparameter likelihood models. Ph.D. dissertation, Center for Statistics, Limburgs Universitair Centrum, Belgium.

Davison, A.C. and Ramesh, N.I. (1998), Local likelihood smoothing of sample extremes, Technical Report.

Fan, J., Härdle, W. and Mammen, E. (1998), Direct estimation of low dimensional components in additive models, *Ann. Statist.* 26, 943–971.

Fan, J., Heckman, N.E. and Wand, M.P. (1995), Local polynomial kernel regression for generalized linear models and quasi-likelihood functions, *J. Am. Stat. Assoc.* 90, 141–150.

Geys, H., Molenberghs, G. and Ryan, L. (1999), Pseudo-likelihood modeling of multivariate outcomes in developmental toxicology, *J. Am. Stat. Assoc.*, to appear.

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

Liang, K.-Y. and Zeger, S.L. (1986), Longitudinal data analysis using generalized linear models, *Biometrika* 73, 13–22.

Linton, O.B. (1997), Efficient estimation of additive nonparametric regression models, *Biometrika* 84, 469–473.

Linton, O.B. (1998), Efficient estimation of generalized additive nonparametric regression models, Technical Report.

Linton, O. and Nielsen, J.P. (1995), A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika* 82, 93–100.

Molenberghs, G. and Ryan, L.M. (1999), Likelihood inference for clustered multivariate binary data, *Environmetrics*, to appear.

Morgan, B.J.T. (1992), *Analysis of Quantal Response Data,* London: Chapman & Hall.

Opsomer, J.D. and Ruppert, D. (1997), Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* 25, 186–211.

Yuan, K.-H. and Jennrich R.I. (1998), Asymptotics of estimating equations under natural conditions, *J. Mult. Anal.* 65, 245–260.

Zhao, L.P. and Prentice, R.L. (1990), Correlated binary regression using a quadratic exponential model, *Biometrika* 77, 642–648.