Kauermann, Opsomer:

A fast method for implementing Generalized Cross-Validation in multi-dimensional nonparametric regression

Sonderforschungsbereich 386, Paper 247 (2001)

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

Projektpartner

A fast method for implementing Generalized Cross-Validation in multi-dimensional nonparametric regression

Göran Kauermann
University of Glasgow
Glasgow, UK

J.D. Opsomer
Iowa State University
Ames, USA

13th June 2001

Abstract

This article presents a modified Newton method for minimizing the Generalized Cross-Validation criterion, a commonly used smoothing parameter selection method in nonparametric regression. The method is applicable to higher dimensional problems such as additive and generalized additive models, and provides a computationally efficient alternative to full grid search in such cases. The implementation of the proposed method requires the estimation of a number of auxiliary quantities, and simple estimators are suggested. This article describes the methodology for local polynomial regression smoothing.
Keywords: local polynomial regression, generalized additive model, Newton method.

1 Introduction

Nonparametric regression techniques are popular statistical tools for exploratory analysis and model building. While these techniques are useful in univariate regression problems, it is in higher dimensional situations that they can provide the most

benefit to the data analyst. One popular class of multidimensional nonparametric regression techniques are the *generalized additive models* and *additive models*, introduced by Hastie and Tibshirani (1987) and extensively discussed in Hastie and Tibshirani (1990). The local scoring and backfitting algorithms implemented in the `gam()` routines in S-Plus provide computationally efficient methods for jointly estimating the additive component functions of such models. Recently, Kauermann and Opsomer (2000) proposed local likelihood backfitting as an alternative for generalized additive model fitting.

The regression function to be estimated in generalized additive models is of the form

$$E(y|x_1, \ldots x_q) = \mu = g\{\alpha + \gamma_1(x_1) + \ldots + \gamma_q(x_q)\}, \tag{1}$$

where $g(\cdot)$ is a known link or response function, y is the response variable, $x_1, \ldots, x_q$ are given covariates and $\gamma_k(\cdot)$, $k = 1, \ldots, q$, are unknown but smooth additive effect functions. If the link function $g(\cdot)$ is the identity function and y is normally distributed, then model (1) reduces to an additive model.

One particular concern in fitting generalized additive models, and indeed in multi-dimensional smoothing in general, is the selection of the "right" values for the smoothing parameters. Fitting a generalized additive model with q covariates requires the selection of q different smoothing parameters, one for each of the smooth component functions $\gamma_k(\cdot)$. A widely accepted criterion for smoothing parameter selection in this context is *generalized cross-validation* or *GCV*, originally proposed by Craven and Wahba (1979). For the purpose of this article, we will assume that the smoothing parameter values minimizing the GCV for a particular dataset are an appropriate smoothing parameter choice for that dataset. Our approach can be extended to other smoothing parameter selection methods such as the commonly used Akaike criterion.

Hastie and Tibshirani (1990), p. 159 discuss a version of the GCV criterion for generalized additive models based on the *deviance*. In this article, we will use the asymptotically equivalent formulation of O'Sullivan, Yandell, and Raynor (1986). Both definitions are exactly equal for the additive model case. Let $\boldsymbol{h} = (h_1, \ldots, h_q)^T$ represent the smoothing parameter vector over which are we are minimizing the GCV. For a random sample of observations $y_1, \ldots, y_n$ following model (1), the GCV objective function can be defined as

$$\mathrm{GCV}(h_1, \ldots h_q) \;=\; \frac{\sum_i v_i (y_i - \widehat{\mu}_i)^2}{n\{1 - df/n\}^2}, \tag{2}$$

where the v_i are weights to be further specified, $\widehat{\mu}_i$ are the estimates of the mean function at the observation points, calculated using $\boldsymbol{h}$, and df a measure of the complexity of the fitted nonparametric model (also dependent on $\boldsymbol{h}$). The adjustment df is used to make the GCV criterion asymptotically unbiased for the Mean Squared Error of the nonparametric estimator. Hastie and Tibshirani (1990) refer to the adjustment df as the *degree of freedom* of the estimator, by noting that it generalizes the degrees of freedom of a parametric model, and we will use that same convention here.

For $q \leq 2$, minimization of (2) is in practice usually carried out by a full grid search. For larger values of q, this grid search approach rapidly becomes computationally expensive, since the full generalized additive model needs to be refitted for each choice of values for $\boldsymbol{h}$.

Some efficient algorithms are available for smoothing parameter selection for additive models. Gu and Wahba (1991) propose a modified Newton procedure for minimizing GCV when smoothing splines are used. Their method does not carry over directly to kernel-based smoothers, the nonparametric regression method discussed in this article. Hastie and Tibshirani (1990) suggest the BRUTO algorithm which is based on iterative univariate minimization of the additive model GCV criterion. While applicable to general linear smoothers, BRUTO can be slow to converge if the covariates display significant concurvity (see Opsomer and Ruppert, 1997 for a discussion of concurvity and its effect on the asymptotically optimal bandwidths for additive models). Opsomer and Ruppert (1998) propose a bandwidth selection procedure for additive models that relies on a plug-in approach instead of GCV minimization. The method is also computationally difficult, even though the computation burden does not increase as fast as for GCV when model dimension q increases. However, none of these methods extend readily to generalized additive models.

In this article, a modified Newton procedure for minimizing (2) for local regression, including additive and generalized additive models, is proposed. We focus on local linear regression as the smoothing method, but the general approach is directly applicable to other kernel-based smoothing methods as well. The basic idea is to solve an approximation of the GCV score equations by an iterative Newton-type procedure. As in Fischer scoring, the Hessian matrix of the GCV criterion required for calculating the iteration step size is replaced by an estimate of the expected Hessian matrix. Both the score equations and the expected Hessian matrix contain

terms that cannot be computed from the data. For these terms, the article therefore proposes estimators that can be readily calculated.

In Section 2, the basic approach is explained for the univariate regression case. Sections 3 and 4 extend the proposed GCV minimization to the additive model and generalized additive model, respectively.

2 Simple Smoothing Models

We explain the approach by considering the univariate regression model $y_i = \mu(x_i) + \epsilon_i$, $i = 1, \ldots, n$ with the ϵ_i independently $N(0, \sigma^2)$ distributed. In this case, the GCV can be minimized efficiently in practice through a grid search procedure. We only present this case to explain the method in a simple setting.

We estimate $\boldsymbol{\mu} = (\mu(x_1), \ldots, \mu(x_n))^T$ by the linear smoother $\widehat{\boldsymbol{\mu}}_h = \boldsymbol{S}_h \boldsymbol{y}$ with $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{S}_h$ a linear smoothing matrix calculated with bandwidth h. We assume that $\boldsymbol{S}_h$ is the smoothing matrix resulting from a local linear fit. The bandwidth h is chosen by minimizing the GCV function

$$GCV(h) = \frac{(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})^T (\boldsymbol{y} - \widehat{\boldsymbol{\mu}})}{n\{1 - \mathrm{tr}(\boldsymbol{S}_h)/n\}^2}. \tag{3}$$

In order to use a Newton algorithm, we propose to minimize $GCV(h)$ by solving the first order derivative equation

$$\frac{\partial GCV(h)}{\partial h} = \frac{2}{n}\left(-\frac{(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_h)^t \frac{\partial \widehat{\boldsymbol{\mu}}_h}{\partial h}}{\{1 - \mathrm{tr}(\boldsymbol{S}_h)/n\}^2} + \frac{(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_h)^t (\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_h) \mathrm{tr}\left(\frac{\partial \boldsymbol{S}_h}{\partial h}\right)}{n\{1 - \mathrm{tr}(\boldsymbol{S}_h)/n\}^3} \right) = 0. \tag{4}$$

Since $\boldsymbol{S}_h$ is a linear smoother, it is easy to see that $\partial \widehat{\boldsymbol{\mu}}_h / (\partial h) = \partial \boldsymbol{S}_h / (\partial h) \boldsymbol{y}$, so that the derivative of $\boldsymbol{S}_h$ is the only unknown quantity in (4). For local linear regression, the elements of $\boldsymbol{S}_h$ are (see e.g. Fan and Gijbels, 1996)

$$\boldsymbol{S}_{h,ij} \;=\; (1, 0) \left(\sum_k w_{h,ik} \boldsymbol{X}_{ik} \boldsymbol{X}_{ik}^T \right)^{-1} w_{h,ij} \boldsymbol{X}_{ij},$$

where $w_{h,ik} = W\{(x_k - x_i)/h\}$ denotes the kernel weight calculated from the unimodal kernel $W(\cdot)$ with h as bandwidth and $\boldsymbol{X}_{ik}^T = (1, x_k - x_i)$.

We will use the following convenient approximation for the elements of $\boldsymbol{S}_h$:

$$\boldsymbol{S}_{h,ij} \;\approx\; K\left(\frac{x_i - x_j}{h}\right) \Big/ \{nhf(x_i)\}. \tag{5}$$

where $f(x_i)$ denotes the design density and $K(\cdot)$ is an equivalent kernel of order 2. In particular, it follows that $\int K(z)dz = 1$ and $\int z^2 K(z)dz = \mu_2(K)$. Using approximation (5), one finds

$$\frac{\partial \boldsymbol{S}_{h,ij}}{\partial h} \approx -\frac{K'\left(\frac{x_i - x_j}{h}\right)\frac{x_i - x_j}{h}}{nh^2 f(x_i)} - h^{-1}\boldsymbol{S}_{h,ij}.$$

Note that this implies $\mathrm{tr}\{\partial \boldsymbol{S}_h/(\partial h)\} = -\mathrm{tr}(\boldsymbol{S}_h)/h$. For the full matrix derivative $\partial \boldsymbol{S}_h/\partial h$, we construct an estimator by employing a second smoothing step. Let $\tilde{h}$ be a second bandwidth with $\tilde{h} = o(h)$, i.e. $\tilde{h}$ tends faster to zero than h, and let $\tilde{K}(\cdot)$ define a second kernel of order 2, with $\tilde{K}(\cdot)$ not necessarily equal to $K(\cdot)$ (as for $K(\cdot)$, this can also be achieved by using local linear regression). Using $\tilde{K}(\cdot)$, we build a smoothing matrix $\tilde{\boldsymbol{S}}_{\tilde{h}}$ with entries $\tilde{\boldsymbol{S}}_{\tilde{h},ij} \approx \tilde{K}\{(x_i - x_j)/\tilde{h}\}/\{nhf(x_i)\}$, and let $\tilde{\boldsymbol{R}}_{\tilde{h}}$ be defined by $\tilde{\boldsymbol{R}}_{\tilde{h},ij} = \tilde{\boldsymbol{S}}_{\tilde{h},ij}(x_i - x_j)/\tilde{h}$. Matrix multiplication and standard smoothing arguments yield

$$\begin{aligned}
[\boldsymbol{S}_h \tilde{\boldsymbol{R}}_{\tilde{h}}]_{ij} &\approx \frac{1}{n^2 \tilde{h} h f(x_i)} \sum_k K\left(\frac{x_i - x_k}{h}\right)\frac{x_k - x_j}{\tilde{h}}\tilde{K}\left(\frac{x_k - x_j}{\tilde{h}}\right)/f(x_k) \\
&= -\frac{\tilde{h}}{h^2}\frac{K'\left(\frac{x_i - x_j}{h}\right)}{nf(x_i)}\{1 + o\left(\frac{\tilde{h}}{h}\right)\},
\end{aligned}$$

where we assumed that $\int z^2 \tilde{K}(z)dz = 1$.

With matrix $\boldsymbol{N}_h$ defined as $\boldsymbol{N}_{h,ij} = [\boldsymbol{S}_h \tilde{\boldsymbol{R}}_{\tilde{h}}]_{ij}(x_i - x_j)/\tilde{h}$, the approximation

$$\frac{\partial \boldsymbol{S}_h}{\partial h} \approx h^{-1}(\boldsymbol{N}_h - \boldsymbol{S}_h) \tag{6}$$

follows. For a given choice of the second bandwidth $\tilde{h}$, expression (6) provides an estimator for $\partial \boldsymbol{S}_h/\partial h$ that can be used in (4).

Solving (4) by applying a Newton algorithm also requires the calculation of the second order derivative. As in Fischer scoring, the *expected* second order derivative is used instead to improve numerical stability of the algorithm. Assuming that we can interchange the order of the integration and differention, the objective is therefore to calculate $\partial^2 \mathrm{E}\{GCV(h)\}/(\partial h)^2$. The expectation of the GCV function can be approximated by

$$\mathrm{E}\{GCV(h)\} = [\sigma^2\{n - \mathrm{tr}(2\boldsymbol{S}_h - \boldsymbol{S}_h\boldsymbol{S}_h^T)\} + \boldsymbol{b}_h^T\boldsymbol{b}_h]\left(\frac{1}{n} + \frac{2\mathrm{tr}(\boldsymbol{S}_h)}{n^2} + \ldots\right),$$

with $\boldsymbol{b}_h = E(\widehat{\boldsymbol{\mu}}_h) - \boldsymbol{\mu} = \boldsymbol{S}_h\boldsymbol{\mu} - \boldsymbol{\mu}$ denoting the smoothing bias. A similar reasoning as above shows that

$$\frac{\partial \mathrm{tr}(2\boldsymbol{S}_h - \boldsymbol{S}_h\boldsymbol{S}_h^T)}{\partial h} \approx -h^{-1}\mathrm{tr}(2\boldsymbol{S}_h - \boldsymbol{S}_h\boldsymbol{S}_h^T). \tag{7}$$

Moreover, the squared bias $\boldsymbol{b}_h^T\boldsymbol{b}_h$ for local linear regression has the asymptotic order h^4, so that $\partial^2(\boldsymbol{b}_h^T\boldsymbol{b}_h)/(\partial h)^2 \approx 12\boldsymbol{b}_h^T\boldsymbol{b}_h/h^2$. Using these approximations, we find

$$\frac{\partial^2 E\{GCV(h)\}}{(\partial h)^2} \approx \frac{1}{nh^2}\{2\sigma^2\mathrm{tr}(2\boldsymbol{S}_h - \boldsymbol{S}_h\boldsymbol{S}_h^T) + 12\boldsymbol{b}_h^T\boldsymbol{b}_h\}. \tag{8}$$

Since the bias $\boldsymbol{b}_h$ is unknown, we replace it by the plug-in estimator $\widehat{\boldsymbol{b}}_h = \boldsymbol{S}_h\widehat{\boldsymbol{\mu}}_h - \widehat{\boldsymbol{\mu}}_h$. Expression (8) therefore provides an estimate for the expected second order derivative that can be used in the GCV minimization.

We propose the following Newton-based algorithm for minimizing the GCV criterion.

Generalized Cross Validation by the Newton algorithm:

i Let h_0 be an initial bandwidth and set $h_t = h_0$. Choose $\tilde{h}_t = o(h_t)$ used in (6) by setting e.g. $\tilde{h}_t = h_t^{5/7}$.

ii For $t = 1, 2, \ldots$ calculate the update h_{t+1} by

$$h_{t+1} := h_t - \frac{\partial \mathrm{GCV}(h_t)}{\partial h} \Big/ \frac{\partial^2 E\{\mathrm{GCV}(h_t)\}}{(\partial h)^2}$$

using (4), (8) and the approximations proposed above.

iii Repeat step ii until changes in $\mathrm{GCV}(h)$ are negligible.

It is interesting to note that a simple first order approximation for $GCV(h_{t+1})$ is available by the Taylor expansion

$$\mathrm{GCV}(h_{t+1}) \approx \mathrm{GCV}(h_t) - 0.5\left(\frac{\partial \mathrm{GCV}(h_t)}{\partial h}\right)^2 \Big/ \left(\frac{\partial^2 \mathrm{GCV}(h_t)}{\partial h^2}\right).$$

This allows to predict the amount of reduction of the generalized cross validation function gained by moving from h_t to h_{t+1}. Moreover, since the second order derivative (8) is positive, it becomes obvious that $GCV(h_{t+1}) < GCV(h_t)$, at least approximately, with equality if the first order derivative (4) is zero.

Example:

For illustration we apply the algorithm to a simple example. We draw $n = 100$

observations from the model $y_i = \mu(x_i) + \epsilon_i$ with $\epsilon_i \sim N(0, 0.5^2)$ and $\mu(x_i) = 2x_i + \cos(\pi x_i)$, where x_i are skewly distributed from a triangular distribution, i.e. x_i^2 are equidistant in $[0, 1]$. Figure 1, left plot, shows the data and the true curve. The right plot shows the function $GCV(h)$ for a local linear fit of the data. The steps of the Newton algorithm are indicated by the numbers 0 to 3, for two different starting points. The algorithm converges quickly to points close to the minimum. It does not reach the GCV minimum exactly since $\partial \boldsymbol{S}_h/(\partial h)$ is estimated by (6), which is based on an asymptotic approximation. For practical purposes, however, this is sufficiently close to the true GCV minimizer to result in a fit $\widehat{\boldsymbol{\mu}}_h$ that is virtually indistinguisable from the one calculated at the GCV minimum.

In order to see the performance of the procedure, we repeat the simulation 100 times and record the optimal bandwidth h_{opt}, which minimizes (3), and the estimated bandwidth h_t resulting after 5 steps with the Newton procedure. We again use two different starting values h_0. Figure 2 shows h_{opt} plotted against h_t for the two starting values used above, and the corresponding values of GCV(h_{opt}) against GCV(h_t). By comparing the two lower plots, it appears that a starting value based on undersmoothing converges faster to the minimal value. This is likely due to the shape of the function GCV(h), as well as to the fact that in the case of a large starting value, the plug-in estimate of the bias does not perform well. This in turn leads to a rough approximation of the second order derivative, which steers the step size in the Newton algorithm. The two upper plots also indicate a tendency towards oversmoothing in the Newton-based method. However, as shown in the lower two plots, the bandwidth values selected by the proposed method result in GCV scores that are almost identical to those found with the optimal grid-search bandwidth. Hence, this slight oversmoothing does not appear to result in a degraded bandwidth selection procedure. Overall, this simulation experiment illustrates that the Newton procedure serves as an appropriate GCV bandwidth selector in this simple case.

3 Additive Models

In this section, we extend the results from the univariate case to normal response additive models of the form $y = \alpha + \mu_1(x_1) + \ldots + \mu_q(x_q) + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$. For identifiability reasons, we assume that the component functions $\mu_k(\cdot)$ have zero mean in the sense $\int \mu_k(x_k) f_k(x_k) dx_k = 0$, with $f_k(\cdot)$ as design density for x_k. Let

$\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{x}_k = (x_{1k}, \ldots, x_{nk})^T$, $k = 1, \ldots q$ denote a random sample, where $(x_{i1}, \ldots, x_{iq})$ is randomly drawn from the joint design density $f_x(\cdot)$ for $i = 1, \ldots, n$. We consider the additive model estimator

$$\widehat{\boldsymbol{\mu}}_k = \boldsymbol{S}^*_{k,h_k}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_{-k}) \tag{9}$$

for $\boldsymbol{\mu}_k = \{\mu_k(x_{1k}), \ldots, \mu_k(x_{nk})\}^T$, with $\widehat{\boldsymbol{\mu}}_{-k} = \sum_{r \neq k} \widehat{\boldsymbol{\mu}}_r$ an estimator for the remaining $q-1$ functions and $\boldsymbol{S}^*_{k,h_k} = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^t/n)\boldsymbol{S}_{k,h_k}$ as centered smoothing matrix, $\boldsymbol{I}$ denoting the identity matrix, $\boldsymbol{1} = (1, \ldots, 1)^t$ and $\boldsymbol{S}_{k,h_k}$ a smoothing matrix as given in (5) with bandwidth h_k. Joint estimation of the component functions $\mu_1, \ldots, \mu_q$ is most often performed by iterating (9) over k. In the additive model context, this is referred to as the *backfitting* procedure (Hastie and Tibshirani, 1990). For simplicity we abbreviate $\boldsymbol{S}_{k,h_k}$ by $\boldsymbol{S}_k$ and assume $\bar{y} = 0$ in the following. Equation (9), combined into the backfitting algorithm, can be jointly written as

$$\boldsymbol{M}\widehat{\boldsymbol{\mu}}_\bullet \;=\; \boldsymbol{S}^*_\bullet \boldsymbol{y}, \tag{10}$$

where $\widehat{\boldsymbol{\mu}}_\bullet = (\widehat{\boldsymbol{\mu}}_1^T, \ldots, \widehat{\boldsymbol{\mu}}_q^T)^T$, $\boldsymbol{S}^*_\bullet = ({\boldsymbol{S}_1^*}^T, \ldots, {\boldsymbol{S}_q^*}^T)^T$ and

$$\boldsymbol{M} \;=\; \begin{pmatrix} \boldsymbol{I} & \boldsymbol{S}^*_1 & \cdots & \boldsymbol{S}^*_1 \\ \boldsymbol{S}^*_2 & \boldsymbol{I} & \cdots & \boldsymbol{S}^*_2 \\ \vdots & & \ddots & \vdots \\ \boldsymbol{S}^*_q & \boldsymbol{S}^*_q & \cdots & \boldsymbol{I} \end{pmatrix}.$$

It is possible to write the k component estimators as $\widehat{\boldsymbol{\mu}}_k = \boldsymbol{Q}_k\boldsymbol{y}$, where $\boldsymbol{Q}_k = \boldsymbol{M}^{kk}\boldsymbol{S}^*_k(\boldsymbol{I} - \boldsymbol{S}^*_{-k})$ with $\boldsymbol{M}^{kk}$ denoting the k-th $n \times n$ block diagonal matrix of $\boldsymbol{M}^{-1}$ and $\boldsymbol{S}^*_{-k} = \sum_{r \neq k} \boldsymbol{S}^*_r$. The estimated mean $\widehat{\boldsymbol{\mu}} = \sum_k \widehat{\boldsymbol{\mu}}_k$ is obtained by $\boldsymbol{Q}_+\boldsymbol{y} = \sum_k \boldsymbol{Q}_k\boldsymbol{y}$. Note that $\boldsymbol{M}$ and $\widehat{\boldsymbol{\mu}}$ depend on the entire bandwidth vector $\boldsymbol{h} = (h_1, \ldots, h_q)$, even though this is suppressed in the notation.

The direct generalization of GCV criterion (3) to the additive model would be to use

$$\mathrm{GCV}(\boldsymbol{h}) = \frac{(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})^T(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})}{n\{1 - \boldsymbol{Q}_+/n\}^2}.$$

However, the matrix $\boldsymbol{Q}_+$ is difficult to compute, since it requires inverting the $nq \times nq$ matrix $\boldsymbol{M}$. We therefore propose selecting $\boldsymbol{h}$ by minimizing an approximation to this criterion, i.e.

$$\mathrm{GCV}(\boldsymbol{h}) = \frac{(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})^T(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})}{n\{1 - \sum_k \mathrm{tr}(\boldsymbol{S}^*_k)/n\}^2}. \tag{11}$$

This was also proposed by Hastie and Tibshirani (1990) for additive model bandwidth selection. If the covariates $x_1, \ldots, x_q$ are independent, this approximation to the degrees of freedom is asymptotically justified, as can be derived using the arguments in Opsomer (2000). In the presence of correlation, the approximation is likely to overestimate the true degrees of freedom of the model. See Buja, Hastie, and Tibshirani (1989) for a further discussion of this approximation.

Differentiation of (11) with respect to h_1, say, yields

$$\frac{\partial \mathrm{GCV}(\boldsymbol{h})}{\partial h_1} = \frac{2}{n}\left(-\frac{(\boldsymbol{y}-\widehat{\boldsymbol{\mu}})^T \frac{\partial \widehat{\boldsymbol{\mu}}}{\partial h_1}}{(1-\sum_k \mathrm{tr}(\boldsymbol{S}_k^*)/n)^2} + \frac{(\boldsymbol{y}-\widehat{\boldsymbol{\mu}})^T(\boldsymbol{y}-\widehat{\boldsymbol{\mu}})\mathrm{tr}\left(\frac{\partial \boldsymbol{S}_1^*}{\partial h_1}\right)}{n(1-\sum_k \mathrm{tr}(\boldsymbol{S}_k^*)/n)^3} \right). \quad (12)$$

As in the previous section, we find an approximation to (12) that can be used in the derivation of the minimization algorithm. Note that we again find $\mathrm{tr}(\partial \boldsymbol{S}_1^*/\partial h_1) = -\mathrm{tr}(\boldsymbol{S}_1^*)/h_1$. The derivative of $\partial\widehat{\boldsymbol{\mu}}/(\partial h_1)$ is obtained by differentiating both sides of (10) with respect to h_1. This yields

$$\begin{pmatrix} 0 & \frac{\partial \boldsymbol{S}_1^*}{\partial h_1} & \cdots & \frac{\partial \boldsymbol{S}_1^*}{\partial h_1} \\ 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \widehat{\boldsymbol{\mu}}_\bullet + \boldsymbol{M}\frac{\partial \widehat{\boldsymbol{\mu}}_\bullet}{\partial h_1} = \begin{pmatrix} \frac{\partial \boldsymbol{S}_1^*}{\partial h_1}\boldsymbol{y} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (13)$$

so that

$$\frac{\partial \widehat{\boldsymbol{\mu}}}{\partial h_1} = \boldsymbol{P}_1 \frac{\partial \boldsymbol{S}_1^*}{\partial h_1}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_{-1}) \quad (14)$$

with $\boldsymbol{P}_1 = (\boldsymbol{I} - \boldsymbol{S}_{-1}^*)\boldsymbol{M}^{11}$. The derivative $\partial \boldsymbol{S}_1^*/(\partial h_1) = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n)\partial \boldsymbol{S}_1/(\partial h_1)$ can now be estimated as in the previous section using approximation (6), setting $\tilde{h}_1 = o(h_1)$.

In order to apply the Newton procedure we derive now an approximation for the expected second order derivative. Taking expectation gives

$$\mathrm{E}\{\mathrm{GCV}(\boldsymbol{h})\} = [\sigma^2\{n - \mathrm{tr}(2\boldsymbol{Q}_+ - \boldsymbol{Q}_+\boldsymbol{Q}_+^t)\} + \boldsymbol{b}_+^T\boldsymbol{b}_+]\{\frac{1}{n} + \frac{2}{n^2}\sum_k \mathrm{tr}(\boldsymbol{S}_k) + \ldots\} \quad (15)$$

where $\boldsymbol{b}_+ = E(\widehat{\boldsymbol{\mu}}) - \boldsymbol{\mu} = \boldsymbol{Q}_+\boldsymbol{\mu} - \boldsymbol{\mu}$ is the overall bias. Analogously to what we did for the degrees of freedom of the GCV, we approximate $\mathrm{tr}(2\boldsymbol{Q}_+ - \boldsymbol{Q}_+\boldsymbol{Q}_+^T)$ in (15) by $\sum_k \mathrm{tr}(2\boldsymbol{S}_k^* - \boldsymbol{S}_k^*\boldsymbol{S}_k^{*^T})$ which holds in an asymptotic sense if the covariates $x = (x_1, \ldots x_q)$ are independent.

Before differentiating (15), we consider the bias $\boldsymbol{b}_+$ in more depth. Assume for the moment that $\boldsymbol{\mu}_{-k} = \sum_{r \neq k} \boldsymbol{\mu}_r$ is known so that the estimate

$$\widehat{\boldsymbol{\mu}}_{k|-k} = \boldsymbol{S}_k^*(\boldsymbol{y} - \boldsymbol{\mu}_{-k}) \tag{16}$$

is a univariate smooth component with $\boldsymbol{\mu}_{-k}$ a non-random offset. We refer to $\widehat{\boldsymbol{\mu}}_{k|-k}$ as the "oracle" estimate of $\boldsymbol{\mu}_k$ (Kauermann and Opsomer, 2000). Using the relations $\widehat{\boldsymbol{\mu}}_{k|-k} = \boldsymbol{S}_k^*(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_{-k}) + \boldsymbol{S}_k^*(\widehat{\boldsymbol{\mu}}_{-k} - \boldsymbol{\mu}_{-k})$ and (9), it is easily seen that

$$\boldsymbol{M}(\widehat{\boldsymbol{\mu}}_\bullet - \boldsymbol{\mu}_\bullet) \;\; = \;\; (\widehat{\boldsymbol{\mu}}_{\bullet|-\bullet} - \boldsymbol{\mu}_\bullet), \tag{17}$$

where $\widehat{\boldsymbol{\mu}}_{\bullet|-\bullet} = (\widehat{\boldsymbol{\mu}}_{1|-1}^T, \ldots, \widehat{\boldsymbol{\mu}}_{q|-q}^T)^T$. Let now $\boldsymbol{b}_k = E(\widehat{\boldsymbol{\mu}}_k) - \boldsymbol{\mu}_k$ be the bias for the k-th component and $\boldsymbol{b}_{k|-k} = E(\widehat{\boldsymbol{\mu}}_{k|-k}) - \boldsymbol{\mu}_k = \boldsymbol{S}_k^*\boldsymbol{\mu}_k - \boldsymbol{\mu}_k$, be the "oracle" bias. Taking expectation on both sides of (17) provides the bias relation

$$\boldsymbol{M}\boldsymbol{b}_\bullet = \boldsymbol{b}_{\bullet|-\bullet} \tag{18}$$

with $\boldsymbol{b}_\bullet = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_q^T)^T$ and $b_{\bullet|-\bullet} = (\boldsymbol{b}_{1|-1}^T, \ldots \boldsymbol{b}_{q|-q}^T)^T$, so that

$$\boldsymbol{b}_+ = \sum_k \boldsymbol{b}_k \;\; = \;\; \boldsymbol{P}_1\boldsymbol{b}_{1|-1} + \ldots \boldsymbol{P}_q\boldsymbol{b}_{q|-q} =: \tilde{\boldsymbol{b}}_1 + \ldots \tilde{\boldsymbol{b}}_q \tag{19}$$

with $\tilde{\boldsymbol{b}}_k = \boldsymbol{P}_k\boldsymbol{b}_{k|-k}$ and $\boldsymbol{P}_k$ as defined above. The oracle bias $\boldsymbol{b}_{k|-k}$ depends only on the bandwidth h_k and in the case of local linear fitting, $\boldsymbol{b}_{k|-k} = O_p(h_k^2)$. As shown in Opsomer and Ruppert (1997) for the case $q = 2$, the matrices $\boldsymbol{P}_k$, $k = 1, \ldots q$ are of asymptotic order $O(\mathbf{1}\mathbf{1}^T/n)$ when $\boldsymbol{M}$ is invertible and under a number of technical regularity conditions. The same order can be derived for $q > 2$ using the results in (Opsomer, 2000). Hence, $\tilde{\boldsymbol{b}}_k = O_p(h_k^2)$ follows and (19) is a decomposition of the bias in components of order $O(h_k^2)$, $k = 1, \ldots, q$. Differentiating the bias then yields

$$\frac{\partial^2 \boldsymbol{b}_+^T\boldsymbol{b}_+}{(\partial h_k)^2} \;\; = \;\; \frac{\partial^2 \tilde{\boldsymbol{b}}_k^T\tilde{\boldsymbol{b}}_k}{(\partial h_k)^2} + 2\frac{\partial^2 \tilde{\boldsymbol{b}}_k^T}{(\partial h_k)^2}\tilde{\boldsymbol{b}}_{-k} \approx h_k^{-2}(12\tilde{\boldsymbol{b}}_k^T\tilde{\boldsymbol{b}}_k + 4\tilde{\boldsymbol{b}}_k^T\tilde{\boldsymbol{b}}_{-k}) \tag{20}$$

$$\frac{\partial^2 \boldsymbol{b}_+^T\boldsymbol{b}_+}{\partial h_k \partial h_l} \;\; \approx \;\; 8(h_k h_l)^{-1}\tilde{\boldsymbol{b}}_k^T\tilde{\boldsymbol{b}}_l \tag{21}$$

by using the same reasoning as in the previous section.

The above results can now be applied to calculate an approximation of the second order derivative of (15). Making use of (7) one finds

$$\begin{aligned}\frac{\partial^2 E\{GCV(h)\}}{\partial h_k^2} \;\; &\approx \;\; \frac{1}{n}\left\{\sigma^2\frac{\partial^2 \mathrm{tr}(2\boldsymbol{S}_k^* - \boldsymbol{S}_k^{*^T}\boldsymbol{S}_k^*)}{\partial h_k^2} + \frac{\partial^2 \boldsymbol{b}_+^T\boldsymbol{b}_h}{\partial h_k^2}\right\} \\ &\approx \;\; \frac{1}{nh_r^2}\{2\sigma^2\mathrm{tr}(\boldsymbol{S}_k^{*^T}\boldsymbol{S}_k^*) + 12\tilde{\boldsymbol{b}}_k^T\tilde{\boldsymbol{b}}_k + 4\tilde{\boldsymbol{b}}_k^T\tilde{\boldsymbol{b}}_{-k}\} \qquad (22)\end{aligned}$$

$$\frac{\partial^2 E\{GCV(h)\}}{\partial h_r \partial h_k} \;\; \approx \;\; \frac{1}{n}\left(\frac{\partial^2 \boldsymbol{b}_+^T\boldsymbol{b}_+}{\partial h_r \partial h_r}\right) \approx \frac{8}{nh_r h_k}\tilde{\boldsymbol{b}}_1^T\tilde{\boldsymbol{b}}_2 \tag{23}$$

Calculation of (22) and (23) requires the estimation of the bias components $\tilde{\boldsymbol{b}}_k$. As in the previous section, we make use of the plug-in estimates $\widehat{\boldsymbol{b}}_k = \boldsymbol{Q}_k\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_k$ which, together with (18) and (19), provide estimates for $\tilde{\boldsymbol{b}}_k$, $k = 1, \ldots, q$. Direct calculation of the matrices $\boldsymbol{Q}_k$ and $\boldsymbol{P}_k$, $k = 1, \ldots q$, is numerically expensive and can be avoided by applying the backfitting idea, i.e. one can calculate $\boldsymbol{Q}_k = \boldsymbol{S}_k^*(\boldsymbol{I} - \sum_{r\neq k}\boldsymbol{Q}_r)$ and $\boldsymbol{P}_k = \boldsymbol{I} - \sum_{r\neq k}\boldsymbol{P}_r\boldsymbol{S}_r$ by iteration over $k = 1, \ldots q$. It is easily checked that at convergence, $\boldsymbol{Q}_k = \boldsymbol{M}^{kk}\boldsymbol{S}^*(\boldsymbol{I} - \boldsymbol{S}_{-k}^*)$ and $\boldsymbol{P}_k = (\boldsymbol{I} - \boldsymbol{S}_{-k}^*)\boldsymbol{M}^{kk}$. In practice, a small number of iteration loops is sufficient to obtain reliable approximations for $\boldsymbol{Q}_k$ and $\boldsymbol{P}_k$. One should also keep in mind that the calculation of $\boldsymbol{P}_k$ and $\boldsymbol{Q}_k$ is only required to get an estimate for the step size of the algorithm. This means that fast and rough approximations of the matrices are usually sufficient to achieve a reasonable performance of the algorithm.

The above results can now readily be applied to define a multivariate version of the Newton GCV procedure.

Generalized Cross Validation by the Multivariate Newton algorithm:

i Let $\boldsymbol{h}_0 = (h_{01}, \ldots, h_{0q})$ be an initial bandwidth and set $\boldsymbol{h}_t = \boldsymbol{h}_0$. Choose $\tilde{\boldsymbol{h}}_t = (\tilde{h}_{t1}, \ldots, \tilde{h}_{tq})$ such that $\tilde{h}_{tr} = o(h_{tr})$ for $r = 1, \ldots, q$ e.g. set $\tilde{h}_{tr} = h_{tr}^{5/7}$.

ii For $t = 1, 2, \ldots$ calculate the update $\boldsymbol{h}_{t+1}$ by

$$\boldsymbol{h}_{t+1} := \boldsymbol{h}_t - \left[\frac{\partial^2 E\{GCV(h_t)\}}{(\partial \boldsymbol{h})(\partial \boldsymbol{h})^T}\right]^{-1}\left[\frac{\partial GCV(h_t)}{\partial \boldsymbol{h}}\right]$$

using (12), (14), (22) and (23) and the approximations proposed above.

iii Repeat step ii until changes in $GCV(\boldsymbol{h})$ are negligible.

During simulation experiments, it was found that the numerical stability of the algorithm was improved by reducing the step size in the first steps of the algorithm by a multiplicative factor δ, with $0 < \delta \leq 1$. This adjustment is particularly useful when the initial values $\boldsymbol{h}_0$ correspond to oversmoothing, since in this case the bias in not estimated reliably by the plug-in estimate. For bandwidth $\boldsymbol{h}$ close to the optimum a step size reduction is not an issue.

Example:
We study the behavior of the proposed method through a simulation experiment.

We generate data from the bivariate additive model $y = \mu_1(x_1) + \mu_2(x_2) + \epsilon_i$ with $\mu_1(x) = x^2$ and $\mu_2(x) = x + 0.3\cos(\pi x)$ and $\epsilon \sim N(0, 0.3^2)$. The covariates x_1 and x_2 are drawn from a truncated bivariate normal distribution on $[-1,1]^2$ with correlation 0.5. Figure 3 shows the function $\mathrm{GCV}(h_1, h_2)$ for a local linear fit for a sample of size $n = 100$. The steps of the Newton algorithm are indicated by 0 to 4 for four different starting points. The algorithm converges quickly to the minimum as desired. We repeat this simulation 100 times. In Figure 4 we plot $\mathrm{GCV}(\boldsymbol{h}_{opt}) = \min\{\mathrm{GCV}(\boldsymbol{h})\}$, calculated from a 7×7 grid, against $\mathrm{GCV}(\boldsymbol{h}_t)$ for the 3rd and the 5th step of the Newton algorithm using $\boldsymbol{h} = (0.4, 0.4)$ as starting value. As in the univariate case, the procedure appears to work well and converges very fast, i.e. after about 3 steps of the Newton algorithm, the value $\mathrm{GCV}(\boldsymbol{h}_t)$ is close to the minimal one. In this and other simulation settings, a step size adjustment of $\delta = 0.5$ performed satisfactorily.

We now consider a more general setting by simulating from $y = \mu_1(x_1)+\mu_2(x_2)+\epsilon_i$ with $\mu_1(x) = x^p$ and $\mu_2(x) = x + 0.3\cos(q\pi x)$ for different values of p and q. The covariates are drawn from a bivariate normal density with mean (0,0), standard deviations $\sigma_1 = \sigma_2 = 1$, correlation ρ and truncated to $[-1,1]^2$. We simulate from the following settings: $p = 1, 2$, $q = 0, 1$ and $\rho = 0, 0.5$ with sample size $n = 200$ and a simulation size of order 150. In Table 1 we give the empirical correlation $cor\{GCV(\widehat{h}), GCV(h_{opt})\}$ after 5 steps of the Newton procedure and $GCV(h_{opt})$ calculated on the 7×7 with $h_k \in \{0.05, 0.125, 0.2, \ldots, 0.5\}$ for $k = 1, 2$. As in Figure 4 the correlation is nearly 1 meaning that the Newton procedure reaches the minimum after a few steps. For the two setting (a) $p = 2, q = 1$ and $\rho = 0$ as well as for (b) $p = 1, q = 1$ and $\rho = 0$ we plot the selected bandwidths exemplary in the histograms shown in Figures 5 and 6. For setting (a) the histograms are very much alike. In contrast, for setting (b) the theoretically optimal bandwidth for h_1 is infinity, since the effect is linear. This can not be accomplished in grid search, but in contrast, the Newton procedure chooses a large value for h_1 in most simulations. For demonstration purposes, we grouped bandwidths chosen larger than 0.5 in the category $[0.5, 0.6]$ in the histogram in Figure 6. The Newton Procedure in this setting clearly uncovers the linear parametric structure of the simulation setting. A similar behaviour was observed in other simulation settings.

Example:
We demonstrate the procedure on a literature data example giving the atmospheric

ozone concentration in the Los Angeles basin 1976. The data are described in Hastie and Tibshirani (1990) who use them to demonstrate different bandwidth selection routines. Covariables considered are `vh` (millibar pressure height, measured at location 1), `wind` (wind speed), `humidity`, `temp` (temperature), `ibh` (temperature inversion height, measured at location 1), `dpg` (pressure gradient from Los Angeles airport LAX), `ibt` (inversion temperature at LAX), `vis` (visibility) and `doy` (day of the year). We start the Newton procedure using for h_0 the empirical standard deviations of each covariate. The steps of the algorithm are shown in Table 2. A step size reduction using $\delta = 0.3$ was applied to stabilize the performance of the algorithm. The final curves are visible from Figure 7. Comparing the plots with the procedures used in Hastie and Tibshirani (1990) shows similarities with the fits obtained by fitting an additive model with 4 degrees of freedom for each smooth component, except of the clear linear shape chosen for `vh`, `humidity` and `ibh` The odd shape for `wind` is determined by the influential point at 21. Overall, the Newton procedure proves to behave rather satisfactory in this high dimensional example, allowing to obtain bandwidths close to the minimal value of $GCV(\boldsymbol{h})$ after already 5 steps.

4 Generalized Additive Models

In this section we extend the Newton-based GCV minimization procedure to generalized additive models of the form (1). The response y for the given predictor $\eta = \alpha + \gamma_1(x_1) + \ldots + \gamma_q(x_q)$ is distributed according to density $f(y|\eta)$ which is assumed to be of exponential family form

$$f(y|\eta) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right), \tag{24}$$

where $\theta = \theta(\eta)$ denotes the natural parameter corresponding to the expectation $h(\eta)$ and ϕ as dispersion parameter which assumed to be known. For simplicity of notation we restrict the presentation to natural links function, i.e. we assume $\theta = \eta$.

Hastie and Tibshirani (1990) propose fitting this model with *local scoring*, a method combining additive model backfitting with a Fischer scoring-type iteration. Because local scoring does not allow a closed-form representation of the estimators, even as an asymptotic approximation, we cannot directly apply the same approach of differentiating and approximating the GCV derivatives as in the previous two

sections. We will therefore derive the method for the local likelihood estimator of Kauermann and Opsomer (2000), since this estimator has a closed-form asymptotic approximation. Local scoring and local likelihood estimators are closely related in both an asymptotic and an algorithmic sense (see Kauermann and Opsomer, 2000 for details) and lead to almost identical fits in practice. Hence, even though the proposed bandwidth selection algorithm cannot be rigorously justified for local scoring, it can be applied for that case as well with no substantial modification in the computations, and should perform equally well.

We introduce the necessary notation. Let $\boldsymbol{V} = \mathrm{diag}(v_1, \ldots, v_n)$ be a diagonal matrix, where $v_i = -\partial^2 \log f(y|\eta)/(\partial\eta)^2 = \partial^2 b(\theta_i)/(\partial\theta_i)^2 \phi^{-1} = \mathrm{Var}(y_i|\eta_i)\phi^2$ and $\eta_i = \alpha + \sum_k \gamma_k(x_{ik})$. Here ϕ_i is the dispersion parameter corresponding to the ith observation. Let $\boldsymbol{\gamma}_k = (\gamma_k(x_{1k}), \ldots, \gamma_k(x_{nk}))^T$ and define the weighted smoothing matrix $\boldsymbol{S}_k$ with elements

$$\boldsymbol{S}_{k,ij} = (1,0)\boldsymbol{F}_{k,i}^{-1} w_{k,ij}(h_k)\boldsymbol{X}_{k,ij}$$

with $w_{k,ij}(h_k) = W\{(x_{kj} - x_{ki})/h_k\}$ for some kernel function $W(\cdot)$, $\boldsymbol{X}_{k,ij}^T = (1, x_{kj} - x_{ki})$ and $\boldsymbol{F}_{k,i} = \sum_j w_{k,ij} v_j \boldsymbol{X}_{k,ij}\boldsymbol{X}_{k,ij}^T$. Let $\boldsymbol{S}_k^* = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T\boldsymbol{V}/\sum_i v_i)\boldsymbol{S}_k$ denote the centered smoothing matrix (note that in the homoskedastic case, this centering adjustment reduces to that of the additive model in Section 3). The entries of $\boldsymbol{S}_{k,ij}$ can again be approximated by

$$\boldsymbol{S}_{k,ij} \approx K\left(\frac{x_{ik} - x_{jk}}{h_k}\right) \Big/ \{n h_k v_k(x_{ik}) f_k(x_{ik})\}, \tag{25}$$

where $K(\cdot)$ denotes a kernel of order 2 and $v_k(x_{ik}) = \int v\{\alpha + \sum_k \gamma_k(x_k)\} f_x(x_{-k}|x_{ik}) dx_{-k}$ is the conditional variance function given x_{ik}. Finally, let

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{S}_1^*\boldsymbol{V} & \cdots & \boldsymbol{S}_1^*\boldsymbol{V} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{S}_q^*\boldsymbol{V} & \boldsymbol{S}_q^*\boldsymbol{V} & \cdots & I \end{pmatrix}. \tag{26}$$

We refer the reader to Appendix A for the formal definition of the local likelihood estimators for the component functions in (1). The essential result from local likelihood estimation is that, similarly to equation (10), one can asymptotically derive a linear form of smoothing, given by

$$\boldsymbol{M}\hat{\boldsymbol{\gamma}}_\bullet \approx \boldsymbol{S}_\bullet^*(\boldsymbol{l}_\eta + \boldsymbol{V}\boldsymbol{\eta}) \tag{27}$$

with $\boldsymbol{l}_\eta = (l_{\eta,1}, \ldots, l_{\eta,n})^T$, $l_{\eta,i}$ the *score* of the ith observation, and $\boldsymbol{S}^*_\bullet = (\boldsymbol{S}^{*T}_1, \ldots, \boldsymbol{S}^{*T}_q)^T$. Since $g(\cdot)$ is assumed to be the natural link function the ith element of the score vector equals the residual $l_{\eta,i} = (y_i - \mu_i)/\phi_i$. From (27) it is possible to write the approximation

$$\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta} \approx \boldsymbol{Q}_+(\boldsymbol{l}_\eta + \boldsymbol{V}\boldsymbol{\eta}), \tag{28}$$

with $\boldsymbol{Q}_+ = \sum \boldsymbol{Q}_k$ and $\boldsymbol{Q}_k = \boldsymbol{M}^{kk}\boldsymbol{S}^*_k\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{S}^*_{-k}\boldsymbol{V})$ defined similar as in the previous section, but using the generalized smoother matrices as in (26). The approximation (28) justifies treating the local likelihood estimator as a weighted additive model.

The generalized additive model GCV criterion is now given by

$$\mathrm{GCV}(\boldsymbol{h}) \;=\; \frac{\widehat{\boldsymbol{l}}^T_\eta \boldsymbol{V}^{-1}\widehat{\boldsymbol{l}}_\eta}{n\{1 - \sum_k \mathrm{tr}(\boldsymbol{S}^*_k\boldsymbol{V})/n\}^2} \;=\; \frac{(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})^T\boldsymbol{\Phi}^{-1}\boldsymbol{V}^{-1}\boldsymbol{\Phi}^{-1}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})}{n\{1 - \sum_k \mathrm{tr}(\boldsymbol{S}^*_k\boldsymbol{V})/n\}^2} \tag{29}$$

where $\boldsymbol{\Phi} = \mathrm{diag}(\phi_1, \ldots, \phi_n)$ and $\widehat{\boldsymbol{l}}_\eta = \boldsymbol{\Phi}^{-1}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})$ with $\widehat{\boldsymbol{\mu}} = g(\widehat{\boldsymbol{\eta}})$. The degrees of freedom approximation in the denominator can be justified using the same reasoning as for the additive model criterion (11), and expression (29) reduces to (11) for the identity link and normal homoskedastic errors.

As in the previous sections, we propose a Newton-type procedure for efficiently minimizing this function. Note first that $\boldsymbol{S}_k$ and $\boldsymbol{V}$ in (29) both depend on the unknown parameters through $v_i, i = 1, \ldots, n$, so that plug-in estimates have to be used to calculate (29). The derivative of (29) is

$$\frac{\mathrm{GCV}(\boldsymbol{h})}{\partial h_k} \;=\; \frac{2}{n}\left(\frac{\widehat{\boldsymbol{l}}^T_\eta \boldsymbol{V}^{-1}\dfrac{\partial \widehat{\boldsymbol{l}}_\eta}{\partial h_k}}{\{1 - \sum_k \mathrm{tr}(\boldsymbol{S}^*_k)/n\}^2} + \frac{\widehat{\boldsymbol{l}}^T_\eta \boldsymbol{V}^{-1}\widehat{\boldsymbol{l}}^T_\eta \,\mathrm{tr}\left(\dfrac{\partial \boldsymbol{S}^*_k}{\partial h_k}\right)}{n\{1 - \sum_k \mathrm{tr}(\boldsymbol{S}^*_k)/n\}^3}\right), \tag{30}$$

where $\partial\widehat{\boldsymbol{l}}_\eta/\partial h_k = -\boldsymbol{\Phi}^{-1}\partial\widehat{\boldsymbol{\mu}}/\partial h_k = -\boldsymbol{V}\partial\widehat{\boldsymbol{\gamma}}_+/\partial h_k$.

Using the first order approximation $\widehat{\boldsymbol{\mu}} \approx \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{\Phi}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})$ and differentiating both sides of (27) as in (13) provides

$$\frac{\partial\widehat{\boldsymbol{\gamma}}_+}{\partial h_r} \;\approx\; \boldsymbol{P}_k\frac{\partial \boldsymbol{S}^*_k}{\partial h_k}\{\boldsymbol{l}_\eta + \boldsymbol{V}\widehat{\boldsymbol{\gamma}}_k\},$$

where $\boldsymbol{P}_k = (\boldsymbol{I} - \boldsymbol{S}^*_{-k}\boldsymbol{V})\boldsymbol{M}^{kk}$. The derivative $\partial\boldsymbol{S}^*_k/\partial h_k$ is again approximated as in (6) in Section 1 which in turn yields an estimator for $\partial\mathrm{GCV}(\boldsymbol{h})/\partial h_k$.

For the second order derivative, we obtain from (28) in first order the approximation

$$\mathrm{E}\{\mathrm{GCV}(\boldsymbol{h})\} \;\approx\; \mathrm{tr}\{\boldsymbol{\Phi} - 2\boldsymbol{Q}_+\boldsymbol{V}\boldsymbol{\Phi} + \boldsymbol{Q}_+^T\boldsymbol{V}\boldsymbol{Q}_+\boldsymbol{V}\boldsymbol{\Phi}\} + \boldsymbol{b}_+^T\boldsymbol{V}\boldsymbol{b}_+/n$$

where $\boldsymbol{b}_+ = E(\widehat{\boldsymbol{\eta}}_+ - \boldsymbol{\eta})$. As in the previous section, we approximate the term $\mathrm{tr}\{2\boldsymbol{Q}_+\boldsymbol{V}\boldsymbol{\Phi} - \boldsymbol{Q}_+^T\boldsymbol{V}\boldsymbol{Q}_+\boldsymbol{V}\boldsymbol{\Phi}\}$ by $\sum_k \mathrm{tr}\{2\boldsymbol{S}_+^{*}\boldsymbol{V}\boldsymbol{\Phi} - \boldsymbol{S}_+^{*^T}\boldsymbol{V}\boldsymbol{S}_+\boldsymbol{V}\boldsymbol{\Phi}\}$. Using the results in Kauermann and Opsomer (2000), (18) can be shown to hold approximately in the generalized additve model as well. Hence, the bias decomposition from before provides $\boldsymbol{b}_+ = \sum_k \tilde{\boldsymbol{b}}_k$ with $\tilde{\boldsymbol{b}}_k = \boldsymbol{P}_k\boldsymbol{b}_{k|-k}$, with $\boldsymbol{b}_{k|-k} = E(\widehat{\boldsymbol{\gamma}}_{k|-k} - \boldsymbol{\gamma}_k)$ as 'oracle' bias. Hence, as in (20) and (21) one obtains

$$\begin{aligned}
\frac{\partial^2 E\{GCV(h)\}}{\partial h_k^2} &\approx \frac{1}{nh_k^2}\{2\sigma^2\mathrm{tr}(2\boldsymbol{S}_k^{*^T}\boldsymbol{V} - \boldsymbol{S}_k^{*^T}\boldsymbol{V}\boldsymbol{S}_k^{*}\boldsymbol{V}\boldsymbol{\Phi}) \qquad (31)\\
&\quad +12\tilde{\boldsymbol{b}}_1^T\boldsymbol{V}\tilde{\boldsymbol{b}}_k + 4\tilde{\boldsymbol{b}}_k^T\boldsymbol{V}\tilde{\boldsymbol{b}}_{-k}\}\\
\frac{\partial^2 E\{GCV(h)\}}{\partial h_r \partial h_k} &\approx \frac{1}{n}\left(\frac{\partial^2 \boldsymbol{b}_+^T\boldsymbol{V}\boldsymbol{b}_+}{\partial h_r \partial h_r}\right) \approx \frac{8}{nh_rh_k}\tilde{\boldsymbol{b}}_r^T\boldsymbol{V}\tilde{\boldsymbol{b}}_k \qquad (32)
\end{aligned}$$

In complete analogy to the previous section, we can now develop a Newton algorithm for bandwidth selection. The only difference occuring here is that quantities involved depend on the unknown variance matrix $\boldsymbol{V}$. One therefore has to replace $\boldsymbol{V}$ in each step of the algorithm by a plug in estimate using the current estimate with bandwidth h_t. Moreover, using approximations as above, a plug-in estimate for the bias is obtained from $\widehat{\boldsymbol{b}}_k = \boldsymbol{Q}_k\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\gamma}}_k$ so that with using (18) and (19) and plug-in estimates for $\boldsymbol{M}$ and $\boldsymbol{P}$ a plug-in estimates for $\tilde{\boldsymbol{b}}_k$ results.

Example:
We apply the procedure to binary response data. The outcome variable describes the occurrence of chronic bronchitis (y=1 for yes, $y = 0$ for no) at workers employed in a mechanical engineering plant in Munich, Germany (see Küchenhoff and Carroll, 2000 for a previous analysis of the data). As explanatory variables we consider the average dust concentration at the worker's workplace, x_1, and the exposure time x_2. To control for the effect of smoking we base our analysis on smokers only. The data are available from the data server at `http://www.stat.uni-muenchen.de`.

We fit the model $P(y = 1|x_1, x_2) = \mathrm{logit}^{-1}\{\alpha + \gamma_1(x_1) + \gamma_2(x_2)\}$ and start the Newton procedure with $h = (h_1, h_2)$ chosen as the empirical standard deviations of x_1 and x_2, respectively. As a second set of starting values we use 1/4 of the empirical

standard deviations. A step size reduction was used for the first setting only, to cope for the effects of oversmoothing. The steps of the routine are plotted in Figure 8 where we show $GCV(h_1, h_2)$ calculated on a grid of points. Obviously, the Newton procedure behaves satisfactory also in the binary case by moving quickly towards the minimum of GCV. This holds for both starting values. The final estimates are shown in Figure 9. In particular, a saturation of the exposure time after about 25 years becomes visible.

A Appendix: Local Likelihood Backfitting

We summarize some of the main ideas of Kauermann and Opsomer (2000) here. Based on (24), the log-likelihood contribution of the jth observation as a function of η is

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

We define the *score* for the jth observation as $l_{\eta,j} = \partial l_j(\eta)/\partial\eta$ evaluated at the true parameter value for the jth observation. The local likelihood estimators $\widehat{\gamma}_k(x_{ki}), k = 1, \ldots, q, i = 1, \ldots, n$ are defined as $\widehat{\gamma}_k(x_{ki}) = (1, 0)\boldsymbol{\beta}_{ki}$, where the $\boldsymbol{\beta}_{ki}$ are the solutions to the system of non-linear score equations

$$\sum_{j=1}^{n} w_{k,ij}\boldsymbol{X}_{k,ij}l_{\eta,j}(\boldsymbol{X}_{k,ij}^T\boldsymbol{\beta}_{ki} + \widehat{\eta}_{-k,j}) = \mathbf{0} \tag{33}$$

for $i = 1, \ldots, n$, $k = 1, \ldots, q$, subject to identifiability constraints for the $\gamma_k(\cdot)$.

Let $\widehat{\boldsymbol{\gamma}}_{k|-k}$ represent the oracle estimator vector for the kth component function evaluated at the observation points, and let $\boldsymbol{l}_\eta = (l_{\eta,1}, \ldots, l_{\eta,n})^T$. (Kauermann and Opsomer, 2000) show that

$$\widehat{\boldsymbol{\gamma}}_{k|-k} \approx \boldsymbol{S}_k^*\boldsymbol{l}_\eta + \boldsymbol{S}_k^*\boldsymbol{V}\boldsymbol{\gamma}_k \tag{34}$$

and provide the following relationship between the oracle and the full local likelihood estimator

$$\boldsymbol{M}(\widehat{\boldsymbol{\gamma}}_\bullet - \boldsymbol{\gamma}_\bullet) \approx \widehat{\boldsymbol{\gamma}}_{\bullet|-\bullet} - \boldsymbol{\gamma}_\bullet \tag{35}$$

where as above $\boldsymbol{\gamma}_\bullet = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_q^T)^T$, $\boldsymbol{\gamma}_{\bullet|-\bullet} = (\boldsymbol{\gamma}_{1|-1}^T, \ldots, \boldsymbol{\gamma}_{n|-n}^T)^T$ and $\boldsymbol{M}$ as given in (26). Approximations (34) and (35) directly lead to (27).

References

Buja, A., T. J. Hastie, and R. J. Tibshirani (1989). Linear smoothers and additive models. *Annals of Statistics 17*, 453–555.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31*, 377–403.

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

Gu, C. and G. Wahba (1991). Minimizing gcv/gml scores with multiple smoothing parameters via the newton method. *SIAM Journal on Scientific and Statistical Computing 12*, 383–398.

Hastie, T. J. and R. J. Tibshirani (1987). Local likelihood estimation. *Journal of the American Statistical Association 82*, 559–567.

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

Kauermann, G. and Opsomer, J. (2000). Local likelihood estimation in generalized additive models. Submitted.

Küchenhoff, H. and R.J. Carroll (1997) Segmented Regression with Errors in Predictors: Semi-Parametric and Parametric Methods. *Stat. in Med.16*, 169–188.

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

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.

O'Sullivan, F., B. Yandell, and W. Raynor (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association 81*, 96–103.

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

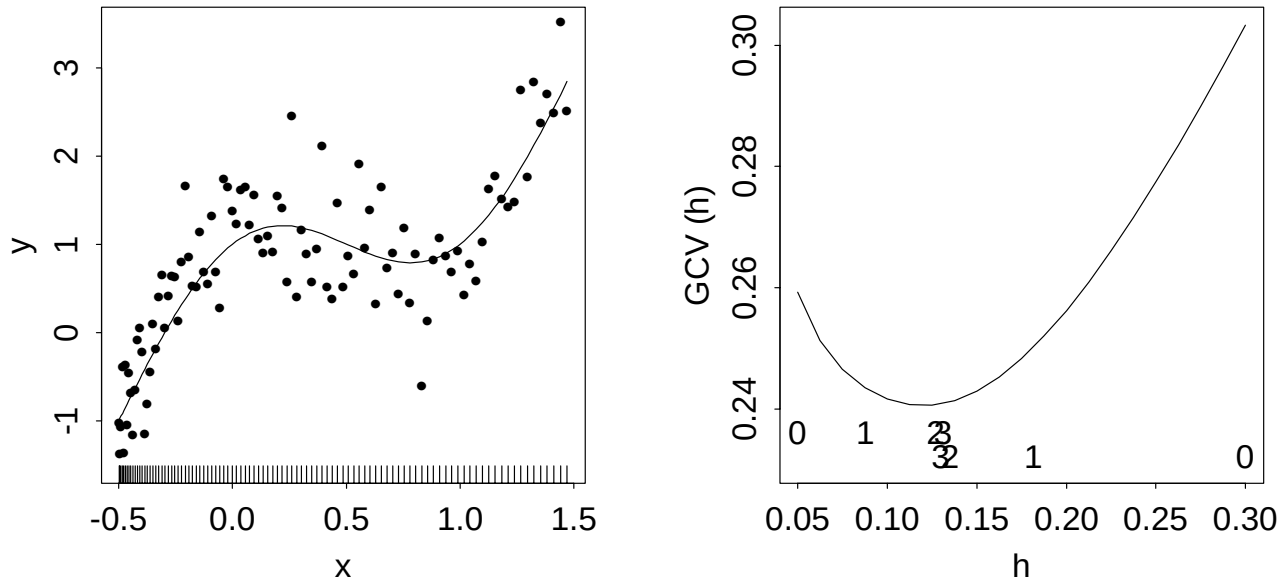

Figure 1: Simulated data with true curve and design density as ticks (left plot), $GCV(h)$ with steps of the Newton procedure, indicated by $0, \ldots 3$, for two different starting values.

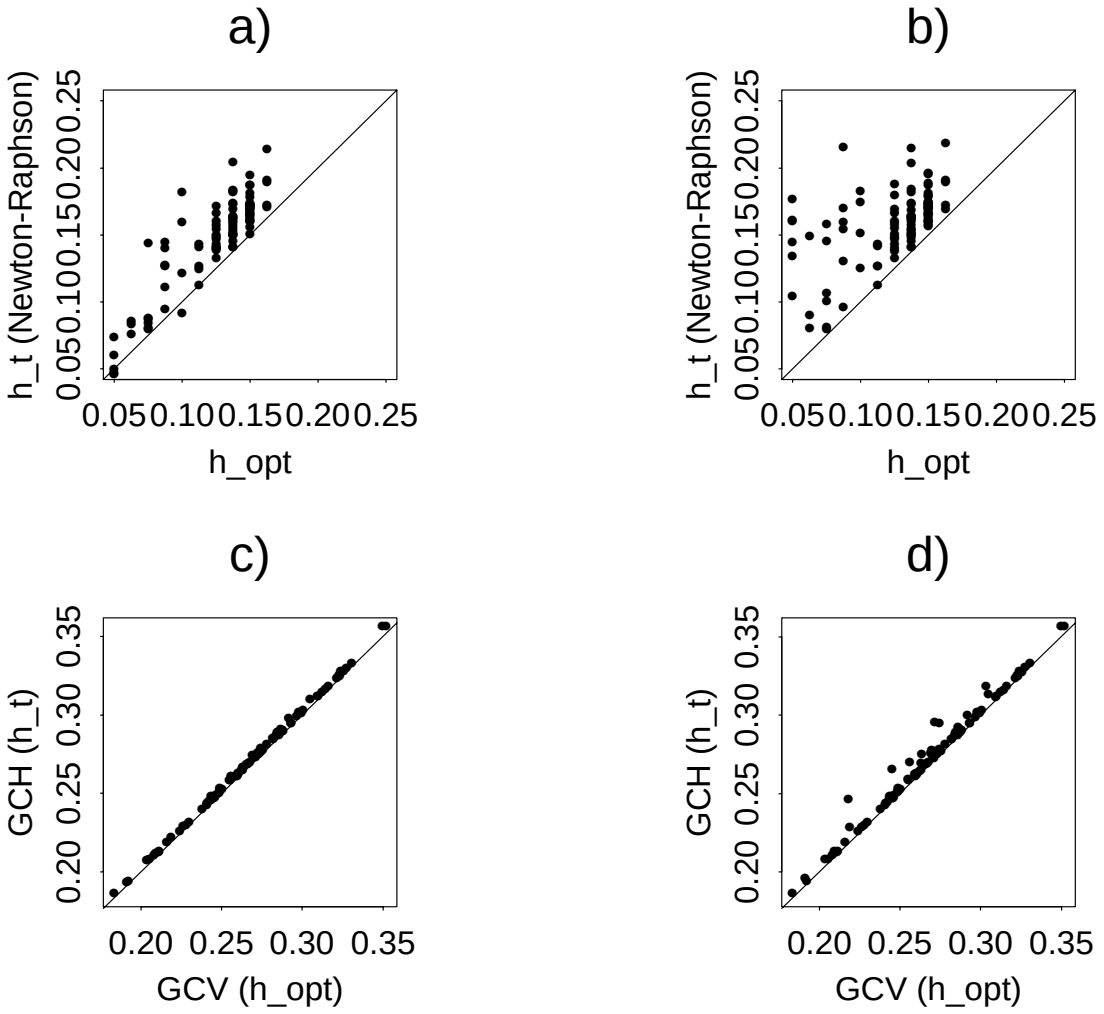

Figure 2: Bandwidth h_{opt} which minimizes $GCV(h)$ plotted against h_t resulting from the 5-th step of the Newton algorithm for starting values a) $h_0 = 0.05$ and b) $h_0 = 0.3$. Lower two plots show $GCV(h_{opt})$ plotted against $GCV(h_t)$, for starting values c) $h_0 = 0.05$ and d) $h_0 = 0.3$.

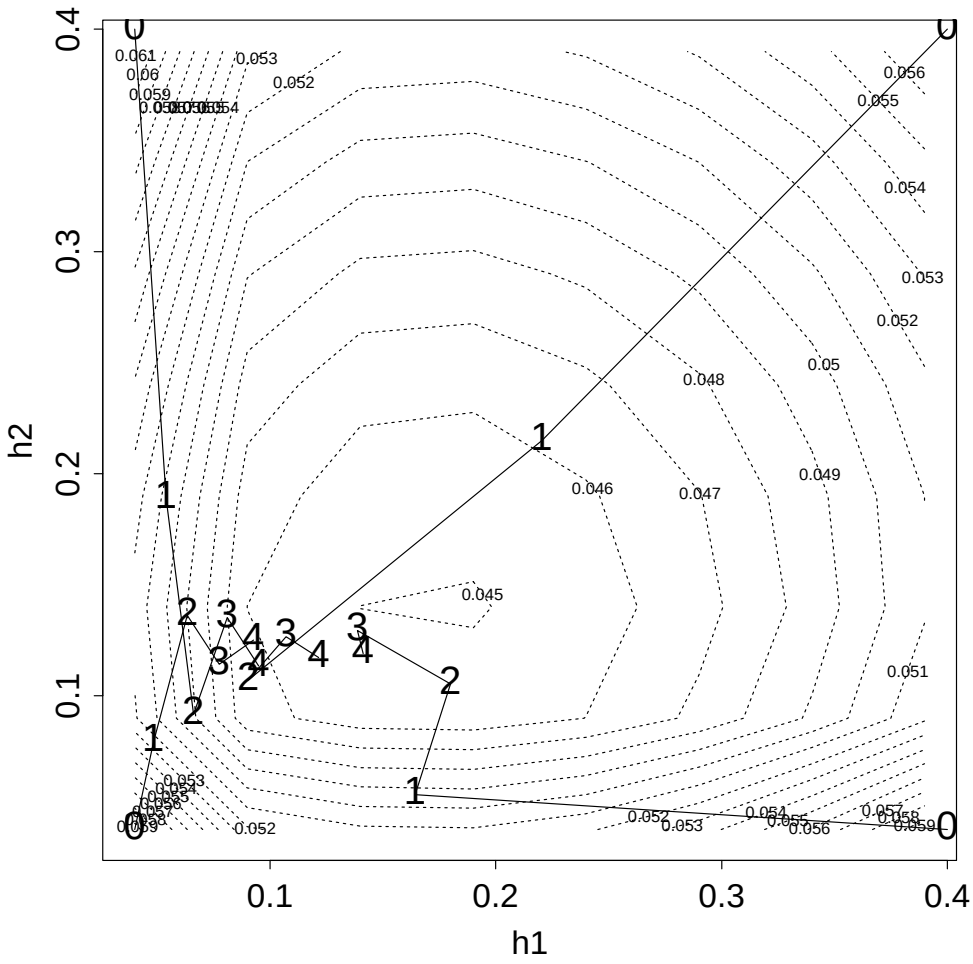

Figure 3: Generalized Cross Validation Function $GCV(h)$ and Newton steps, denoted by 0 to 4, for 4 different starting points.

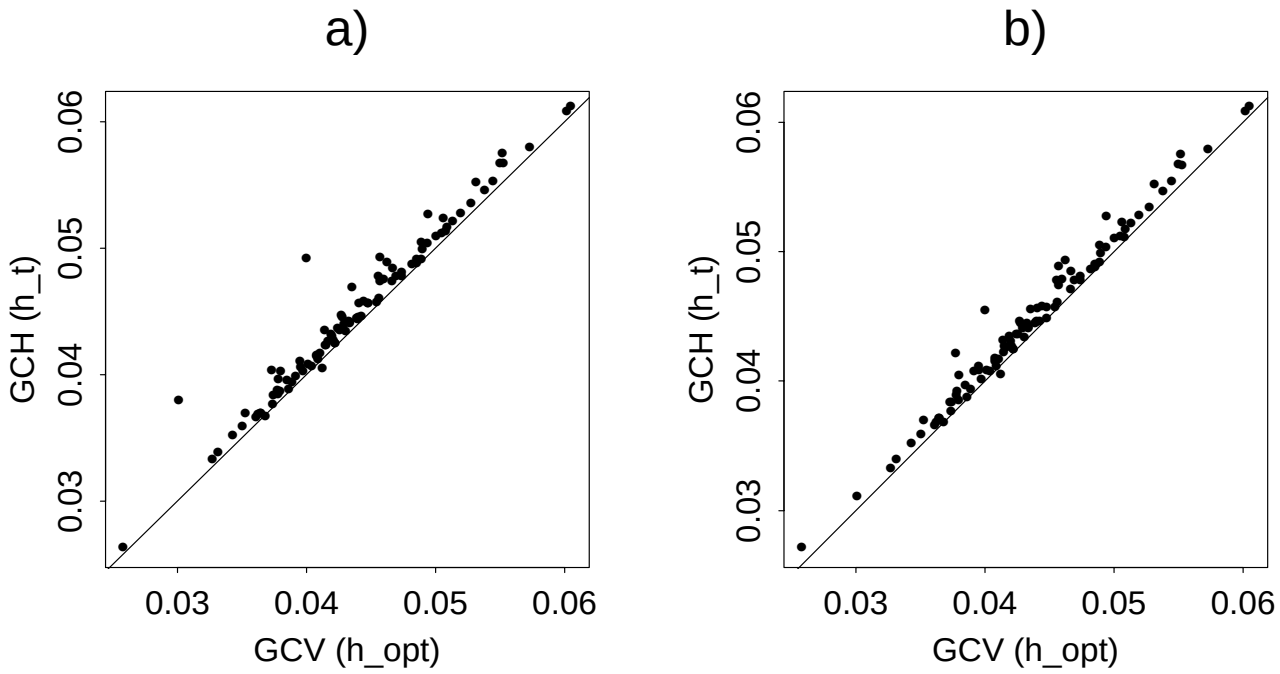

Figure 4: $GCV(h_{opt})$ plotted against $GCV(h_t)$ with h_t as a) 3-rd and b) 5-th step of the Newton algorithm.

	$p=1$		$p=2$	
	$q=0$	$q=1$	$q=0$	$q=1$
$\rho=0$	.995	.998	.996	.998
$\rho=0.5$	.990	.996	.995	.996

Table 1: Empirical correlation $cor\{GCV(\widehat{h}), GCV(h_{opt})\}$ for different simulation settings each based on 150 replicates

step	GCV	vh	wind	humidity	temp	ibh	dpg	ibt	vis	doy
0	17.65	105	2.29	19.9	14.4	1803	35.7	76.7	79.3	106.1
1	16.16	.	.	.	.	.	.	.	.	.
2	15.75	.	.	.	.	.	.	.	.	.
3	15.59	473	2.79	47.6	9.50	4905	37.5	115	69.7	45.6
4	15.48	.	.	.	.	.	.	.	.	.
5	15.41	988	3.00	69.5	8.46	7623	38.9	133	66.9	41.6

Table 2: Value of $GCV(h)$ for ozone data and the first 5 steps of the Newton procedure.

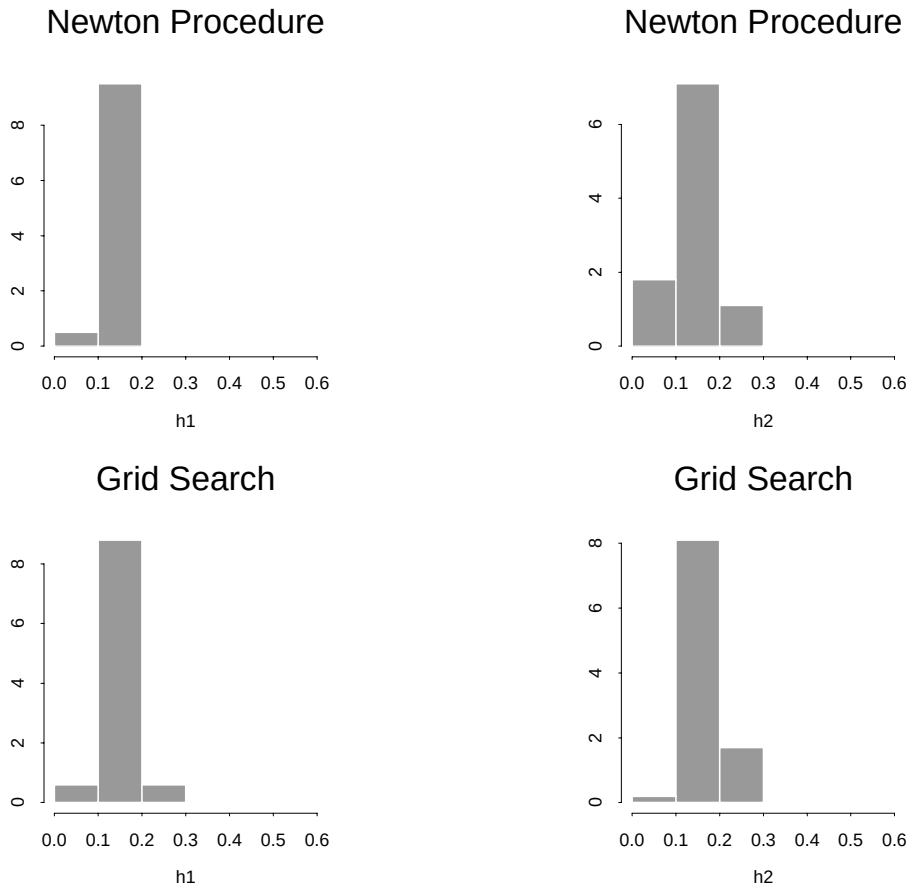

Figure 5: Selected Bandwidth for $p = 2$, $q = 1$, $\rho = 0$.

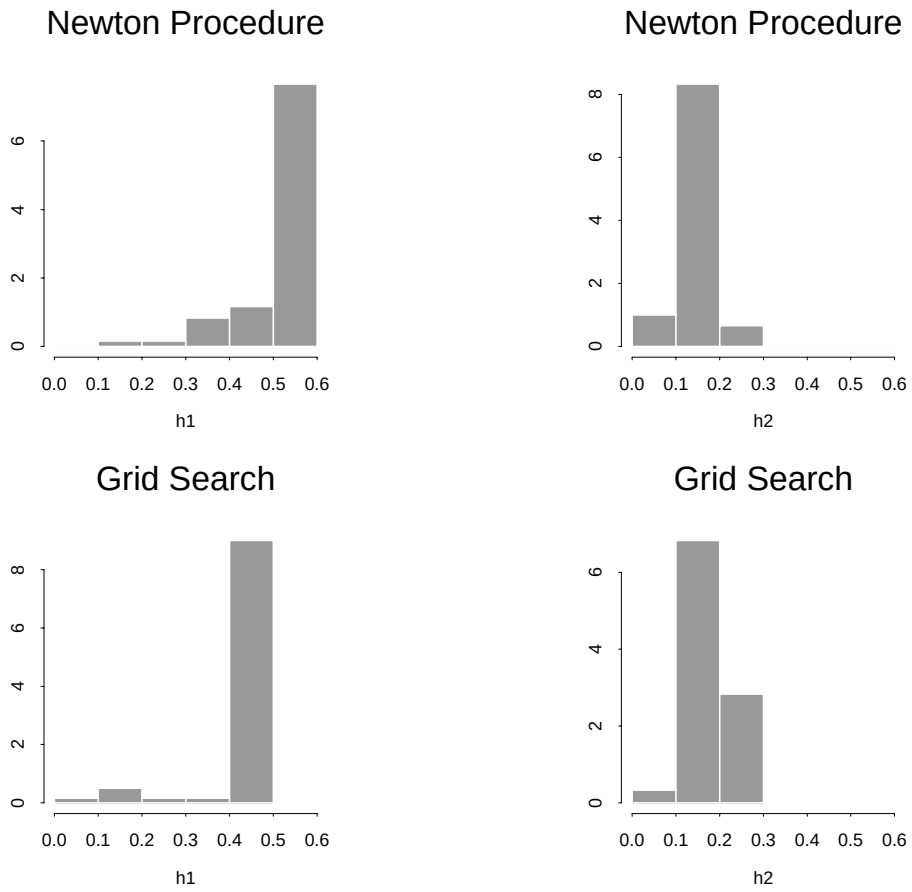

Figure 6: Selected Bandwidth for $p = 1$, $q = 1$, $\rho = 0$ (Bandwidth >0.5 are set displayed in group $[0.5, 0.6]$).

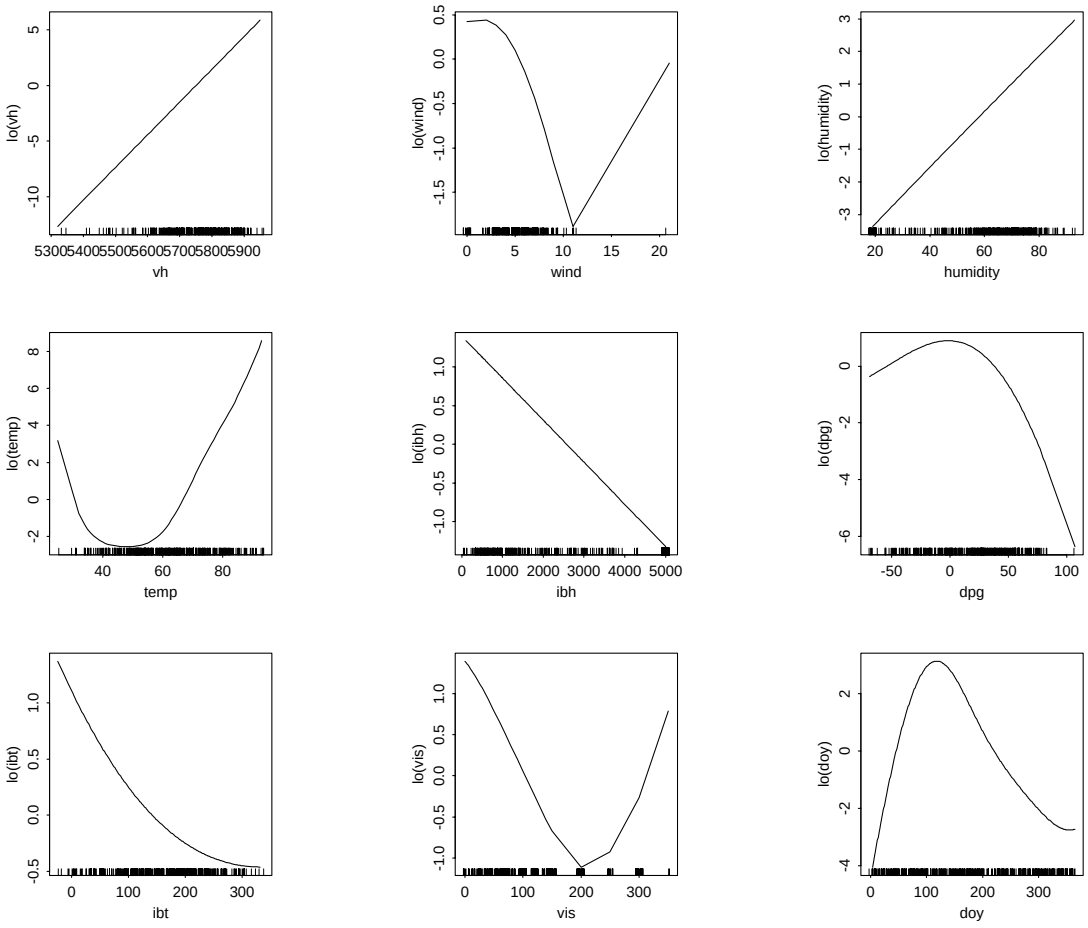

Figure 7: Fitted Generalized Additive Model for ozone data.

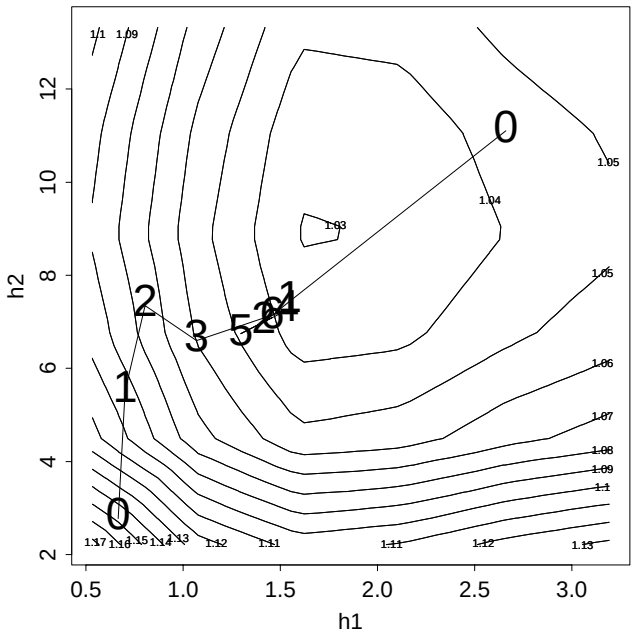

Figure 8: $GCV(h_1, h_2)$ for brochitis data.

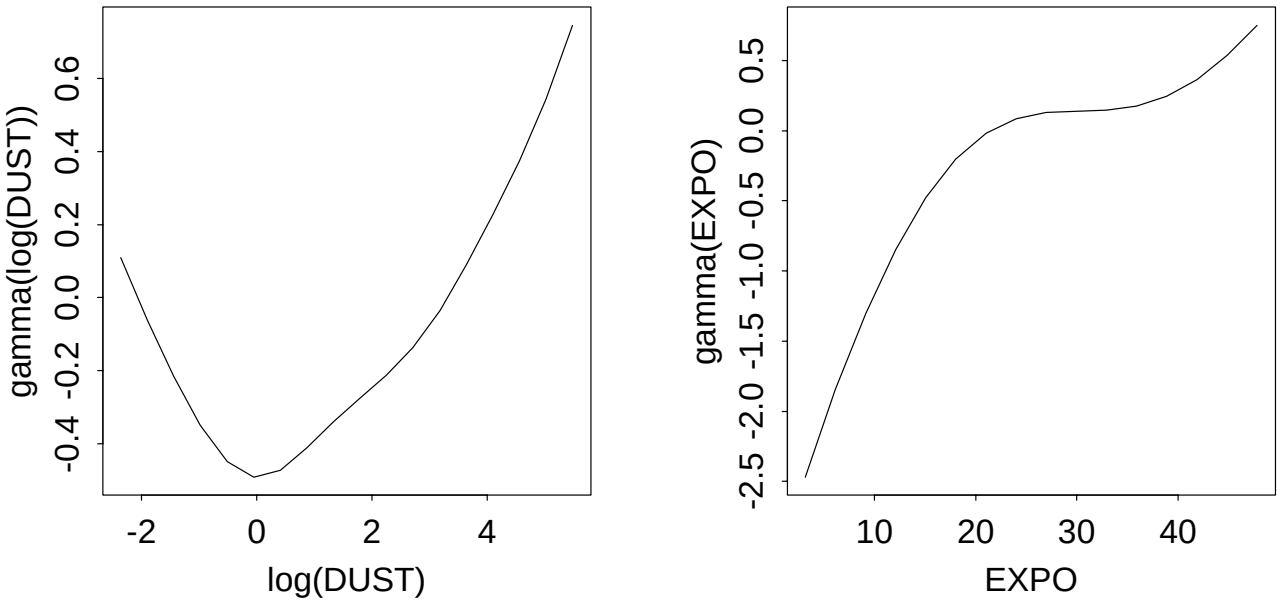

Figure 9: Fitted Generalized Additive Model for bronchitis data.