Kauermann, Opsomer:

Local Likelihood Estimation in Generalized Additive Models

Sonderforschungsbereich 386, Paper 190 (2000)

Online unter: http://epub.ub.uni-muenchen.de/

LOCAL LIKELIHOOD ESTIMATION IN GENERALIZED ADDITIVE MODELS

Göran Kauermann
University of Glasgow
Glasgow, UK and
Ludwig-Maximilians-Universität
München, Germany

J.D. Opsomer
Iowa State University
Ames, USA

September 5, 2000

Abstract

Generalized additive models are a popular class of multivariate nonparametric regression models, due in large part to the ease of use of the local scoring estimation algorithm. However, the theoretical properties of the local scoring estimator are poorly understood. In this article, we propose a local likelihood estimator for generalized additive models that is closely related to the local scoring estimator fitted by local polynomial regression. We derive the statistical properties of the estimator and show that it achieves the same asymptotic convergence rate as a one-dimensional local polynomial regression estimator. We also propose a wild bootstrap estimator for calculating pointwise confidence intervals for the additive component functions. The practical behavior of the proposed estimator is illustrated through simulation experiments and an example.

Keywords: backfitting, bootstrapping, generalized additive models, local likelihood, local polynomial regression, local scoring, wild bootstrap.

1 Introduction

Generalized additive models (Hastie and Tibshirani, 1990) are a popular approach for fitting multivariate data with known link functions. The `gam()` set of fitting routines and accompanying one-dimensional smoothing methods in S-Plus (Hastie, 1992) are often used for that purpose. These routines implement the *local scoring* algorithm described in Hastie and Tibshirani (1990). Local scoring generalizes Fisher scoring, the most commonly used fitting method for generalized linear models (see, for instance, McCullagh and Nelder, 1989). First result on backfitting estimates in additive models with normal response are found in Buja, Hastie and Tibshirani (1989).

While local scoring as a fitting method is popular in practice, the theoretical properties of the resulting estimators are not well understood. Even such basic properties as consistency have not been generally established. The major difficulty in developing theoretical results for local scoring is that they can only be defined implicitly as the solution to a complicated iterative algorithm. In this sense, the study of generalized additive models is more complicated than that of additive models, where a set of normal equations solved by the estimators can be written down (see Opsomer, 2000). Some results on the properties of local scoring when the univariate smoothers are local polynomials are in Opsomer and Kauermann (2000).

In this article, we propose a new estimator for generalized additive models based on local likelihood estimation (Fan et al. 1995). To differentiate it from local scoring, we will refer to this estimator as the *local likelihood estimator* in what follows. Similarly to local scoring, the local likelihood estimator is a natural extension the local polynomial regression backfitting estimator for additive models, discussed in Opsomer and Ruppert (1997), to the generalized regression model context. For models with linear link link and Gaussian errors, these estimators are all equivalent. Unlike the local scoring estimator, however, the local likelihood estimator is the solution to a set of well-defined normal equations, allowing us to study its statistical properties. In this article, we provide explicit asymptotic bias and variance approximations for the local likelihood estimator and show that the estimator avoids the "curse of dimensionality" by achieving the same convergence rates as one-dimensional nonparametric regression.

Since the normal equations defining the local scoring estimator represent a very large set of simultaneous nonlinear equations and may be difficult to solve in practice, we also propose an iterative algorithm whose solution at convergence is consistent for the "true" local likelihood estimator. Under certain uniqueness conditions, the local scoring estimator at convergence is similarly consistent for the local likelihood estimator, so that the asymptotic properties of the local likelihood estimator provide some insights for the local scoring estimator fitted by local polynomial smoothing.

Linton and Nielsen (1995) introduced a non-iterative fitting method for generalized additive models based on marginal integration, and derived its statistical properties. Using this estimator as a starting point, Linton (2000) proposed a local likelihood procedure to find an fully efficient estimator for individual additive component functions when the remaining component functions are suitably undersmoothed (see also Fan, Mammen and Härdle, 1998). Recently Mammen, Linton and Nielsen (1999) suggest to improve the efficiency of estimates for additive models with normal response based on pilot mean estimate by making use of backfitting. An important difference between these estimation approaches and both local scoring and the local likelihood estimation in the current article is that the latter two methods estimate all additive components simultaneously.

In this article, we consider the generalized additive model

$$E(y|\boldsymbol{x}) = h(\eta) = h\{\alpha + \gamma_1(x_1) + \ldots + \gamma_q(x_q)\} \tag{1}$$

where $h(\cdot)$ is some known link function, $\boldsymbol{x} = (x_1, \ldots x_q)$ are given covariates and $\gamma_r(\cdot)$ are smooth, unknown functions. The response y (for given $\boldsymbol{x}$) is assumed to be distributed according to the exponential family

$$f(y|\boldsymbol{x}) \quad = \quad \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{2}$$

where $\theta = \theta(\eta)$ denotes the natural parameter corresponding to the expectation $h(\eta)$. The dispersion parameter $a(\phi)$ is for ease of notation assumed to be known, as is commonly done in the generalized regression literature. We suppose that a sample $(\boldsymbol{x}_j, y_j), j = 1, \ldots, n$ is available, where y_j is assumed to be drawn from (2) and $\boldsymbol{x}_j = (x_{1j}, \ldots x_{qj})$ is fixed or distributed according to $f_x(\boldsymbol{x})$. The local likelihood estimators for the functions $\gamma_r(\cdot)$ based on such a sample are the topic of this article. We also propose a bootstrap approach for assessing the variability of the estimates by generalizing the *wild boostrap* method of Härdle and Marron (1991) to the generalized additive model.

The remainder of the article is structured as follows. In Section 2, we review univariate local likelihood estimation and introduce the local likelihood estimators for the generalized additive model (1). Section 3 discusses the asymptotic properties of the estimators, and Section 4.1 proposes an algorithm whose solution approximates the local likelihood estimator. In Section 4.2, we explain the relationship between the local likelihood and local scoring estimators. In Section 5, a bootstrap-based inference method for the estimator is proposed. The practical behavior of the method is illustrated in simulation experiments and an example in Section 6.

2 Local Likelihood Estimation

2.1 Univariate Local Likelihood Estimation

Before defining the local likelihood estimators for the generalized additive model (1), we review local likelihood estimation for univariate models. Define the *likelihood contribution* of the jth observation as a function of η:

$$l_j(\eta) = [y_j\theta(\eta) - b\{\theta(\eta)\}]/a(\phi).$$

If $h(\cdot)$ is the canonical link, then the likelihood contribution takes the simple form $l_j(\eta) = \{y_j\eta + b(\eta)\}/a(\phi)$. We define the *score* for the jth observation by $l_{\eta,j}(\eta) = dl_j(\eta)/d\eta$, and drop the parameter argument and write $l_{\eta,j}$ if $l_{\eta,j}(\eta_j)$ is evaluated at the true parameter value for the jth observation. For the canonical link one has $l_{\eta,j}(\eta) = \{y_j - h(\eta)\}/a(\phi)$.

When $q = 1$ and $\alpha = 0$ in (1), the *local likelihood estimator* for the function $\eta \equiv \gamma(\cdot)$ at location x is defined as the maximizer of the local likelihood

$$\max_\beta \sum_{j=1}^n K\left(\frac{x_j - x}{h}\right) l_j(\beta) \tag{3}$$

for a kernel function K and bandwidth h. The maximizer of (3) is found by solving the score equation

$$\sum_{j=1}^n K\left(\frac{x_j - x}{h}\right) l_{\eta,j}(\widehat{\beta}) = 0$$

for $\widehat{\beta}$, so that $\widehat{\gamma}(x) = \widehat{\beta}$. In the Gaussian case with an identity link function, $\widehat{\gamma}(x)$ reduces to the familiar Nadaraya-Watson kernel regression estimator.

Similarly to kernel regression, the local likelihood estimator can readily be generalized to a local polynomial likelihood estimator. Specifically, let $\widehat{\boldsymbol{\beta}} = \{\widehat{\beta}_0, \ldots, \widehat{\beta}_p\}^T$

represent the maximizer of

$$\max_{\boldsymbol{\beta}} \sum_{j=1}^{n} K\left(\frac{x_j - x}{h}\right) l_j(\boldsymbol{X}_{x,j}^T \boldsymbol{\beta}) \tag{4}$$

with $\boldsymbol{X}_{x,j} = \{1, (x_j - x), \ldots, (x_j - x)^p\}^T$. The local polynomial likelihood estimator at location x is defined as $\widehat{\gamma}(x) = \boldsymbol{e}_1^T \widehat{\boldsymbol{\beta}} = \widehat{\beta}_0$, with $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T$. The $p+1$ score equations corresponding to (4) are written jointly as

$$\sum_{j=1}^{n} K\left(\frac{x_j - x}{h}\right) \boldsymbol{X}_{x,j} l_{\eta,j}(\boldsymbol{X}_{x,j}^T \widehat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{5}$$

The properties of $\widehat{\gamma}(\cdot)$ for the simple and local polynomial case have been studied by Fan et al. (1998), Carroll et al. (1998) and Kauermann, Müller, and Carroll (1998). For $p = 0$, Kauermann and Tutz (2000) extend local likelihood estimation to varying coefficient models.

Local likelihood estimators are defined as the solution to non-linear equations, so that in general, no explicit expressions are available for the estimators. However, asymptotic approximations are available, and are useful in clarifying the statistical properties of the estimator. In particular, under suitable assumptions on the underlying statistical model and for sample size n sufficiently large, we can expand (5) about $\boldsymbol{X}_{x,j}\beta$ and apply series inversion as given e.g. in Barndorff-Nielsen and Cox (1989) to find

$$\begin{aligned} 0 &= \sum_j K\left(\frac{x_j - x}{h}\right) \boldsymbol{X}_{x,j} \left[l_{\eta,j}(\boldsymbol{X}_{x,j}\boldsymbol{\beta}) + l_{\eta\eta,j}(\boldsymbol{X}_{x,j}\beta)\{\boldsymbol{X}_{x,j}^T \widehat{\boldsymbol{\beta}} - \boldsymbol{X}_{x,j}\boldsymbol{\beta}\} + \ldots \right] \\ \Leftrightarrow \widehat{\gamma}(x) &= \gamma(x) + \left\{ \boldsymbol{e}_1 \boldsymbol{F}_x^{-1} \sum_j K\left(\frac{x_j - x}{h}\right) \boldsymbol{X}_{x,j} l_{\eta,j} + b(x) \right\} \{1 + o_p(1)\} \end{aligned} \tag{6}$$

where $l_{\eta\eta,j} = \partial l_{\eta,j}(\eta)/\partial\eta$ and $\boldsymbol{F}_x = \sum_j K\{(x_j - x)/h\} v_j \boldsymbol{X}_{x,j} \boldsymbol{X}_{x,j}^T$ is the local Fisher matrix with $v_j = -E(l_{\eta\eta,j})$ and $b(x)$ is the bias component. The bias thereby decomposes to $b(x) = b_{(1)}(x) + b_{(2)}(x)$ with

$$b_{(1)}(x) = \boldsymbol{e}_1 \boldsymbol{F}_x^{-1} \sum_j K\left(\frac{x_j - x}{h}\right) \boldsymbol{X}_{x,j} v_j \{\gamma(x_j) - \boldsymbol{X}_{x,j}\boldsymbol{\beta}\} \tag{7}$$

$$b_{(2)}(x) = -\frac{1}{2} \boldsymbol{e}_1 \boldsymbol{F}_x^{-1} \sum_j K\left(\frac{x_j - x}{h}\right) \boldsymbol{X}_{x,j} v_j' \{\gamma(x_j) - \boldsymbol{X}_{x,j}\boldsymbol{\beta}\}^2 \tag{8}$$

with $v_j' = \partial v_j(\eta_j)/\partial\eta$. Under the usual conditions, i.e. assuming $\gamma(\cdot)$ sufficiently smooth, it follows that $b_{(1)}(x) = O(h^{2\lfloor 1+p/2 \rfloor})$, where $\lfloor \cdot \rfloor$ denotes the largest integer fraction. For the second bias term one finds $b_{(2)}(x) = O(h^{2(1+p/2)})$ for even degree p while $b_{(2)}(x) = O(h^{2\lfloor 1+p/2 \rfloor +2})$ for odd values of p. Hence, the second bias term (8) has negligible asymptotic order compared to (7) if p is odd. For more details, see e.g. Kauermann and Tutz (2000).

We have set $\alpha = 0$ in the local likelihoods (3) and (4) and estimated η as a smooth function of a single covariate. When $q > 1$ and η is assumed to be an additive function of several covariates, the component functions $\gamma_r(\cdot)$ in (1) are not identifiable without additional constraints. For normally distributed y_i and identity link function, the standard constraints are

$$\mathrm{E}_x\{\gamma_r(x_r)\} = 0 \quad \text{for } r = 1, \ldots, q,$$

where E_x denotes the expectation with respect to the design density $f_x(\boldsymbol{x})$, combined with the inclusion of an intercept, so that $\mathrm{E}_x(\eta) = \alpha$. For the general model (1), we will generalize this restriction by making use of the Kullback-Leibler discrepancy measure.

In the univariate case ($q = 1$), suppose that we are interested in uniquely decomposing an unrestricted function $\gamma^o(x)$ in model $E(y|x) = h\{\gamma^o(x)\}$ into α and $\gamma(x)$, such that $\gamma^o(x) = \alpha + \gamma(x)$. This can be done by requiring α to be the minimizer of the Kullback-Leibler distance $\mathcal{K}\{\gamma^o(x), \alpha\} = E_x[E_y\{l(\alpha)|\eta = \gamma^o(x)\}] + const$ with respect to α, where $l(\cdot)$ denotes the log-likelihood function and the inner expectation is carried out using density (2) with $\eta = \gamma^o(x)$ while the outer expectation uses the design density $f_x(\cdot)$. The intercept is therefore defined as the solution to

$$0 \quad = \quad E_x[E_y\{l_\eta(\alpha)|\eta = \gamma^o(x)\}], \tag{9}$$

and we set $\gamma(\cdot) = \gamma^o(\cdot) - \alpha$.

Centered estimators $\widehat{\alpha}$ and $\widehat{\gamma}(\cdot)$ are found by replacing the expectations in (9) by empirical moments. The estimating equation for $\widehat{\alpha}$ is given by

$$0 = \sum_i l_{\eta,i}(\widehat{\alpha}). \tag{10}$$

Hence, the resulting estimator $\widehat{\alpha}$ is the maximum likelihood estimator in the simplified model $E(y|x) = h(\alpha)$ and $\widehat{\gamma}(\cdot) = \widehat{\gamma}^o(\cdot) - \widehat{\alpha}$, where $\widehat{\gamma}^o(\cdot)$ is the solution to (5).

This definition is not directly applicable for centering the additive component functions when $q > 1$, however. From (9), we get by expanding $l_\eta(\alpha)$ about η in first order

$$0 \;=\; E_x[E_y\{l_{\eta\eta}(\eta)\gamma(x)|\eta = \gamma^o(x)\}] + \ldots, \tag{11}$$

since $\mathrm{E}_y\{l_\eta(\eta)\} = 0$. This yields an alternative, asymptotically equivalent identifiability restriction for $\widehat{\gamma}(x)$: by replacing the expectation with respect to x in (11) by empirical moments once again, one obtains $\sum_i v_i\widehat{\gamma}(x_i) = 0$ with $v_i = -E_y(l_{\eta\eta,i})$. Hence, $\widehat{\gamma}^o(x_i)$ is centered by $\widehat{\gamma}(x_i) = \widehat{\gamma}^o(x_i) - \sum_i v_i\widehat{\gamma}(x_i)/\sum v_i$, where in practice, v_i is substituted by an estimate. This centering adjustment now generalizes readily to the case $q > 1$, as will be shown in Section 2.2. For normally distributed y_i with $v_i \equiv \sigma^2$, this also yields the standard mean-centered adjustment used in Opsomer and Ruppert (1998) for the additive model.

Before tackling the estimation of the full generalized additive model (1) in the next section, we consider the extension of the univariate estimator (4) to the so-called "oracle estimator". In this case the model is q-dimensional but all the components functions except the rth function are assumed to be known. The resulting estimate of the rth function will be denoted by $\widehat{\gamma}_{r|-r}(\cdot)$. This means we consider the function $\eta_{-r} = \alpha + \sum_{k\neq r}\gamma_k(\cdot)$ in (1) as known offset and $\gamma_r(\cdot)$ is an unknown univariate function which has to be estimated by local likelihood. The results from univariate local polynomial likelihood fitting given above are readily extended to this situation. In particular, the oracle estimator of $\gamma_r(\cdot)$ at the observation point x_{ri} is $\widehat{\gamma}_{r|-r}(x_{ri}) = \boldsymbol{e}_1^T\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ solves

$$\sum_{j=1}^{n} w_{r,ij}\boldsymbol{X}_{r,ij}l_{\eta,j}(\boldsymbol{X}_{r,ij}^T\widehat{\boldsymbol{\beta}} + \eta_{-r,j}) = \boldsymbol{0}, \tag{12}$$

with $w_{r,ij} = K\{(x_{rj} - x_{ri})/h_r\}$, $\boldsymbol{X}_{r,ij} = \{1, (x_{rj} - x_{ri}), \ldots, (x_{rj} - x_{ri})^{p_r}\}^T$ and $\eta_{-r,j} = \alpha + \sum_{k\neq r}\gamma_k(x_{kj})$. As in (6), expansion of (12) about the true parameter values $\gamma_r(x_{rj})$ yields

$$\widehat{\gamma}_{r|-r}(x_{ri}) \;=\; \gamma_r(x_{ri}) + \left\{\boldsymbol{e}_1^T\boldsymbol{F}_{r,i}^{-1}\sum_j w_{r,ij}\boldsymbol{X}_{r,ij}l_{\eta,j} + b_{r|-r}(x_{ri})\right\}\{1 + o_p(1)\} \tag{13}$$

where $l_{\eta,j} = l_{\eta,j}(\eta_j)$, $v_j = -E(l_{\eta\eta,j})$, $\boldsymbol{F}_{r,i} = \sum_j w_{r,ij}v_j\boldsymbol{X}_{r,ij}\boldsymbol{X}_{r,ij}^T$ and the bias $b_{r|-r}(x_{ri}) = b_{r|-r(1)}(x_{ri}) + b_{r|-r(2)}(x_{ri})$ as in (7) and (8). Defining $\boldsymbol{S}_r$ as weighted smoothing matrix with ijth element

$$[\boldsymbol{S}_r]_{ij} \;=\; w_{r,ij}\boldsymbol{e}_1^T\boldsymbol{F}_{r,i}^{-1}\boldsymbol{X}_{r,ij}. \tag{14}$$

we can rewrite (13) more compactly as

$$\widehat{\boldsymbol{\gamma}}_{r|-r} \;=\; (\boldsymbol{S}_r\boldsymbol{l}_\eta + \boldsymbol{S}_r\boldsymbol{V}\boldsymbol{\gamma}_r + \boldsymbol{b}_{r|-r(2)})\{1+o_p(1)\} \tag{15}$$

with $\widehat{\boldsymbol{\gamma}}_{r|-r} = \{\widehat{\gamma}_{r|-r}(x_{r1}),\ldots,\widehat{\gamma}_{r|-r}(x_{rn})\}^T$, $\boldsymbol{l}_\eta = (l_{\eta,1},\ldots,l_{\eta,n})^T$, $\boldsymbol{V} = \mathrm{diag}(v_1,\ldots v_n)$, $\boldsymbol{\gamma}_r = \{\gamma_r(x_{r1}),\ldots,\gamma_r(x_{rn})\}^T$ and $\boldsymbol{b}_{r|-r(2)} = \{b_{r|-r(2)}(x_{r1}),\ldots,b_{r|-r(2)}(x_{r1})\}^T$, since $\boldsymbol{b}_{r|-r(1)} = \boldsymbol{S}_r\boldsymbol{V}\boldsymbol{\gamma}_r - \boldsymbol{\gamma}_r$. As argued above, $\boldsymbol{b}_{r|-r(2)}$ is of negligible asymptotic order if p_r is odd. This is assumed throughout the remaining of the paper. The estimated function $\widehat{\boldsymbol{\gamma}}_{r|-r}(\cdot)$ is not centered. Using the same approach as above, this is achieved by replacing $\boldsymbol{S}_r$ in (15) by $\boldsymbol{S}_r^+ = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T\boldsymbol{V}/\{\sum_i v_i\})\boldsymbol{S}_r$, where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{1}$ is the $n\times 1$ vector $(1,\ldots 1)^T$.

Lemma A.1 in the Appendix gives asymptotic bias and variance approximations for the oracle estimator. These resuls are a direct generalization of those for univariate local likelihood estimators. They will be used in the derivation of the statistical properties of the full local likelihood estimator in the next section.

2.2 Local Likelihood Estimation for Generalized Additive Models

We now extend the univariate local polynomial likelihood results from Section 2.1 to the generalized additive model (1), with all $q > 1$ additive component functions unknown. Suppose that we are interested in using local polynomial likelihood estimation with odd degrees $p_r, r = 1,\ldots,q$. For all r, the estimators for $\gamma_r(x_{ri}), i = 1,\ldots,n$, are formally defined as $\widehat{\gamma}_r(x_{ri}) = \boldsymbol{e}_1^T\widehat{\boldsymbol{\beta}}_{ri}$, where $\widehat{\boldsymbol{\beta}}_{ri} = \{\widehat{\beta}_{ri,0},\ldots,\widehat{\beta}_{ri,p_r}\}^T$ is the maximizer of the local polynomial likelihood

$$\max_{\boldsymbol{\beta}_{ri}} \sum_{j=1}^n w_{r,ij} l_j(\boldsymbol{X}_{r,ij}^T\boldsymbol{\beta}_{ri} + \widehat{\eta}_{-r,j}), \tag{16}$$

with $\widehat{\eta}_{-r,j} = \widehat{\alpha} + \sum_{k\neq r}\widehat{\gamma}_k(x_{kj})$, $j = 1,\ldots,n$. Maximization (16) has to be done jointly for all covariates, subject to the constraints

$$\sum_{i=1}^n \widehat{v}_i\widehat{\gamma}_r(x_{ri}) = 0 \quad \text{for } r = 1,\ldots,q, \tag{17}$$

where $\widehat{v}_i = v_i(\widehat{\eta}_i)$ with $\widehat{\eta}_i = \widehat{\alpha} + \widehat{\gamma}_{1i}(x_{1i}) + \ldots, \widehat{\gamma}_{qi}(x_{qi})$. These constraints are satisfied when a solution of (16) is centered by replacing it with $\widehat{\gamma}_r(x_{ri}) - \sum_i \widehat{v}_i\widehat{\gamma}_r(x_{ri})/\sum_i \widehat{v}_i$. The intercept estimator $\widehat{\alpha}$ is the solution of

$$\sum_{i=1}^n l_{\eta,j}(\widehat{\alpha}) = 0 \tag{18}$$

as shown in the previous section.

The local polynomial likelihood estimators for the generalized additive model solve a $n \times \sum_{r=1}^{q}(p_r+1)+q+1$ dimensional system of equations composed of the non-linear score equations

$$\sum_{j=1}^{n} w_{r,ij}\boldsymbol{X}_{r,ij} l_{\eta,j}(\boldsymbol{X}_{r,ij}^T \widehat{\boldsymbol{\beta}}_{ri} + \widehat{\eta}_{-r,j}) = \mathbf{0} \tag{19}$$

for $i = 1, \dots, n$, $r = 1, \dots, q$, subject to (17), and the score equation (18) for $\widehat{\alpha}$. Solving this system of equations represents a formidable task if attempted directly. In Section 4.1, we propose a more practical backfitting estimator that approximates the solution corresponding to these equations. We assume for now that a solution to (19) can be found and is unique, so that the local likelihood estimator is well-defined, and we first discuss its statistical properties.

The local likelihood estimator is related to the oracle estimators, in a manner that will be made precise in Theorem 2.1 below. Before stating the theorem, we list the assumptions used in this section. Let $f_r(\cdot)$ represent the marginal density of $x_r, r = 1, \dots, q$, and $\mu_p(K) = \int u^p K(u)du$.

- **A1** *For p_r odd, the (p_r+1)th derivative of $\gamma_r(\cdot)$ exist for $r = 1, \dots, q$ and they are continuous and bounded.*
- **A2** *The kernel K is bounded and continuous, has compact support and for $r = 1, \dots, q$, we have $\mu_{p_r+1}(K) \neq 0$.*
- **A3** *The marginal design densities $f_r(\cdot)$, $r = 1, \dots, q$ have compact support and their first derivatives are continuous and bounded.*
- **A4** *The functions $v(\eta) = E\{-l_{\eta\eta}(\eta)\}$ and $v_r(x_r) = E_x\{v(\eta)|X_r = x_r\}$ are positive, continuously differentiable, bounded and bounded away from zero.*
- **A5** *As $n \to \infty$, $h_r \to 0$ and $nh_r \to \infty$ for all $r = 1, \dots, q$.*

We define the $nq \times 1$ vectors $\boldsymbol{\gamma}_{\bullet} = (\boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_q^T)^T$, $\widehat{\boldsymbol{\gamma}}_{\bullet} = (\widehat{\boldsymbol{\gamma}}_1^T, \dots, \widehat{\boldsymbol{\gamma}}_q^T)^T$ and $\widehat{\boldsymbol{\gamma}}_{\bullet|-\bullet} = (\widehat{\boldsymbol{\gamma}}_{1|-1}^T, \dots, \widehat{\boldsymbol{\gamma}}_{q|-q}^T)^T$, and the $nq \times nq$ matrix

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{S}_1^+\boldsymbol{V} & \cdots & \boldsymbol{S}_1^+\boldsymbol{V} \\ \boldsymbol{S}_2^+\boldsymbol{V} & \boldsymbol{I} & \cdots & \boldsymbol{S}_2^+\boldsymbol{V} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{S}_q^+\boldsymbol{V} & \boldsymbol{S}_q^+\boldsymbol{V} & \cdots & \boldsymbol{I} \end{bmatrix}.$$

With $\boldsymbol{h}_{\bullet} = (h_1\mathbf{1}^T, \ldots, h_q\mathbf{1}^T)^T$ we denote the vector of bandwidths, where $\mathbf{1}$ is the $n \times 1$ vector $(1, \ldots, 1)^T$. Finally, with $\boldsymbol{A}^{[p]}$ we denote the component-wise pth power of matrix $\boldsymbol{A}$, i.e. the matrix with ijth element A_{ij}^p.

Theorem 2.1 *Under assumptions A1–A5, the local likelihood estimator and the oracle estimator are related through the following asymptotic equality*

$$\widehat{\boldsymbol{\gamma}}_{\bullet|-\bullet} - \boldsymbol{\gamma}_{\bullet} = \boldsymbol{M}(\widehat{\boldsymbol{\gamma}}_{\bullet} - \boldsymbol{\gamma}_{\bullet})\left\{1 + O\left(\boldsymbol{h}_{\bullet}^{[2]}\right)\right\} + O\left(\{\boldsymbol{M}(\widehat{\boldsymbol{\gamma}}_{\bullet} - \boldsymbol{\gamma}_{\bullet})\}^{[2]}\right) \tag{20}$$

where the $O(\cdot)$ terms hold component-wise. If $\boldsymbol{M}$ is invertible, (20) is equivalent to $\widehat{\boldsymbol{\gamma}}_{\bullet} - \boldsymbol{\gamma}_{\bullet} = \boldsymbol{M}^{-1}(\widehat{\boldsymbol{\gamma}}_{\bullet|-\bullet} - \boldsymbol{\gamma}_{\bullet})\{1 + o_p(1)\}$

Remarks:

1. By Lemma A.1 in the Appendix, we know that the oracle estimators in $\widehat{\boldsymbol{\gamma}}_{\bullet|-\bullet}$ are consistent for $\boldsymbol{\gamma}_{\bullet}$ under assumptions A1–A5, so that both sides of (20) converge to 0. Since the rate of the approximation term on the right-hand side is faster than that of the leading term, we can conclude that $\widehat{\boldsymbol{\gamma}}_{\bullet}$ is also consistent for $\boldsymbol{\gamma}_{\bullet}$.

2. The matrix $\boldsymbol{M}$ has the same structure as that used to define the backfitting estimators for additive models in Opsomer and Ruppert (1999) and for weighted additive models in Opsomer and Kauermann (2000). In both applications, the existence of the estimators was shown to depend on the invertability of this matrix. Similarly, explicit expressions for the asymptotic properties of the local likelihood estimator will require invertability of $\boldsymbol{M}$. This will be discussed in the next section.

3. The $O\left(\boldsymbol{h}_{\bullet}^{[2]}\right)$ rate of the approximation in Theorem 2.1 holds for local likelihood estimators of arbitrary degree $p_1, \ldots, p_q$. It should be noted that this approximation rate represents a *relative* rate for the difference between the convergence rates of the oracle and the local likelihood estimators. Both these rates indeed depend directly on the p_r (see next section).

4. Theorem 2.1 also holds for even degrees p_r. However, since the matrix form (15) for the oracle estimator is disturbed by an additional bias term, we do not pursue even degrees here.

3 Asymptotic Properties

In this section, we use Theorem 2.1 to approximate the conditional bias and variance of the local likelihood estimators for generalized additive models. For simplicity, we only discuss the case $q = 2$ and p_1, p_2 odd. As for additive models, the bias and variance expressions become much more complicated for models with $q > 2$ (see Opsomer, 2000).

We write $v(x_1, x_2) = v\{\alpha + \gamma_1(x_1) + \gamma_2(x_2)\}$, the variance function evaluated at (x_1, x_2), and $v_r(x_r) = \mathrm{E}_x(v(X_1, X_2)|X_r = x_r)$. Define $\boldsymbol{T}^*_{12}$ as the $n \times n$ matrix whose ijth element is

$$[\boldsymbol{T}^*_{12}]_{ij} = \frac{1}{n} \frac{f_x(x_{1i}, x_{2j})}{f_1(x_{1i}) f_2(x_{2j})} \frac{v(x_{1i}, x_{2j})}{v_1(x_{1i}) v_2(x_{2j})} v(x_{1j}, x_{2j}) - \frac{1}{n}.$$

Finally, let $R(K) = \int K(u)^2 du$. We replace assumptions A2–A5 by the following

- **A2**' *The kernel K is bounded and continuous, it has compact support and its first derivative has a finite number of sign changes over its support. Also, $\mu_{p_r+1}(K) \neq 0$ for $r = 1, 2$.*
- **A3**' *The densities f_x, f_1, f_2 are bounded and continuous, they have compact support and their first derivatives have a finite number of sign changes over their supports. Also, $f_1(x_1), f_2(x_2) > 0$ for all $(x_1, x_2) \in supp(f)$.*
- **A4**' *The functions v, v_1, v_2 for are bounded, continuous and differentiable. Their first derivatives have a finite number of sign changes. $v_1(x_1), v_2(x_2) > 0$ for all $(x_1, x_2) \in supp(f)$.*
- **A5**' *As $n \to \infty$, $h_1, h_2 \to 0$ and $nh_1/\log(n), nh_2/\log(n) \to \infty$.*
- **A6**' *There exists a matrix norm $\|\cdot\|$, such that $\|\boldsymbol{T}^*_{12}\| < 1$.*

Theorem 3.1 *Under the assumptions A1, A2' –A6', the conditional bias of $\widehat{\gamma}_1(x_{1i})$ is approximated by*

$$E(\widehat{\gamma}_1(x_{1i}) - \gamma_1(x_{1i})|\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{e}_i^T(\boldsymbol{I} - \boldsymbol{T}^*_{12})^{-1} \times$$
$$\left\{ h_1^{p_1+1} \frac{\mu_{p_1+1}(K)}{(p_1+1)!} \left(\boldsymbol{\gamma}_1^{(p_1+1)} - \frac{E_x(v(X_1, X_2)\gamma_1^{(p_1+1)}(X_1))}{E_x(v(X_1, X_2))} \right) \right.$$

$$+h_2^{p_2+1}\frac{\mu_{p_2+1}(K)}{(p_2+1)!}\left(E(\gamma_2^{(p_2+1)}(X_2)|\boldsymbol{X}_1)-\frac{E_x(v(X_1,X_2)\gamma_2^{(p_2+1)}(X_2))}{E_x(v(X_1,X_2))}\right)\Bigg\}$$
$$+o_p(h_1^{p_1+1}+h_2^{p_2+1}).$$

The conditional variance of $\widehat{\gamma}_1(x_{1i})$ *is approximated by*

$$Var(\widehat{\gamma}_1(x_{1i})|\boldsymbol{X}_1,\boldsymbol{X}_2)=\frac{1}{nh_1}R(K)v_1(x_{1i})^{-1}f_1(x_{1i})^{-1}(1+o_p(1)). \tag{21}$$

The proof of Theorem 3.1 as well as two required technical lemmas are in the Appendix. The results for $\widehat{\gamma}_2(x_{2i})$ and $\widehat{\eta}_i$ are completely analogous. Recursive expressions for $q>2$ can be derived using the approach of (Opsomer, 2000).

Remarks:

1. For the identity link and Gaussian errors, the asymptotic bias and variance in Theorem 3.1 simplify to those found for the additive model described in Opsomer and Ruppert (1997). In that sense, the local likelihood estimator proposed here is the direct extension of the local polynomial regression backfitting approach of Opsomer and Ruppert (1997) to the generalized additive model context.

2. The theorem also shows that the local likelihood estimator for generalized additive models shares the desirable property of *dimension reduction* with additive model backfitting: the convergence rates of the estimator of a q-dimensional model are the same as those for a one-dimensional model. Stone (1986) found the same result for generalized additive models fitted by maximum likelihood using additive regression splines.

We end this section on the statistical properties of the local polynomial likelihood estimator with the following corrolary.

Corollary 3.1 *If* $\boldsymbol{M}$ *is invertible and under assumptions A1 and A2'–A5', the local likelihood estimator* $\widehat{\gamma}_r(x_{ri})$ *has the following asymptotic distribution:*

$$\widehat{\gamma}_r(x_{ri})\ \overset{\mathcal{L}}{\longrightarrow}\ \mathcal{N}(\gamma_r(x_{ri})+B(x_{ri}),\ \tfrac{1}{nh_r}R(K)v_r(x_{ri})^{-1}f_r(x_{ri})^{-1})$$

with $B(x_{ri})$ *the conditional bias approximation as given in Theorem 3.1.*

The result follows easily from the asymptotic normality of the oracle estimates, which in turn results from standard arguments available from univariate smoothing (see e.g. Fan and Gijbels, 1996).

4 Estimation algorithms

4.1 Local Likelihood Backfitting

In this section, we propose a backfitting algorithm for generalized additive models based on local polynomial likelihood estimation. In order to differentiate the solution of this algorithm from the local likelihood estimator discussed so far, we will refer to it as the *local likelihood backfitting estimator.* We will show below that, while both estimators are not exactly equal, they are asymptotically equivalent under certain conditions.

The algorithm is displayed in Figure 1. Step 2 is the main step in the algorithm. It directly results from (15) and it resembles a one-step Fisher scoring for solving (19), i.e.

$$\begin{aligned}
\widehat{\gamma}_r^{(t+1)}(x_{ri}) &= \widehat{\gamma}_r^{(t)}(x_{ri}) + \boldsymbol{e}_1^T \widehat{\boldsymbol{F}}_{r,i}^{(t)^{-1}} \sum_j w_{r,ij} \boldsymbol{X}_{r,ij} l_{\eta,j}(\boldsymbol{X}_{r,ij}^T \widehat{\boldsymbol{\beta}}_{ri}^{(t)} + \widehat{\eta}_{-r,j}^{(t)}) \\
&= \widehat{\gamma}_r^{(t)}(x_{ri}) + e_1 \widehat{\boldsymbol{F}}_{r,i}^{(t)^{-1}} \sum_j w_{r,ij} \boldsymbol{X}_{r,ij} [\widehat{l}_{\eta,j}^{(t)} - \widehat{v}_j^{(t)} \{\boldsymbol{X}_{r,ij}^T \widehat{\boldsymbol{\beta}}_{ri}^{(t)} - \widehat{\gamma}_r^{(t)}(x_{rj})\}] + \ldots \\
&\approx \boldsymbol{e}_1^T \widehat{\boldsymbol{F}}_{r,i}^{(t)^{-1}} \sum_j w_{r,ij} \boldsymbol{X}_{r,ij} \{\widehat{l}_{\eta,j}^{(t)} + \widehat{v}_j^{(t)} \widehat{\gamma}_r^{(t)}(x_{rj})\} \qquad (22)
\end{aligned}$$

with $\widehat{\boldsymbol{F}}_{r,i}^{(t)}$ as plug in estimate of $\boldsymbol{F}_{r,i}$, $\widehat{v}_t^{(t)} = v_j(\widehat{\eta}_j^{(t)})$ and $\widehat{l}_\eta^{(t)} = l_{\eta,j}(\widehat{\eta}_j)$. This means that, instead of solving (19) explicitly, we update $\widehat{\gamma}_r^{(t+1)}(x_{ri})$ by a one-step Fisher scoring only.

Suppose that the local likelihood backfitting algorithm has fully converged, so that $\widehat{\boldsymbol{\gamma}}_r^{(t+1)} = \widehat{\boldsymbol{\gamma}}_r^{(t)}$ for $r = 1, \ldots, q$, and that this solution is unique. Let $\widehat{\boldsymbol{\gamma}}_r^{(\infty)}, r = 1, \ldots, q$ denote this estimator. Suppose also that the solutions to the score equations (19) and the associated identification restrictions exist and are unique. Let $\widehat{\boldsymbol{\gamma}}_r, r = 1, \ldots, q$ denote this estimator. In general, both estimators will not be exactly equal to each other, but for sufficiently large samples they are asymptotically equivalent, as the following theorem shows.

```
Generalized Additive Model Local Likelihood Algorithm

1 Initialize the parameters e.g. by some parametric model fit and
  denote the resulting preliminary fits by $\widehat{\alpha}^{(t)}$ and $\widehat{\boldsymbol{\eta}}^{(t)} = \sum_l \widehat{\gamma}_l^{(t)} + \widehat{\alpha}^{(t)}$ .

2 For $1 \leq r \leq q$ update $\widehat{\boldsymbol{\gamma}}_r^{(t)}$ in the following way: Calculate $\widehat{\boldsymbol{S}}_r^{+(t)}$ and
  $\widehat{\boldsymbol{V}}^{(t)} = \mathrm{diag}(\widehat{v}_1, \ldots, \widehat{v}_n)$ by using $\widehat{\boldsymbol{\eta}}^{(t)}$ as plug-in estimate, and compute

      $\widehat{\boldsymbol{\gamma}}_r^{(t+1)} = \widehat{\boldsymbol{S}}_r^{+(t)} \widehat{\boldsymbol{l}}_{\boldsymbol{\eta}}^{(t)} + \widehat{\boldsymbol{S}}_r^{+(t)} \widehat{\boldsymbol{V}}^{(t)} \widehat{\boldsymbol{\gamma}}_r^{(t)}$

  where $\widehat{\boldsymbol{l}}_\eta^{(t)} = \boldsymbol{l}_\eta(\widehat{\boldsymbol{\eta}}^{(t)})$ is the score with plug-in estimates.

3 Update $\widehat{\alpha}$ by setting $\widehat{\alpha}^{(t+1)} = \widehat{\boldsymbol{P}}^{(t)^T} \widehat{\boldsymbol{l}}_\eta^{(t)}$ with $\widehat{\boldsymbol{P}}^{(t)} = (\mathbf{1}^T \widehat{\boldsymbol{V}}^{(t)} \mathbf{1})^{-1} \mathbf{1}$.

4 Set $\widehat{\boldsymbol{\eta}}^{(t+1)} = \sum_l \widehat{\gamma}_l^{(t+1)} + \widehat{\alpha}^{(t+1)}$.

5 Iterate Steps 2–4 until convergence is achieved.
```

Figure 1: Local likelihood backfitting algorithm for generalized additive models.

Theorem 4.1 *Assuming A.1 to A.5 hold, that the local likelihood estimator defined through (19) is unique and that the local likelihood backfitting estimate* $\widehat{\boldsymbol{\gamma}}_{\bullet}^{(\infty)} = (\widehat{\boldsymbol{\gamma}}_{\bullet}^{(\infty)}, \ldots, \widehat{\boldsymbol{\gamma}}_{\bullet}^{(\infty)})^T$ *resulting from the algorithm in Figure 1 exists and is unique. Then,*

$$\begin{aligned} 0 &= \widehat{\boldsymbol{M}}(\widehat{\boldsymbol{\gamma}}_{\bullet} - \widehat{\boldsymbol{\gamma}}_{\bullet}^{(\infty)})\{1 + O(\boldsymbol{h}^{[2]})\} \\ &\quad + O\left(\left\{\widehat{\boldsymbol{M}}(\widehat{\boldsymbol{\gamma}}_{\bullet} - \widehat{\boldsymbol{\gamma}}_{\bullet}^{(\infty)})\right\}^{[2]}\right) + O\left(\sum_{r=1}^{q} \{\boldsymbol{S}_r(\widehat{\boldsymbol{\gamma}}_r - \widehat{\boldsymbol{\gamma}}_r^{(\infty)})\}^{[2]}\right), \end{aligned} \tag{23}$$

where the hat notation indicates plug-in estimates. If $\widehat{\mathbf{M}}$ *is invertible, this is equivalent to* $\widehat{\boldsymbol{\gamma}}_{\bullet} = \widehat{\boldsymbol{\gamma}}_{\bullet}^{(\infty)}\{1 + o_p(1)\}$.

The proof is similar to the proof of Theorem 2.1, as shown the appendix.

4.2 Relationship with local scoring estimators

There are a number of similarities and important differences between local likelihood backfitting estimation and local scoring as described by Hastie and Tibshirani

(1990), page 141. At each iteration step, local scoring calculates

$$z_i^{(t)} \;=\; \widehat{\eta}_i^{(t)} + \left(\frac{\partial h(\widehat{\eta}_i^{(t)})}{\partial \eta}\right)^{-1} \{y_i - h(\widehat{\eta}_t^{(t)})\}$$

and weights $\widehat{v}_i^{(t)} = \{\partial h(\widehat{\eta}_i^{(t)})/\partial \eta\}^2 Var(y_i|\widehat{\eta}_i^{(t)})^{-1}$, followed by a weighted backfitting, where

$$\widehat{\boldsymbol{\gamma}}_r^{(t+1)} = \boldsymbol{S}_r^{(t)} \widehat{\boldsymbol{V}}^{(t)} (\boldsymbol{z}^{(t)} - \widehat{\boldsymbol{\eta}}_{-r}^{(t+1)}) \qquad r = 1, \ldots, q \tag{24}$$

are iterated to convergence, with $\boldsymbol{z}^{(t)} = (z_1^{(t)}, \ldots z_n^{(t)})^T$. Because $\widehat{v}_i^{(t)} z_i^{(t)} = \widehat{v}_i^{(t)} \widehat{\eta}_i^{(t)} + l_{\eta,i}(\widehat{\eta}_i^{(t)})$, step 2 in the local likelihood algorithm can be rewritten as

$$\widehat{\boldsymbol{\gamma}}_r^{(t+1)} = \boldsymbol{S}_r^{(t)} \widehat{\boldsymbol{V}}^{(t)} (\boldsymbol{z}^{(t)} - \widehat{\boldsymbol{\eta}}_{-r}^{(t)}).$$

This is very similar to calculating (24) in local scoring. The critical difference is that at each iteration of the algorithm, local scoring calculates this step to convergence among the q additive component functions, whereas local likelihood backfitting only performs a one-step update. Since this calculation is itself done on a first-order approximation of the data (the z_i), it seems reasonable to assume that full convergence of the additive model step might represent unnecessary precision and slow down the algorithm. This computational efficiency issue is outside the scope of the current article.

The same reasoning as in Theorem 4.1 also applies to the fully converged local scoring estimator, since (24) can be written as (22) when the values for t and $t+1$ are identical. Therefore, the asymptotic results of Sections 2 and 3 also hold for both the local likelihood backfitting and the local scoring estimators.

5 Inference for local likelihood estimators

For point-wise inference about the estimates, local variance bands of the form $\widehat{\gamma}_r(x_{ir}) \pm z_{\alpha/2}\sqrt{\mathrm{Var}(\widehat{\gamma}_r(x_{ir}))}$ are desired, for some distribution with quantiles $z_{\alpha/2}$. In this type of pointwise bands, the bias is usually ignored and it is customary to use the Gaussian distribution as an approximation to the true distribution, as done in Corollary 3.1. In order to use these confidence bands, an estimate for the variance of $\widehat{\gamma}_r(x_{ir})$ is required, and the fully asymptotic approximation in Theorem 3.1 is rarely satisfactory in this respect. From Theorem 2.1 and equation (15), it is easy to derive the first order approximation

$$\widehat{\boldsymbol{\gamma}}_{\bullet} \approx \boldsymbol{M}^{-1}\boldsymbol{S}_{\bullet}^{+}(\boldsymbol{l}_{\eta} + \boldsymbol{V}\boldsymbol{\eta})$$

with $\boldsymbol{S}_{\bullet}^{+} = (\boldsymbol{S}_1^{+^T}, \ldots, \boldsymbol{S}_q^{+^T})^T$. For the rth component,

$$\widehat{\gamma}_r \approx \boldsymbol{Q}_r(\boldsymbol{l}_{\eta} + \boldsymbol{V}\boldsymbol{\eta}) \tag{25}$$

where $\boldsymbol{Q}_r = \boldsymbol{M}^{rr}\boldsymbol{S}_r(\boldsymbol{I} - \boldsymbol{V}\boldsymbol{S}_{-r}^{+})$ with $\boldsymbol{M}^{rr}$ as rth $n \times n$ block diagonal matrix of $\boldsymbol{M}^{-1}$ and $\boldsymbol{S}_{-r}^{+} = \sum_{k \neq r} \boldsymbol{S}_{-k}^{+}$. Hence, the variance equals in first order approximation $\text{Var}(\widehat{\gamma}_{r,i}) = \boldsymbol{Q}_r\boldsymbol{V}\boldsymbol{Q}_r^T\{1 + o_p(1)\}$. This can be estimated by plugging in estimators for v_j, η_j as required. In practice, matrix $\boldsymbol{Q}_r$ is numerically difficult to calculate directly, since it requires inverting an $nq \times nq$ matrix. Instead, using the definition of $\boldsymbol{Q}_r$, it follows directly that $\boldsymbol{Q}_r$ can be calculated by solving the fixpoint equations

$$\boldsymbol{Q}_r \;=\; \boldsymbol{S}_r^{+}(\boldsymbol{I} - \boldsymbol{V}\sum_{l \neq r}\boldsymbol{Q}_l) \tag{26}$$

for $r = 1, \ldots q$, which can be applied iteratively in a backfitting fashion. It should be noted that Hastie and Tibshirani (1990) use an approximation similar to (25) to derive a variance estimator for local scoring, which is also implemented e.g. in the `gam()` procedure in Splus.

An alternative approximation is given by $\text{Var}(\widehat{\boldsymbol{\gamma}}_r) \approx \widehat{\boldsymbol{S}}_r^{+}\boldsymbol{V}\widehat{\boldsymbol{S}}_r^{+^T}$, based on the fact that the asymptotic variance of the oracle estimator in Lemma A.1 and that of the local likelihood estimator in Theorem 3.1 are identical. This can be easily estimated by plugging in estimators for v_j, η_j. We will compare both approximations in the following section.

Even if such asymptotic variance approximations are reasonable, the confidence intervals based on them ignore the bias and the distribution of the estimators, and are difficult to generalize to simultaneous intervals. We are therefore interested in providing an alternative inference approach based on bootstrapping. This will require the use of different bandwidths to ensure convergence of the bootstrap (Efron and Tibshirani, 1993). Because of this, we include an additional subscripts $\{\boldsymbol{h}\}$ indicating the bandwidth used for fitting, e.g. $\widehat{\boldsymbol{\gamma}}_{r,\{\boldsymbol{h}\}}$ is the backfitting estimate calculated with bandwidth $\boldsymbol{h} = (h_1, \ldots, h_q)$, or $\widehat{\boldsymbol{\gamma}}_{r|-r,\{h_r\}}$ is the oracle estimate calculated with bandwidth h_r. Moreover $\boldsymbol{S}_{r,\{h_r\}}$ is the smoothing matrix calculated with bandwidth h_r and $\boldsymbol{Q}_{r,\{\boldsymbol{h}\}}$ solves (26) for $S_{k,\{h_k\}}$, $k = 1, \ldots q$.

We now apply the bootstrap principle to (25). This means we replace the left hand side of the equation by bootstrap replicates and replace the parameters on the right hand side by their corresponding estimates. This leads to the *local likelihood bootstrap*

$$\widehat{\boldsymbol{\gamma}}_r^* \;=\; \widehat{\boldsymbol{Q}}_{r,\{\boldsymbol{h}\}}(\widehat{\boldsymbol{l}}^*_{\eta,\{\boldsymbol{h}\}} + \widehat{\boldsymbol{V}}\widehat{\boldsymbol{\eta}}_{\{\boldsymbol{g}\}}) \tag{27}$$

where plug-in estimates are used for $\boldsymbol{Q}_r$ and $\boldsymbol{V}$, $\widehat{\boldsymbol{l}}^*_{\eta,\{h\}}$ is a bootstrap sample from $\widehat{\boldsymbol{l}}_{\eta,\{h\}} \equiv l_\eta(\widehat{\boldsymbol{\eta}}_{\{h\}})$ and $\widehat{\boldsymbol{\eta}}_{\{\boldsymbol{g}\}}$ is calculated using a second bandwidth $\boldsymbol{g} = (g_1, \dots g_q)$ which tends to zero more slowly than the bandwidth $\boldsymbol{h}$ as specified below. In principle there are various ways to draw $\widehat{\boldsymbol{l}}^*_{\eta,\{h\}}$ in order to obtain appropriate convergence. A convenient method is to use *wild bootstrapping* as introduced by Härdle and Marron (1991) for normal response smoothing models and extended to generalized smoothing models in Galindo et al. (2000). In wild bootstrapping, the elements of $\widehat{\boldsymbol{l}}^*_{\eta,\{h\}}$ are drawn independently from a two point distribution with masspoints $\widehat{\boldsymbol{l}}_{\eta,\{h\},i} \times \{(1-5^{1/2}), (1+5^{1/2})\}$, $i = 1, \dots n$, and corresponding masses $(c, 1-c)$ where $c = (5+5^{1/2})/10$. This guarantees that the first three bootstrap moments match the empirical ones. The following theorem shows that the local likelihood bootstrap $\widehat{\boldsymbol{\gamma}}^*_{r,\{h\}}$ converges in distribution.

Theorem 5.1 *Under the assumptions A1, A2' –A6' and for bandwidth $\boldsymbol{g} = (g_1, \dots g_q)$ chosen such that the $(p_r{+}1)$th order derivatives are fitted consistently, i.e. $\widehat{\boldsymbol{\gamma}}^{(p_r+1)}_{r,\{g\}}(x_{ri}) - \boldsymbol{\gamma}^{(p_r+1)}_{r,\{g\}}(x_{ri}) = o_p(1)$, for $i = 1, \dots, n$ and $r = 1, \dots q$ as $n \to \infty$, the local likelihood bootstrap $\widehat{\boldsymbol{\gamma}}^*_{r,\{h\}}$ converges in distribution to $\widehat{\boldsymbol{\gamma}}_{r,\{h\}}$, i.e.*

$$\{\widehat{\gamma}^*_{r,\{h\}}(x_{ri}) - \widehat{\gamma}_{r,\{g\}}(x_{ri})\} \overset{\mathcal{L}}{\to} \{\widehat{\gamma}_{r,\{h\}}(x_{ri}) - \gamma_r(x_{ri})\} \tag{28}$$

In the same fashion one can also expand (18) to derive bootstraps for the intercept α. This is not further explored here.

In (15), the score vector $\boldsymbol{l}_\eta$ serves as vector of residuals while in (27) we bootstrap from the fitted scores which mirror fitted residuals. As usual in regression models, squared residuals underestimate the true squared errors in mean. It is therefore advisable to increase the components of $\widehat{\boldsymbol{l}}_{\eta,\{\boldsymbol{h}\}}$ by some multiplicator $a_i > 1$, say, such that

$$E(a_i \widehat{l}^2_{\eta,i,\{h\}}) \approx E(l^2_{\eta,i}) \tag{29}$$

holds up to the second asymptotic order. In particular one finds for the ith component of $\widehat{\boldsymbol{l}}_{\eta,\{\boldsymbol{h}\}}$ by simple expansion

$$E(\widehat{l}^2_{\eta,i,\{\boldsymbol{h}\}}) = E(\{l_\eta - v_i\boldsymbol{Q}_{+,i\cdot}\boldsymbol{l}_\eta + \ldots\}^2) \approx v_i - 2v_i^2\boldsymbol{Q}_{+,ii} + v_i\boldsymbol{Q}_{+,i\cdot}\boldsymbol{V}\boldsymbol{Q}^T_{+,i\cdot}v_i$$

where $\boldsymbol{Q}_+ = \sum_{r=1}^q \boldsymbol{Q}_r$, with $\boldsymbol{Q}_{+,i\cdot}$ and $\boldsymbol{Q}_{+,ii}$ the ith row and the ith diagonal element of $\boldsymbol{Q}_+$, respectively. Using $\boldsymbol{Q}_r \approx \boldsymbol{S}$ (based on Theorem 3.1) and the fact that $\boldsymbol{S}_{r,ii} = O\{(nh_r)^{-1}\}$, we find that a_i can be chosen as $a_i = (1 - v_i\sum_{r=1}^k \boldsymbol{S}_{r,ii} + v_i/2\sum_{r=1}^k \boldsymbol{S}_{r,i\cdot}\boldsymbol{V}\boldsymbol{S}^T_{r,i\cdot})$ such that equality (29) holds asympotically up to the second order.

6 Simulation and Example

Simulation

The behavior of the local likelihood backfitting estimator is illustrated in a simulation experiment and a real application. We generate data from the bivariate additive logistic model

$$\text{logit}\{E(y|x)\} \quad = \quad \gamma_1(x_1) + \gamma_2(x_2) \tag{30}$$

with $\gamma_1(x_1) = x_1/2 + \exp(-4x^2)/2$ and $\gamma_2(x_2) = \cos(x\pi/2)$, where $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$ are centered in the usual way. We consider both a random and a fixed design for the x_1, x_2. For the random design case, we draw (x_1, x_2) from a truncated bivariate normal distribution with mean 0, variance 1 and correlation levels specified below, truncated such that $(x_1, x_2) \in [-1, 1]^2$. For the fixed design case, we select (x_1, x_2) as a grid of 15^2 equidistant design points on $[-1, 1]$.

In the random design, we can evaluate how close the asymptotic variance approximations $\widehat{\boldsymbol{S}}_r^+\boldsymbol{V}\widehat{\boldsymbol{S}}_r^{+^T}$ and $\boldsymbol{Q}_r\boldsymbol{V}\boldsymbol{Q}_r^T$ are relative to the real variance of the estimators for model (30). Figure 2 shows the approximation error when using $\widehat{\boldsymbol{S}}_r^+\boldsymbol{V}\widehat{\boldsymbol{S}}_r^{+^T}$ as variance compared to $\boldsymbol{Q}_r\boldsymbol{V}\boldsymbol{Q}_r^T$ for different correlation among the covariates. It appears that the simpler approximation behaves unsatisfactory for highly correlated covariates, while the more complicated one remains close to the true variance of the estimators.

For each design setting above, we draw now 200 replicates. The estimates are fitted by local linear likelihood backfitting and in each simulation the bandwidth in chosen as minimizer of the Akaike criterion (see Hastie and Tibshirani, 1990) $\max 2 \sum_i l_r(\widehat{\eta}_i) - 2df$, where $df = 1 + \sum_{r=1}^{q} \operatorname{tr}(\boldsymbol{S}_r \boldsymbol{V})$ is chosen as measure for the degree of freedom. To reduce the computational effort, however, we minimize the Akaike criterion over the 2×2 grid $(0.15, 0.3)^2$ only. In what follows, we will use the wild bootstrap for inference around the estimated additive functions, using $B = 200$ bootstrap replicates. The second bandwidth $g = (g_1, g_2)$ for bootstrapping is chosen by the rule of thumb $g_r = h_r^{(2p_r+3)/(2p_r+5)}$. Figure 3 shows a typical simulation of the random design setting with corresponding pointwise bootstrap bands. Table 1 gives the simulated coverage probabilities at selected design points for random design setting based on the 200 simulations. It appears that the bootstrap behaves rather satisfactory, though it shows slight undercoverage at the boundaries and at point $x_1 = 0$ for $\gamma_1(\cdot)$, which is at the local peak as seen from Figure 3. In contrast, slight overcoverage is seen for inner points of x_2. The same behavior is found for the fixed design case as shown in Figure 4. In general, the coverage behavior exhibited in this simulation is not surprising and is in line with that of bootstrapping in simpler nonparametric regression settings.

Example We briefly discuss a data example considering the creditworthiness of customers of a bank. The German Hypo-bank provided data of 700 sucessful $(y = 1)$ and 300 failed credits $(y = 0)$, with explanatory quantities x_1: the amount (in DM), x_2: the period of the credit (in months) and x_3: the age of the borrower (in years). Additionally we take x_4: the gender of the borrower into account which is included as parametric fit. The resulting model is

$$E(y|x_1, \ldots x_4) = h\{\gamma_1(x_1) + \gamma_2(x_2) + \gamma_3(x_3) + x_4\beta\}$$

with $h(\cdot)$ as logit link. The data are made public on the webserver `http://www.stat.uni-muenchen.de`.

Figure 5 shows the resulting local polynomial backfitting estimates in a logistic model with pointwise bootstrap confidence intervals. The bandwidth for the first three plots is chosen as $h = (3500, 30, 15)$ by an Akaike criterion, for the gender effect we fitted single effect for each category, i.e. we chose h_4 small such that the resulting weights equal 1 or 0, i.e. the gender effect is fited as factorial effect. It appears that the period has basically a linear effect while age has a quadratic type effect, showing young borrowers to be more risky in paying back a credit.

7 Discussion

In the paper we give a general discussion of asymptotic properties of backfitting estimates in generalized additive models. We propose the local likelihood backfitting estimate which shows as a numerically simpler form of the local scoring algorithm. We proof consistency for both estimates, where our theoretical arguments are based on properties of the oracle estimate. From a practical viewpoint, the results given in the paper provide the so far missing theoretical justification for variance expressions for estimates in generalized additive models based local scoring. For practioneers this means, that confidence bands as e.g. plotted by the `gam()` in Splus result from a rigorous theoretical reasoning.

Moreover a bootstrap procedure is suggested which extends available bootstrap approaches to generalized additive models.

A Proof of Theorems

Proof of Theorem 2.1:
Let $\boldsymbol{X}_{r,ij}^T\boldsymbol{\beta}_{r,i}$ be a Taylor approximation of $\gamma_r(x_{rj})$ and define $\delta_{r,ij} = \gamma_r(x_{rj}) - \boldsymbol{X}_{r,ij}^T\boldsymbol{\beta} = \gamma_r^{(p_r+1)}(x_{ri})(x_{rj} - x_{ri})^{p_r+1}/(p_r + 1)! + \ldots$ as approximation bias, where $\delta_{r,ij}{=}o\{(x_{r,j} - x_{r,i})^{p_r}\}$. Note that $\sum_j w_{r,ij}\delta_{r,ij}/\sum_k w_{r,ik} = O(h^{2\lfloor 1+p_r/2\rfloor})$. The local likelihood estimate $\widehat{\boldsymbol{\beta}}_{r,i}$ is obtained from (19) while the oracle estimate $\widehat{\boldsymbol{\beta}}_{r|-r,i}$ solves (12). To avoid confusion in the following we use brackets of the type $(\cdot)$ if we refer to parameter arguments while brackets $\{\cdot\}$ and $[\cdot]$ are used for arithmetic reasons. We define $\widehat{\eta}_{(r),ij} = \boldsymbol{X}_{r,ij}^T\widehat{\boldsymbol{\beta}}_{r,i} + \widehat{\eta}_{-r,j}$ and $\eta_{(r),ij} = \boldsymbol{X}_{r,ij}^T\boldsymbol{\beta}_{r,i} + \eta_{-r,j}$ with $\eta_{-r,j} = \alpha + \sum_{k\neq r}\gamma_k(x_{kj})$, so that $\eta_j = \alpha + \gamma_{r,j} + \gamma_{-r,j} = \eta_{(r),ij} + \delta_{r,ij}$. Expansion of the local estimating equation (19) about the oracle estimate $\widehat{\boldsymbol{\beta}}_{r|-r,i}$ and the true parameters $\eta_{-r,j}$ yields

$$\begin{aligned}
\mathbf{0} &= \sum_{j=1}^{n} w_{r,ij}\boldsymbol{X}_{r,ij} l_{\eta,j}(\boldsymbol{X}_{r,ij}^T\widehat{\boldsymbol{\beta}}_{r,i} + \widehat{\eta}_{-r,j}) \\
&= \sum_{j=1}^{n} w_{r,ij}\boldsymbol{X}_{r,ij} l_{\eta,j}(\boldsymbol{X}_{r,ij}^T\widehat{\boldsymbol{\beta}}_{r|-r,i} + \eta_{-r,j}) \\
&\quad -\sum_{j=1}^{n} w_{r,ij}\boldsymbol{X}_{r,ij} v_j(\boldsymbol{X}_{r,ij}^T\widehat{\boldsymbol{\beta}}_{r|-r,i} + \eta_{-r,j})[\boldsymbol{X}_{r,ij}^T\{\widehat{\boldsymbol{\beta}}_{r|-r,i} - \widehat{\boldsymbol{\beta}}_{r,i}\} - \{\widehat{\eta}_{-r,j} - \eta_{-r,j}\}]
\end{aligned}$$

$$\begin{aligned}
&-\frac{1}{2}\sum_{j=1}^{n} w_{r,ij}\boldsymbol{X}_{r,ij}v_j'(\boldsymbol{X}_{r,ij}^T\widehat{\boldsymbol{\beta}}_{r|-r,i}+\eta_{-r,j})[\boldsymbol{X}_{r,ij}^T\{\widehat{\boldsymbol{\beta}}_{r|-r,i}-\widehat{\boldsymbol{\beta}}_{r,i}\}-\{\widehat{\eta}_{-r,j}-\eta_{-r,j}\}]^2+\dots \\
=\;& \mathbf{0}+\boldsymbol{F}_{r,i}\{\widehat{\boldsymbol{\beta}}_{r|-r,i}-\widehat{\boldsymbol{\beta}}_{r,i}\}-\sum_{j=1}^{n} w_{r,ij}\boldsymbol{X}_{r,ij}v_j\{\widehat{\eta}_{-r,j}-\eta_{-r,j}\} \qquad (31)\\
&+\sum_{j=1}^{n} w_{r,ij}\boldsymbol{X}_{r,ij}v_j'[\boldsymbol{X}_{r,ij}^T\{\widehat{\boldsymbol{\beta}}_{r|-r,i}-\boldsymbol{\beta}_{r,i}\}-\delta_{r,ij}][\boldsymbol{X}_{r,ij}^T\{\widehat{\boldsymbol{\beta}}_{r|-r,i}-\boldsymbol{\beta}_{r,i}\}-(\widehat{\eta}_{(r),ij}-\eta_{(r),jj})] \\
&-\frac{1}{2}\sum_{j=1}^{n} w_{r,ij}\boldsymbol{X}_{r,ij}v_j'[\boldsymbol{X}_{r,ij}^T\{\widehat{\boldsymbol{\beta}}_{r|-r,i}-\boldsymbol{\beta}_{r,i}\}-(\widehat{\eta}_{(r),ij}-\eta_{(r),jj})]^2+\dots.
\end{aligned}$$

We now group terms in (31) to obtain a polynomial equation in $\widehat{\boldsymbol{\beta}}_{r|-r,i}-\boldsymbol{\beta}_{r|i}$. We define the p_r dimensional vector

$$\begin{aligned}
\mathbf{A}_{r,i} \;:=\;& -\boldsymbol{F}_{r,i}\{\widehat{\boldsymbol{\beta}}_{r,i}-\boldsymbol{\beta}_{r,i}\}-\sum_{j=1}^{n} w_{r,ij}\boldsymbol{X}_{r,ij}v_j\{\widehat{\eta}_{-r,j}-\eta_{-r,j}\} \\
&+\sum_{j=1}^{n} w_{r,ij}v_j'\boldsymbol{X}_{r,ij}\delta_{r,ij}\{\widehat{\eta}_{(r),ij}-\eta_{(r),ij}\} \\
&+\frac{1}{2}\sum_{j=1}^{n} w_{r,ij}v_j'\boldsymbol{X}_{r,ij}\{\widehat{\eta}_{(r),ij}-\eta_{(r),ij}\}^2,
\end{aligned}$$

the $p_r \times p_r$ dimensional matrix

$$\mathbf{B}_{r,i} \;:=\; \boldsymbol{F}_{r,i}+\sum_{j=1}^{n} w_{r,ij}v_j'\delta_{r,ij}\boldsymbol{X}_{r,ij}\boldsymbol{X}_{r,ij}^T$$

and the $p_r \times p_r \times p_r$ dimensional array

$$\mathbf{C}_{r,i,(stu)} \;:=\; \frac{1}{2}\sum_{j=1}^{n} w_{r,ij}v_j'\boldsymbol{X}_{r,ij,(s)}\boldsymbol{X}_{r,ij,(t)}\boldsymbol{X}_{r,ij,(u)},$$

where the bracketed subscripts here and in the following indicate the element of the array with $0 \le s,t,u \le p_r$, e.g. $\boldsymbol{X}_{r,ij,(s)}=(x_{ri}-x_{rj})^{s-1}$. This allows us to write the sth element of (31) as

$$\begin{aligned}
0 \;=\;& \mathbf{A}_{r,i,(s)}+\sum_t \mathbf{B}_{r,i,(st)}\{\widehat{\boldsymbol{\beta}}_{r|-r,i}^{(t)}-\boldsymbol{\beta}_{r,i}^{(t)}\} \qquad (32)\\
&+\sum_{t,u}\mathbf{C}_{r,i,(stu)}\{\widehat{\boldsymbol{\beta}}_{r|-r,i}^{(t)}-\boldsymbol{\beta}_{r,i}^{(t)}\}\{\widehat{\boldsymbol{\beta}}_{r|-r,i}^{(s)}-\boldsymbol{\beta}_{r,i}^{(s)}\}+\dots
\end{aligned}$$

where superscript $\widehat{\boldsymbol{\beta}}_{r|-r,i}^{(t)}$ refers to the tth element of $\widehat{\boldsymbol{\beta}}_{r|-r,i}$. We show now that

$$\mathbf{B}_{r,i,(st)}=\boldsymbol{F}_{r,i,(st)}\{1+O(h_r^{2\lfloor 1+p_r/2\rfloor})\}. \qquad (33)$$

Considering $\boldsymbol{F}_{r,i}$ one finds

$$\boldsymbol{F}_{r,i,(st)} = \begin{cases} O(nh_r^{s+t-1}) & \text{for } s+t \text{ even} \\ O(nh_r^{s+t}) & \text{for } s+t \text{ odd} \end{cases} \tag{34}$$

For the second component in $\mathbf{B}_{r,i}$ one obtains

$$\begin{aligned} & [\sum_{j=1}^{n} w_{r,ij}\boldsymbol{X}_{r,ij}v_i'\delta_{r,ij}\boldsymbol{X}_{r,ij}^T]_{(st)} \\ = \; & n\int K\left(\frac{x_{ri}-x_r}{h_r}\right)(x_{ri}-x_r)^{s+t+p_r-1}v_r'(x_r)\{\frac{1}{(p_r+1)!}\gamma_r^{(p_r+1)}(x_{ri})+\ldots\}f_r(x_r)dx_r \\ = \; & nh_r^{s+t+p_r}\int K(z)z^{s+t+p_r-1}v_r'(x_{r,i}-zh)\{\frac{1}{(p+1)!}\gamma_r^{(p+1)}(x_{r,i})+\ldots\}f_r(x_{ri}-zh_r)dz \\ = \; & \begin{cases} O(nh_r^{s+t+p_r}) & \text{for } s+t+p_r \text{ odd} \\ O(nh_r^{s+t+p_r+1}) & \text{for } s+t+p_r \text{ even} \end{cases} \end{aligned}$$

where $v_r'(x_r) = \int v'(\gamma_r(x_{r,i}) + \eta_{-r}(x_{-r}) + \alpha)f(x_{-r}|x_r)dx_{-r}$. Comparing these quantities with (34) immediately shows (33). In the same fashion one can show that all components involving $\delta_{r,ij}$ are of negligible asymptotic order. Defining $\mathbf{B}_{r,i}^{(st)}$ as the (s,t)th element of the inverse of $\mathbf{B}_{r,i}$, we find by standard series inversion of (32) (see e.g. McCullagh, 1987[chapter 7])

$$\begin{aligned} \{\widehat{\boldsymbol{\beta}}_{r|-r,i}^{(s)} - \boldsymbol{\beta}_{r,i}^{(s)}\} &= -\sum_t \mathbf{B}_{r,i}^{(st)}\mathbf{A}_{r,i,(t)} \\ &\quad - \sum_{t,u,v,w,z} \mathbf{B}_{r,i}^{(st)}\mathbf{B}_{r,i}^{(uv)}\mathbf{B}_{r,i}^{(wz)}\mathbf{C}_{r,i,(tuw)}\mathbf{A}_{r,i,(v)}\mathbf{A}_{r,i,(z)} + \ldots. \end{aligned} \tag{35}$$

We are interested in the approximation for $\widehat{\gamma}_{r|-r,i} - \gamma_{r,i} = \widehat{\boldsymbol{\beta}}_{r|-r,i}^{(1)} - \boldsymbol{\beta}_{r,i}^{(1)}$. Using (33) and the definition of $\mathbf{A}_{r,i}$ we find that the first component in (35) decomposes to

$$\begin{aligned} -\sum_t \boldsymbol{B}_{r,i}^{(1t)}\mathbf{A}_{r,i,(t)} &= \left[\widehat{\gamma}_{r,i} - \gamma_{r,i} + \mathbf{e}_1^T\boldsymbol{F}_{r,i}^{-1}\sum_{j=1}^n w_{r,ij}\boldsymbol{X}_{r,ij}v_j\{\widehat{\eta}_{-r,j} - \eta_{-r,j}\}\right] & (36) \\ &\quad \times\{1 + O(h_r^{2[1+p_r/2]})\} \\ &\quad + \boldsymbol{e}_1^T\boldsymbol{F}_{r,i}^{-1}\Big[\sum_{j=1}^n w_{r,ij}\boldsymbol{X}_{r,ij}v_j'\delta_{r,ij}O(\widehat{\eta}_{(r),ij} - \eta_{(r),ij}) & (37) \\ &\qquad + \sum_{j=1}^n w_{r,ij}\boldsymbol{X}_{r,ij}v_j'O(\{\widehat{\eta}_{(r),ij} - \eta_{(r),ij}\}^2) + \ldots\Big]. & (38) \end{aligned}$$

The components in (36) are equal to $\widehat{\gamma}_{r,i} - \gamma_{r,i} + \boldsymbol{S}_{r,i\cdot}\boldsymbol{V}\{\widehat{\boldsymbol{\eta}}_{-r} - \boldsymbol{\eta}_{-r}\}$, where $\boldsymbol{S}_{r,i\cdot}$ denotes the i-th row of $\boldsymbol{S}_r$. For the centered estimate $\widehat{\boldsymbol{\gamma}}_r$ the smoothing matrix $\boldsymbol{S}_r$

is replaced by the centered version $\boldsymbol{S}_r^+$. Since (37) involves the approximation bias $\delta_{r,ij}$, this component is of negligible order, using similar arguments as applied above. Finally, (38) results from quadratic terms, e.g. from the latter component in $\boldsymbol{A}_{r,i}$ one gets

$$\begin{aligned} & \boldsymbol{e}_1 \boldsymbol{F}_{r,i}^{-1} \sum_{j=1}^{n} w_{r,ij} v'_j \boldsymbol{X}_{r,ij} \left[\boldsymbol{X}_{r,ij}^T \{ \widehat{\boldsymbol{\beta}}_{r,i} - \boldsymbol{\beta}_{r,i} \} + \{ \widehat{\eta}_{-r,j} - \eta_{-r,j} \} \right]^2 \qquad (39) \\ = \; & O\left(\{ \widehat{\gamma}_{r,i} - \gamma_{r,i} \}^2 + \boldsymbol{S}_{r,i \cdot} \boldsymbol{V}' \{ \widehat{\boldsymbol{\eta}}_{-r} - \boldsymbol{\eta}_{-r} \} \{ \widehat{\gamma}_{r,i} - \gamma_{r,i} \} + \boldsymbol{S}_{r,i \cdot} \boldsymbol{V}' \{ \widehat{\boldsymbol{\eta}}_{-r} - \boldsymbol{\eta}_{-r} \}^2 \right). \end{aligned}$$

where $\boldsymbol{V}' = \mathrm{diag}(v'_1, \ldots, v'_n)$. From A4 we get $v'_i = O(v_i)$ for $i = 1, \ldots n$ and it is easily seen that $O(\boldsymbol{S}_{r,i \cdot} \boldsymbol{V}' \{ \widehat{\boldsymbol{\eta}}_{-r} - \boldsymbol{\eta}_{-r} \}^2]) = O([\boldsymbol{S}_{r,i \cdot} \boldsymbol{V} \{ \widehat{\boldsymbol{\eta}}_{-r} - \boldsymbol{\eta}_{-r} \}]^2)$. Similar terms as in (39) result from the second component in (35), which together with (39) are contained in the correction components in (20).

The components containing quadratic terms can jointly be written as a quadratic form $\sum_{j,k=1}^{nq} \boldsymbol{D}_{r,i}^{(jk)} (\widehat{\boldsymbol{\gamma}}_{\bullet,(j)} - \boldsymbol{\gamma}_{\bullet,(j)})(\widehat{\boldsymbol{\gamma}}_{\bullet,(k)} - \boldsymbol{\gamma}_{\bullet,(k)})$ with indices j and k refering to the corresponding element of $\boldsymbol{\gamma}_{\bullet}$ and $\widehat{\boldsymbol{\gamma}}_{\bullet}$, respectively. Here $\boldsymbol{D}_{r,i}^{(jk)}$ is an array of dimension $(nq) \times (nq)$. Combining this, we can write (36) jointly for all $i = 1, \ldots, n$ and $r = 1, \ldots, q$ as matrix form

$$\widehat{\boldsymbol{\gamma}}_{\bullet|-\bullet} - \boldsymbol{\gamma}_{\bullet} \;=\; M(\widehat{\boldsymbol{\gamma}}_{\bullet} - \boldsymbol{\gamma}_{\bullet}) + \sum_{j,k=1}^{n} \boldsymbol{D}^{(jk)} (\widehat{\boldsymbol{\gamma}}_{\bullet,(j)} - \boldsymbol{\gamma}_{\bullet,(j)})(\widehat{\boldsymbol{\gamma}}_{\bullet,(k)} - \boldsymbol{\gamma}_{\bullet,(k)}) \qquad (40)$$

with $\boldsymbol{D}$ as $(nq)^3$ dimensional array build from $\boldsymbol{M}$ and $\boldsymbol{D}_{r,i}$ so that (20) follows. If $\boldsymbol{M}$ is invertible, we can invert (40) and find

$$\widehat{\boldsymbol{\gamma}}_{\bullet} - \boldsymbol{\gamma}_{\bullet} \;=\; M^{-1}(\widehat{\boldsymbol{\gamma}}_{\bullet|-\bullet} - \boldsymbol{\gamma}_{\bullet}) + \sum_{j,k=1}^{n} \widetilde{\boldsymbol{D}}^{(jk)} (\widehat{\boldsymbol{\gamma}}_{\bullet|-\bullet,(j)} - \boldsymbol{\gamma}_{\bullet,(j)})(\widehat{\boldsymbol{\gamma}}_{\bullet|-\bullet,(k)} - \boldsymbol{\gamma}_{\bullet,(k)})$$

with $\widetilde{\boldsymbol{D}}^{(jk)}$ as $(nq)^3$ dimensional array build from $\boldsymbol{M}$ and $\boldsymbol{D}$. Since $\widehat{\boldsymbol{\gamma}}_{\bullet|-\bullet}$ is consistent by Lemma A.1 below, the second part of Theorem 2.1 follows.

■

Lemma A.1 *Under assumptions A1–A5, the conditional bias of the centered oracle estimator* $\widehat{\gamma}_{r|-r}(x_{ri})$ *is approximated by*

$$\begin{aligned} & E(\widehat{\gamma}_{r|-r}(x_{ri}) - \gamma_{r|-r}(x_{ri}) | \boldsymbol{X}_1, \boldsymbol{X}_2) = \\ & \quad h_r^{p_r+1} \frac{\mu_{p_r+1}(K)}{(p_r+1)!} \left(\gamma_r^{(p_r+1)}(x_{ri}) - \frac{E_x(v(\eta)\gamma_r^{(p_r+1)}(X_r))}{E_x(v(\eta))} \right) + o_p(h_r^{p_r+1}). \qquad (41) \end{aligned}$$

The conditional variance of $\widehat{\gamma}_{r|-r}(x_{ri})$ *is approximated by*

$$Var(\widehat{\gamma}_{r|-r}(x_{ri})|\boldsymbol{X}_1, \boldsymbol{X}_2) = \frac{1}{nh_r} R(K) v_r(x_{ri})^{-1} f_r(x_{ri})^{-1}(1 + o_p(1))$$

The conditional covariance of $\widehat{\gamma}_{r|-r}(x_{ri})$ *and* $\widehat{\gamma}_{r|-r}(x_{rj})$*,* $i \neq j$*, is approximated by*

$$Cov(\widehat{\gamma}_{r|-r}(x_{ri}), \widehat{\gamma}_{r|-r}(x_{rj})|\boldsymbol{X}_1, \boldsymbol{X}_2) = \frac{1}{nh_r} K{*}K \left(\frac{x_{rj} - x_{ri}}{h_r}\right) v_r(x_{ri})^{-1} f_r(x_{ri})^{-1}(1{+}o_p(1)),$$

where $K * K$ *denotes the convolution of* K *with itself. Finally, the conditional covariance of* $\widehat{\gamma}_{r|-r}(x_{ri})$ *and* $\widehat{\gamma}_{s|-s}(x_{sj})$*,* $r \neq s$*, is approximated by*

$$Cov(\widehat{\gamma}_{r|-r}(x_{ri}), \widehat{\gamma}_{s|-s}(x_{sj})|\boldsymbol{X}_1, \boldsymbol{X}_2) = \frac{1}{n} \frac{v_{rs}(x_{ri}, x_{sj}) f_{rs}(x_{ri}, x_{sj})}{v_r(x_{ri}) v_s(x_{sj}) f_r(x_{ri}) f_s(x_{sj})} (1 + o_p(1)),$$

where f_{rs} *denotes the bivariate marginal density for* (X_r, X_s) *and* $v_{rs}(x_r, x_s) = E_x(v(\eta)|X_r = x_r, X_s = x_s)$*.*

Proof of Lemma A.1:
The approximations follow from (15) and the asymptotic results in Opsomer and Kauermann (2000) for weighted local polynomial regression.

■

Lemma A.2 *Under assumptions A2'–A6',*

$$\Pr\{\textit{there exists } N \textit{ such that } \boldsymbol{M} \textit{ is invertible for all} \, n \geq N\} = 1$$

and

$$\begin{aligned}(\boldsymbol{I} - \boldsymbol{S}_1^+ \boldsymbol{V} \boldsymbol{S}_2^+ \boldsymbol{V})^{-1} &= (\boldsymbol{I} - \boldsymbol{T}_{12}^*)^{-1} + o(\boldsymbol{1}\boldsymbol{1}^T/n) \;\; \textit{a.s.} \\ &= \boldsymbol{I} + O(\boldsymbol{1}\boldsymbol{1}^T/n) \;\; \textit{a.s.}\end{aligned}$$

Proof of Lemma A.2:
This lemma is a direct generalization of Lemmas 3.1 and 3.2 in Opsomer and Ruppert (1997) for the weighted local polynomial smoothers discussed in Opsomer and Kauermann (2000). $\boldsymbol{M}$ is invertible if $(\boldsymbol{I} - \boldsymbol{S}_1^+ \boldsymbol{V} \boldsymbol{S}_2^+ \boldsymbol{V})^{-1}$ exists. The approximation

$$\boldsymbol{S}_1^+ \boldsymbol{V} \boldsymbol{S}_2^+ \boldsymbol{V} = \boldsymbol{T}_{12}^* + o(\boldsymbol{1}\boldsymbol{1}^T/n) \;\; \text{a.s.} \tag{42}$$

follows from assumptions A2'–A5', using standard kernel moment approximations and the uniform convergence theory from Chapter 2 of Pollard (1984). Assumption A6' and (42) then lead to the invertability of $\boldsymbol{M}$.

Assume now that $(\boldsymbol{I} - \boldsymbol{S}_1^{+}\boldsymbol{V}\boldsymbol{S}_2^{+}\boldsymbol{V})^{-1}$ exists. As in the proof of Lemma 3.2 in Opsomer and Ruppert (1997), the two approximations from the lemma follow if we can show that $\sum_{k=1}^{\infty}\boldsymbol{T}_{12}^{*k} = O(11^T/n)$ a.s. Using Corollary 5.6.13 in Horn and Johnson (1985) and assumption A6', we know that there exist $\epsilon > 0$ and $C > 0$, such that

$$\max_{i,j} |[\boldsymbol{T}_{12}^{*k}]_{ij}| \leq \frac{C}{n}(1-\epsilon)^k.$$

Hence, $\max_{i,j} |\sum_{k=1}^{\infty}[\boldsymbol{T}_{12}^{*k}]_{ij}| \leq K/n$ as desired.

■

Proof of Theorem 3.1:
For $q = 2$, the matrix $\boldsymbol{M}^{-1}$ exists for sufficiently large n and can be approximated by

$$\boldsymbol{M}^{-1} = \begin{bmatrix} (\boldsymbol{I} - \boldsymbol{T}_{12}^{*})^{-1} & -(\boldsymbol{I} - \boldsymbol{T}_{12}^{*})^{-1}\boldsymbol{S}_1^{+}\boldsymbol{V} \\ -(\boldsymbol{I} - \boldsymbol{T}_{12}^{*})^{-1}\boldsymbol{S}_2^{+}\boldsymbol{V} & (\boldsymbol{I} - \boldsymbol{T}_{12}^{*})^{-1} \end{bmatrix} (1 + o(1/n)) \text{ a.s.},$$

using the formula for the inverse of a partitioned matrix (Horn and Johnson, 1985, p.18) and Lemma A.2. Now, using Lemma A.1,

$$\begin{aligned}&(\boldsymbol{I} - \boldsymbol{T}_{12}^{*})^{-1}\mathrm{E}(\widehat{\boldsymbol{\gamma}}_{1|-1} - \boldsymbol{\gamma}_1|\boldsymbol{X}_1, \boldsymbol{X}_2) = \\ &\qquad h_1^{p_1+1}\frac{\mu_{p_1+1}(K)}{(p_1+1)!}\left((\boldsymbol{I} - \boldsymbol{T}_{12}^{*})^{-1}\boldsymbol{\gamma}_1^{(p_1+1)} - \frac{\mathrm{E}_x\{v(\eta)\gamma_1^{(p_1+1)}(x_1)\}}{\mathrm{E}_x\{v(\eta)\}}\right) + o_p(h_1^{p_1+1})\end{aligned}$$

since $\boldsymbol{T}_{12}^{*}\mathbf{1} = \mathbf{0}$. Similarly,

$$\begin{aligned}&(\boldsymbol{I} - \boldsymbol{T}_{12}^{*})^{-1}\boldsymbol{S}_1^{+}\boldsymbol{V}\mathrm{E}(\widehat{\boldsymbol{\gamma}}_{2|-2} - \boldsymbol{\gamma}_2|\boldsymbol{X}_1, \boldsymbol{X}_2) = \\ &\qquad h_2^{p_2+1}\frac{\mu_{p_2+1}(K)}{(p_2+1)!}\left((\boldsymbol{I} - \boldsymbol{T}_{12}^{*})^{-1}\mathrm{E}\{\gamma_2^{(p_2+1)}(x_2)|\boldsymbol{X}_1\} - \frac{\mathrm{E}_x\{v(\eta)\gamma_2^{(p_2+1)}(x_2)\}}{\mathrm{E}_x\{v(\eta)\}}\right) + o_p(h_2^{p_2+1}),\end{aligned}$$

leading directly to the desired bias approximation.

The variance-covariance matrix of $\widehat{\boldsymbol{\gamma}}_1$ is

$$\begin{aligned}\mathrm{Var}(\widehat{\boldsymbol{\gamma}}_1|\boldsymbol{X}_1, \boldsymbol{X}_2) &= (\boldsymbol{I} - \boldsymbol{T}_{12}^{*})^{-1} \times \Big\{\mathrm{Var}(\widehat{\boldsymbol{\gamma}}_{1|-1}) - \boldsymbol{S}_1^{+}\boldsymbol{V}\mathrm{Cov}(\widehat{\boldsymbol{\gamma}}_{1|-1}, \widehat{\boldsymbol{\gamma}}_{2|-2}) \\ &\quad -\mathrm{Cov}(\widehat{\boldsymbol{\gamma}}_{1|-1}, \widehat{\boldsymbol{\gamma}}_{2|-2})\boldsymbol{V}\boldsymbol{S}_1^{+T} + \boldsymbol{S}_1^{+}\boldsymbol{V}\mathrm{Var}(\widehat{\boldsymbol{\gamma}}_{2|-2})\boldsymbol{V}\boldsymbol{S}_1^{+T}\Big\} \times (\boldsymbol{I} - \boldsymbol{T}_{12}^{*})^{-T}(1 + o_p(1)).\end{aligned}$$

Using Lemmas A.1 and A.2 and standard kernel moment approximations, it can be shown that the terms involving $\mathrm{Cov}(\widehat{\boldsymbol{\gamma}}_{1|-1}, \widehat{\boldsymbol{\gamma}}_{2|-2})$ and $\mathrm{Var}(\widehat{\boldsymbol{\gamma}}_{2|-2})$ are $o_p(1/nh_1)$, and that

$$\mathrm{Var}(\widehat{\boldsymbol{\gamma}}_1 | \boldsymbol{X}_1, \boldsymbol{X}_2) = \mathrm{Var}(\widehat{\boldsymbol{\gamma}}_{1|-1})(1 + o_p(1)),$$

proving the variance approximation in the theorem.

■

Proof of Theorem 4.1:

As above we use brackets $(\cdot)$ to refer to parameter arguments while brackets $\{\cdot\}$ and $[\cdot]$ are used for arithmetic reasons. Moreover we use the hat notations to refer to plug-in estimates. Taylor expansion and simple calculation allows to derive from (22) at convergence

$$\begin{aligned}
0 &= \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{l}_{\eta,j}^{(\infty)} + \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j^{(\infty)} (\widehat{\gamma}_{r,j}^{(\infty)} - \boldsymbol{X}_{r,ij}^T \widehat{\boldsymbol{\beta}}_{r,i}^{(\infty)}) \\
&= \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{l}_{\eta,j} - \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j (\widehat{\gamma}_{r,j}^{(\infty)} - \widehat{\gamma}_{r,j} + \widehat{\eta}_{-r,j}^{(\infty)} - \widehat{\eta}_{-r,j}) \\
&\quad - \frac{1}{2} \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j' (\widehat{\gamma}_{r,j}^{(\infty)} - \widehat{\gamma}_{r,j} + \widehat{\eta}_{-r,j}^{(\infty)} - \widehat{\eta}_{-r,j})^2 \\
&\quad + \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j (\widehat{\gamma}_{r,j}^{(\infty)} - \boldsymbol{X}_{r,ij}^T \widehat{\boldsymbol{\beta}}_{r,i}^{(\infty)}) \\
&\quad + \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j' (\widehat{\gamma}_{r,j}^{(\infty)} - \boldsymbol{X}_{r,i}^T \widehat{\boldsymbol{\beta}}_{r,i}^{(\infty)}) (\widehat{\gamma}_{r,j}^{(\infty)} - \widehat{\gamma}_{r,j} + \widehat{\eta}_{-r,j}^{(\infty)} - \widehat{\eta}_{-r,j}) + \dots \\
\Leftrightarrow 0 &= \widehat{\gamma}_{r,i} - \widehat{\gamma}_{r,i}^{(\infty)} + \boldsymbol{e}_1 \boldsymbol{F}^{-1} \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j (\widehat{\eta}_{-r,j} - \widehat{\eta}_{-r,j}^{(\infty)}) \\
&\quad - \frac{1}{2} \boldsymbol{e}_1 \boldsymbol{F}^{-1} \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j' \{\boldsymbol{X}_{r,ij} (\widehat{\boldsymbol{\beta}}_{r,i} - \widehat{\boldsymbol{\beta}}_{r,i}^{(\infty)}) + \widehat{\eta}_{-r,j} - \widehat{\eta}_{-r,j}^{(\infty)}\}^2 \\
&\quad + \boldsymbol{e}_1 \boldsymbol{F}^{-1} \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j' \widehat{\delta}_{r,ij} \{\boldsymbol{X}_{r,ij} (\widehat{\boldsymbol{\beta}}_{r,i} - \widehat{\boldsymbol{\beta}}_{r,i}^{(\infty)}) + \widehat{\eta}_{-r,j} - \widehat{\eta}_{-r,j}^{(\infty)}\} \qquad (43) \\
&\quad + \frac{1}{2} \boldsymbol{e}_1 \boldsymbol{F}^{-1} \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j' \widehat{\delta}_{r,ij}^{(\infty)^2} + \dots . \qquad (44)
\end{aligned}$$

where $\widehat{\delta}_{r,ij} = \widehat{\gamma}_{r,j} - \boldsymbol{X}_{r,ij} \widehat{\beta}_{r,ij}$ and analogous definition for $\widehat{\delta}_{r,ij}^{(\infty)}$. Since γ_r is assumed to be sufficiently smooth the component (43) are of negligible order. This follows since $\widehat{\delta}_{r,ij} \to \delta_{r,ij}$ and as shown above, components including $\delta_{r,ij}$ vanish with order

$O(h^2)$. Finally, (44) is decomposed to

$$
\begin{aligned}
& \boldsymbol{e}_1 \boldsymbol{F}^{-1} \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j' \widehat{\delta}_{r,ij}^{(\infty)^2} \\
= \; & \boldsymbol{e}_1 \boldsymbol{F}^{-1} \{ \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j' \widehat{\delta}_{r,ij}^2 + \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j' \widehat{\delta}_{r,ij} \{ \widehat{\gamma}_{r,j}^{(\infty)} - \widehat{\gamma}_{r,j} + \boldsymbol{X}_{r,ij} (\widehat{\boldsymbol{\beta}}_{r,i}^{(\infty)} - \widehat{\boldsymbol{\beta}}_{r,i}) \} \\
& + \sum_{j=1}^{n} w_{r,ij} \boldsymbol{X}_{r,ij} \widehat{v}_j' \{ \widehat{\gamma}_{r,j}^{(\infty)} - \widehat{\gamma}_{r,j} + \boldsymbol{X}_{r,ij} (\widehat{\boldsymbol{\beta}}_{r,i}^{(\infty)} - \widehat{\boldsymbol{\beta}}_{r,i}) \}^2 .
\end{aligned}
$$

The first two terms are of negligible order using the same arguments as above. The last component builds the extra correction term given in (23).

■

Proof of Theorem 5.1:
As will be shown below, the bootstrap (27) is directly obtained from the "oracle bootstrap"

$$
\widehat{\boldsymbol{\gamma}}_{r|-r}^{*} \;=\; \widehat{\boldsymbol{S}}_{r,\{h\}} \widehat{\boldsymbol{l}}_{\eta,\{h\}}^{*} + \widehat{\boldsymbol{S}}_{r,\{h\}} \widehat{\boldsymbol{V}} \widehat{\boldsymbol{\gamma}}_{r,\{g\}} . \tag{45}
$$

Note that (45) results from applying the bootstrap principle to (15), i.e. replacing the parameters by estimates, and the estimates by bootstrap replicates. We show first that the oracle bootstrap converges in distribution to $\widehat{\boldsymbol{\gamma}}_{r|-r,\{h\}}$. Let $G_{r,i}^{*}(\cdot)$ denote the distribution function of $(nh_r)^{1/2}\{\widehat{\gamma}_{r|-r,\{h\}}^{*}(x_{ir}) - \widehat{\gamma}_{r,\{g\}}(x_{ir})\}$ and $G_{r,i}(\cdot)$ be the distribution function of $(nh_r)^{1/2}\{\widehat{\gamma}_{r|-r,\{h_r\}}(x_{ir}) - \gamma_r(x_{ir})\}$. We show by expansion of $G_{r,i}^{*}(\cdot)$ about $G_{r,i}(\cdot)$ that

$$
G_{r,i}^{*}(\cdot) \to G_{r,i}(\cdot) \tag{46}
$$

for $n \to \infty$. Let $\widehat{\kappa}_l^{*}$, $l = 1, 2, \ldots$ denote the cumulants of $G_{r,i}^{*}(\cdot)$. For $\widehat{\kappa}_1^{*}$ we find with the definition of the wild bootstrap and by using (20) and (41), assuming p_r to be odd,

$$
\begin{aligned}
\widehat{\kappa}_1^{*} \;&=\; (nh_r)^{1/2} \{ \widehat{\boldsymbol{S}}_{r,\{h_r\},i\cdot}^{+} \widehat{\boldsymbol{V}} \widehat{\boldsymbol{\gamma}}_{r,\{g\}} - \widehat{\boldsymbol{\gamma}}_{r,\{g\}} \} \{1 + o_p(1)\} \\
&=\; (nh_r)^{1/2} h_r^{p_r+1} \widehat{\gamma}_{r,\{g\}}^{(p_r+1)}(x_{ri}) \, \frac{\mu_{p_r+1}(K)}{(p_r+1)!} \{1 + o_p(1)\}
\end{aligned}
$$

where $\widehat{\gamma}_{r,\{g\}}^{(p_r+1)}(\cdot)$ is the $p_r + 1$-th order derivative of the estimated function $\widehat{\boldsymbol{\gamma}}_{r,\{g\}}(\cdot)$, and $\boldsymbol{S}_{r,\{h_r\},i\cdot}$ is the ith row of the smoothing matrix. Moreover, the second order cumulant results by

$$
\widehat{\kappa}_2^{*} \;=\; nh_r \widehat{\boldsymbol{S}}_{r,\{h_r\},i\cdot} \mathrm{diag}(\widehat{\boldsymbol{l}}_{\eta,\{h\}}^{2}) \widehat{\boldsymbol{S}}_{r,\{h_r\},i\cdot}^{T} \{1 + o_p(1)\}.
$$

and in close analogy we obtain higher order cumulants, where it is not difficult to check that the third order cumulant has negligible order $O_p(n^{-1/2}h_r^{-1/2})$. With κ_l we denote the cumulants of $G_{r,i}(\cdot)$ which are found as

$$\begin{aligned}
\kappa_1 &= (nh_r)^{1/2}h_r^{p_r+1}\gamma_r^{(p_r+1)}(x_{ri})\,\frac{\mu_{p_r+1}(K)}{(p_r+1)!}\{1+o_p(1)\} \\
\kappa_2 &= nh_r\boldsymbol{S}_{r,\{h_r\},i\cdot}\boldsymbol{V}\boldsymbol{S}_{r,\{h_r\},i\cdot}^T\{1+o_p(1)\},
\end{aligned}$$

and $\kappa_3 = O(n^{-1/2}h^{-1/2})$. Finally, we define $\widehat{\delta}_1 := \widehat{\kappa}_1^* - \kappa_1$ and $\widehat{\delta}_2 = \widehat{\kappa}_2^* - \kappa_2 + \widehat{\delta}_1\widehat{\delta}_1$ and $\widehat{\delta}_3 = \widehat{\kappa}_3^* - \kappa_3 + 3\widehat{\delta}_2\widehat{\delta}_1 + \widehat{\delta}_1^3$ and similar definitions for higher order terms (see McCullagh (1987, page 144) for more details). Considering $\widehat{\delta}_1$ we find

$$\widehat{\delta}_1 \;=\; O(n^{1/2}h_r^{p_r+3/2})O\{\widehat{\gamma}_{r,\{g\}}^{(p_r+1)}(x_{ir}) - \gamma_r^{(p_r+1)}(x_{ir})\}.$$

Taking p_r as odd with $h_r = O(n^{-1/(2p_r+3)})$ as optimal bandwidth (see Fan and Gijbels, 1996), we find $\widehat{\delta}_1 = o_p(1)$ if $\widehat{\gamma}_{r,\{g\}}^{(p_r+1)}(x_{ir}) - \gamma_r^{(p_r+1)}(x_{ir}) = o_p(1)$. This means $\widehat{\delta}_1$ vanishes asymptotically if the (p_r+1)-th order derivative is estimated consistently. This holds if g tends slower to zero than h (see e.g. Fan and Gijbels (1996) or Gasser and Müller (1984)). Similarly, for $\widehat{\delta}_2$ we find by standard arguments

$$\begin{aligned}
\widehat{\delta}_2 &= nh_r\boldsymbol{S}_{r,\{h_r\},i\cdot}\{\mathrm{diag}(\widehat{\boldsymbol{l}}_\eta^2) - \boldsymbol{V}\}\boldsymbol{S}_{r,\{h_r\},i\cdot}^T + \widehat{\delta}_1\widehat{\delta}_1 \\
&= O_p(n^{-1/2}h_r^{-1/2}) + O_p(\widehat{\delta}_1^2)
\end{aligned}$$

which is $o_p(1)$ if $\widehat{\delta}_1$ vanishes. Finally, $\widehat{\delta}_3 = O_p(n^{-1}h^{-1}) + O_p(n^{-1/2}h^{-1/2})O(\widehat{\delta}_1) + O_p(\widehat{\delta}_1^3)$. Making now use of an Edgeworth series as given for instance in (McCullagh, 1987, chapter 5) (see also (Davis, 1976)) we find

$$G_{r,i}^*(\cdot) \;=\; G_{r,i}(\cdot) + \sum_l G_{r,i}^{(l)}(\cdot)\widehat{\delta}_l/l! \tag{47}$$

where $G_{r,i}^{(l)}(\cdot)$ denotes the l-th derivative of $G_{r,i}(\cdot)$. The second component in (47) is $o_p(1)$, as shown above, which in turn proves (46).

Using (20), we defines the bootstrap equation

$$\widehat{\boldsymbol{\gamma}}_{\bullet|-\bullet,\{h\}}^* - \widehat{\boldsymbol{\gamma}}_{\bullet,\{g\}} \;=\; \widehat{\boldsymbol{M}}_{\{h\}}(\widehat{\boldsymbol{\gamma}}_{\bullet,\{h\}}^* - \widehat{\boldsymbol{\gamma}}_{\bullet,\{g\}}) \tag{48}$$

and inserting (45) in (48) leads with simple matrix algebra to (27). Hence, if $\boldsymbol{M}$ is invertable, the local likelihood bootstrap results by a linear combination from the consistent oracle bootstrap which finally proves the results.

■

References

Barndorff-Nielsen, O. E. and D. R. Cox (1989). *Asymptotic Techniques for use in Statistics*. Chapman & Hall.

Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models *Annals of Statistics 17*, 453–510, (with discussion).

Carroll, R. J., D. Ruppert, and A. H. Welsh (1998). Local estimating equations. *Journal of the American Statistical Association 93*, 214–227.

Davis, A. (1976). Statistical distribution in univariate and multivariate edgeworth populations. *Biometrika 63*, 661–670.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Fan, J., M. Farmen, and I. Gijbels (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Association, Series B 60*, 591–608.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Fan, J., N. E. Heckman, and M. P. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association 90*, 141–150.

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

Galindo, C. D., G. Kauermann, H. Liang, and R. J. Carroll (2000). Bootstrap confidence intervals for local likelihood, local estimating equations and varying coefficient models. Discussion Paper 205, Institut für Statistik, Universität München.

Gasser, T. and H.-G. Müller (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics 11*, 171–185.

Härdle, W. and J. S. Marron (1991). Bootstrap simultaneous error bars for nonparametric regression. *Annals of Statistics 19*(2), 778–796.

Hastie, T. (1992). Generalized additive models. In J. Chambers and T. Hastie (Eds.), *Statistical models in S*, pp. 249–308. Pacific Grove, CA: Wadsworth and Brooks/Cole.

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

Horn, R. A. and C. A. Johnson (1985). *Matrix Analysis.* Cambridge, U.K.: Cambridge University Press.

Kauermann, G., M. Müller, and R. Carroll (1998). The efficiency of bias-corrected estimators for nonparametric kernel estimation based on local estimation functions. *Stat. and Prob. Letters 37*, 41–47.

Kauermann, G. and G. Tutz (2000). Local likelihood estimation in varying-coefficient models including additive bias correction. *Journal of Nonparametric Statistics*, (to appear).

Linton, O. (2000). Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory 16*, 502–523.

Linton, O. and J.P. Nielsen (1995). Estimating structured nonparametric regression by the kernel method. *Biometrika 82*, 93–101.

Mammen, E., O. Linton and J. Nielsen (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics 27*, 1443–1490.

McCullagh, P. (1987). *Tensor Methods in Statistics.* London: Chapman & Hall.

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

Opsomer, J. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis 73*, 166–179.

Opsomer, J. and G. Kauermann (2000). Weighted local polynomial regression, weighted additive models and local scoring. Preprint 00–7, Department of Statistics, Iowa State University. Submitted to *Statistics and Probability Letters.*

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

Opsomer, J.-D. and D. Ruppert (1998). A fully automated bandwidth selection method for fitting additive models by local polynomial regression. *Journal of the American Statistical Association 93*, 605–619.

Opsomer, J.-D. and D. Ruppert (1999). A root-n consistent estimator for semiparametric additive modelling. *Journal of Computational and Graphical Statistics 8*, 715–732.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York, NY: Springer-Verlag.

Ruppert, D. and M. P. Wand (1994). Multivariate locally weighted least squares regression. *Annals of Statistics 22*, 1346–1370.

Stone, C. (1986). The dimensionality reduction principle for generalized additive models. *Annals of Statistics 14*, 590–606.

$\gamma_1(x_1)$			$\gamma_2(x_2)$		
	nominal level			nominal level	
x_1	0.90	0.95	x_2	0.90	0.95
-0.75	.885	.935	-0.75	.840	.890
-0.5	.910	.945	-0.5	.880	.945
-0.25	.905	.940	-0.25	.940	.970
0	.780	.890	0	.985	.990
0.25	.955	.980	0.25	.935	.955
0.5	.910	.950	0.5	.887	.950
0.75	.895	.955	0.75	.840	.905

Table 1: Simulated pointwise coverage probability for uniform design (b)

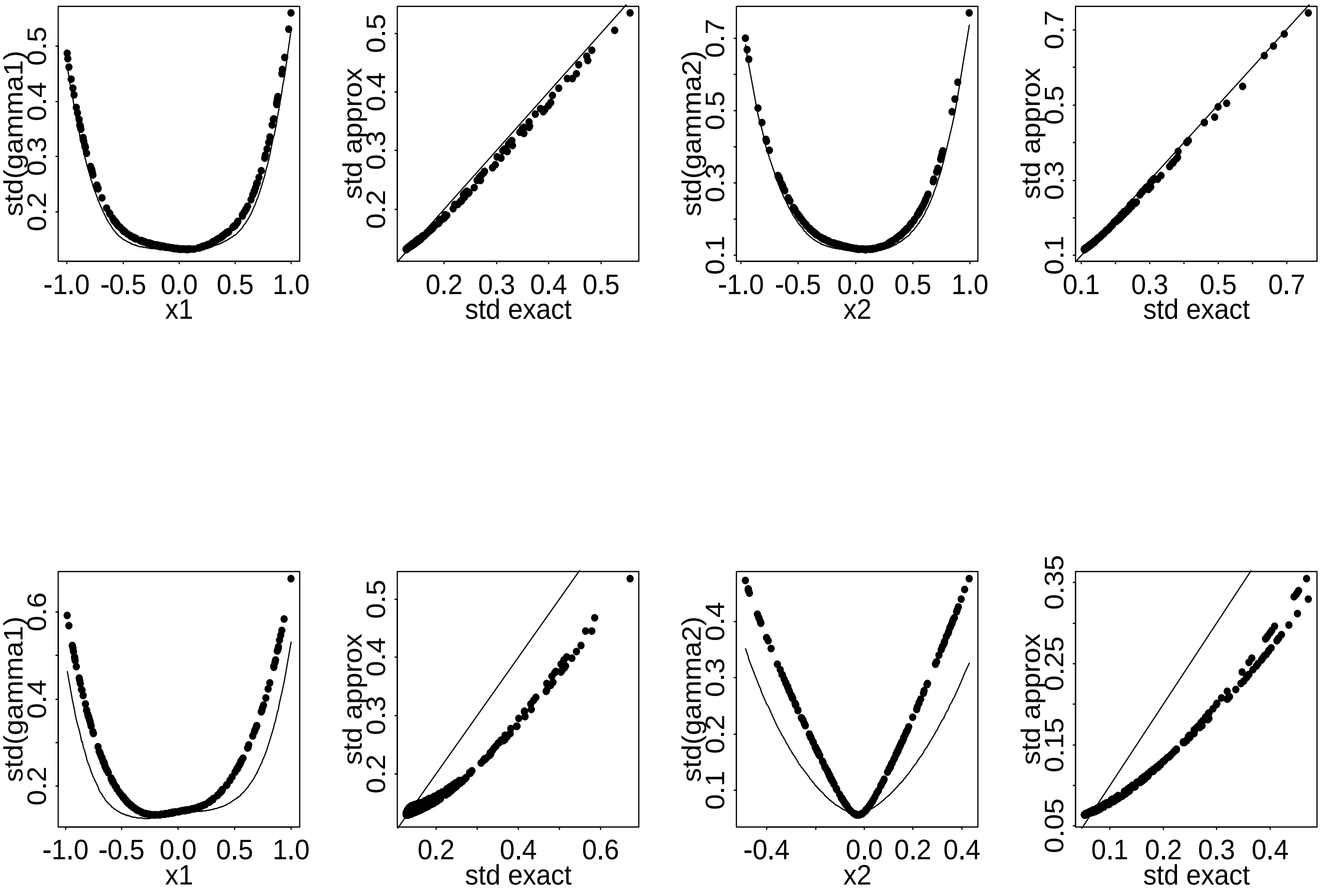

Figure 2: Calculated standard deviations using the approximative formula $(\widehat{\boldsymbol{S}}_r^{+} \boldsymbol{V} \widehat{\boldsymbol{S}}_r^{+^T})^{1/2}$ (lines) compared to $(\boldsymbol{Q}_r \boldsymbol{V} \boldsymbol{Q}_r^T)^{1/2}$ (dots) for different values for the correlation ρ among the covariates (upper row $\rho = 0.45$, lower row $\rho = 0.9$) and bandwidths $h_1 = h_2 - 0.3$.

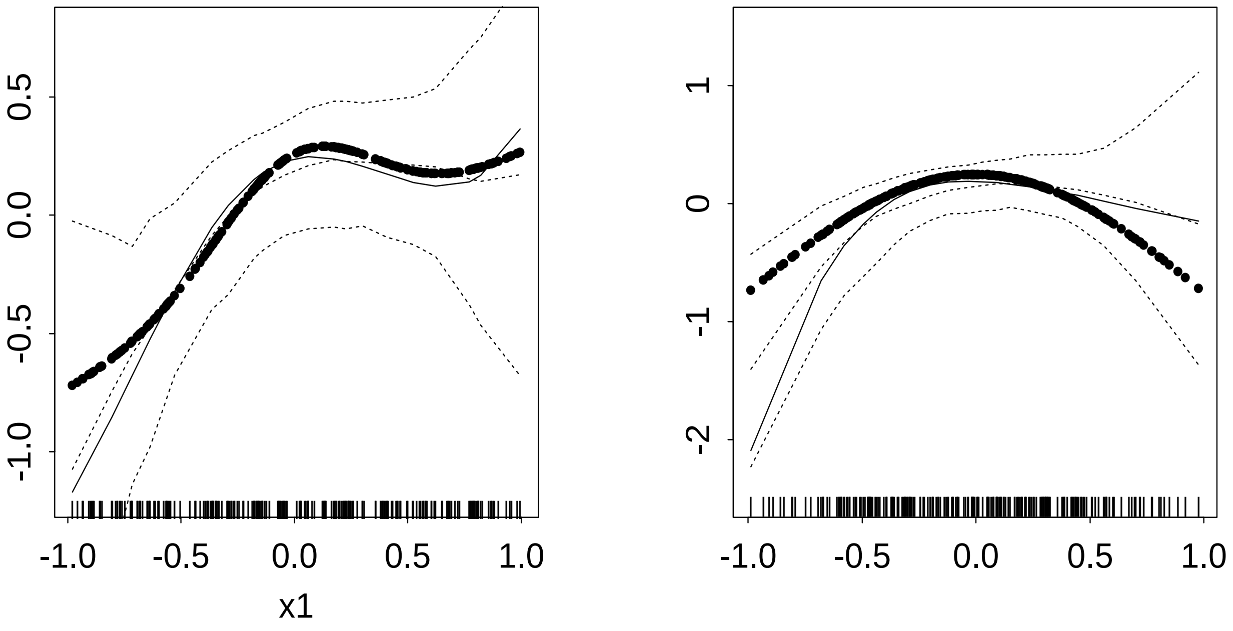

Figure 3: Simulated local backfitting estimate (solid line) and pointwise $.05, .5$ and $.95$ bootstrap quantiles (dashed lines). Dotted line shows true curve.

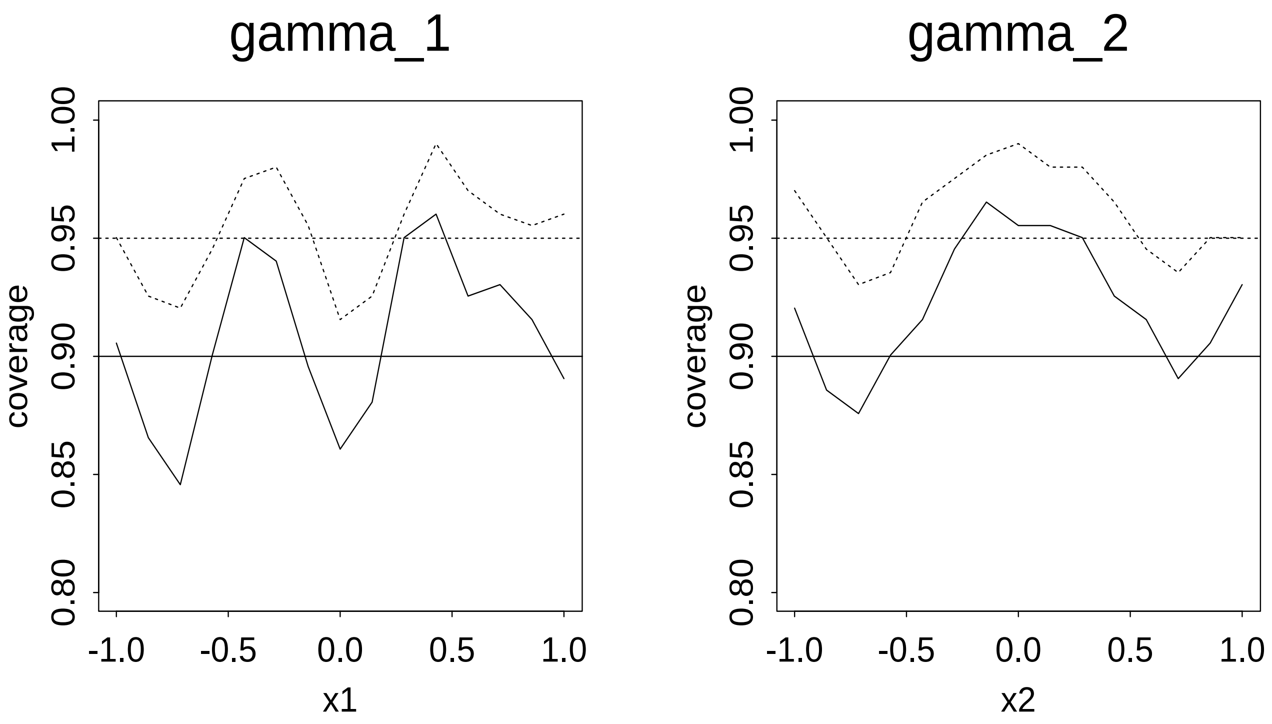

Figure 4: Simulated pointwise coverage probability for fixed design, nominal level .90 as solid line, nominal level .95 as dotted line.

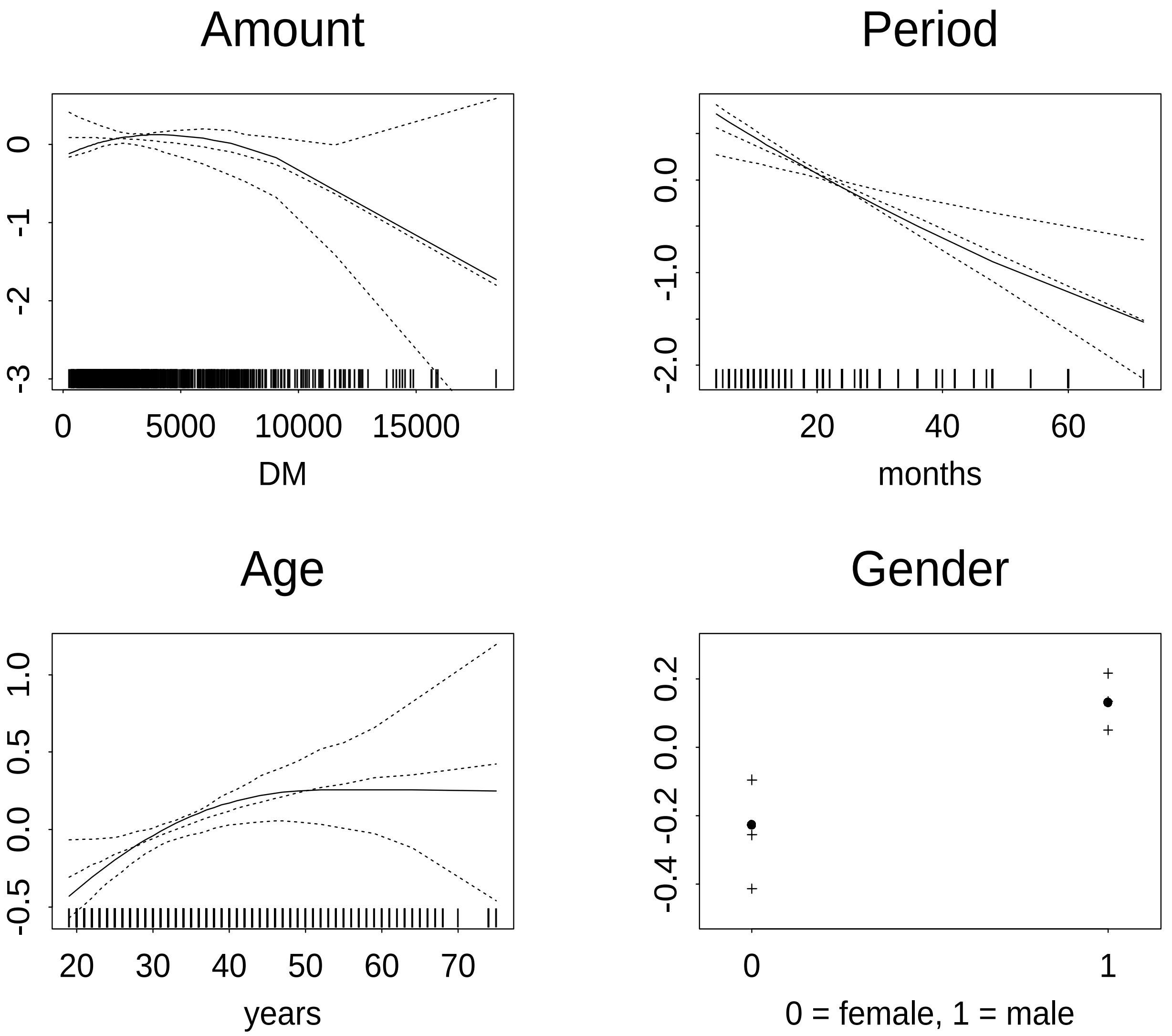

Figure 5: Backfitting estimates in credit data (solid line) with pointwise $.05, .5$ and $.95$ bootstrap quantiles (dashed lines).