ELSEVIER

Journal of Statistical Planning and Inference 103 (2002) 455–470

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Some theory for penalized spline generalized additive models

M. Aerts[a], G. Claeskens[b], M.P. Wand[c,*]

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

[b]*Department of Statistics, Texas A&M University, College Station, TX 77843, USA*

[c]*Department of Biostatistics, School of Public Health, Harvard University, 665 Huntington Avenue, Boston, MA 02115, USA*

Abstract

Generalized additive models have become one of the most widely used modern statistical tools. Traditionally, they are fit through scatterplot smoothing and the backfitting algorithm. However, a more recent development is the direct fitting through the use of low-rank smoothers (Hastie, J. Roy. Statist. Soc. Ser. B 58 (1996) 379). A particularly attractive example of this is through use of penalized splines (Marx and Eilers, Comput. Statist. Data Anal. 28 (1998) 193). Such an approach has a number of advantages, particularly regarding computation. In this paper, we exploit the explicitness of penalized spline additive models to derive some useful and revealing theoretical approximations. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Asymptotic approximation; Automatic smoothing parameter selection; Degrees of freedom; Nonparametric regression

1. Introduction

Generalized additive models (GAM) are among the most practically used modern statistical techniques. Examples of their use in applications includes political science (Beck and Jackman, 1998), economics (Linton and Härdle, 1996) and environmental epidemiology (Schwartz, 1994). The main catalysts for this widespread use by practitioners is the exemplary monograph on the topic, Hastie and Tibshirani (1990), and the availability of the function `gam()` in the `S-PLUS` language for fitting such models (see e.g. Chambers and Hastie, 1991).

One aspect of GAM that has been slow to develop is the statistical properties of the estimation strategies for fitting such a model. Perhaps the main reason for this is the nonexplicit nature of the most common type of GAM fitting procedure: *backfitting* combined with *local scoring* (Hastie and Tibshirani, 1990). This nonexplicitness

[*] Corresponding author.
E-mail address: mwand@hsph.harvard.edu (M.P. Wand).

PII: S0378-3758(01)00237-3

appears to be the main motivation for the *marginal integration* approach to additive model fitting (Linton and Nielsen, 1995) and a sophisticated theory now exists for this strategy (e.g. Fan et al., 1998). The statistical properties of additive models based on backfitting have since been derived by Opsomer and Ruppert (1997), Opsomer (2000) and Wand (1999a). Claeskens and Aerts (2000) examine extensions of this theory to generalized additive models.

An attractive alternative to backfitting and marginal integration is direct fitting based on low-rank smoothers (Hastie, 1996). Marx and Eilers (1998) demonstrate how this can be achieved using penalized splines, also known as P-splines. This approach has chiefly been motivated by computational expediency. However, the directness of the method also means that the estimator has an explicit form. This paper exploits this fact. We derive simple closed form approximations to risk and degrees of freedom of the estimator and its components, not just for ordinary additive models but for GAM.

In Section 2, the ordinary additive model is treated. Recursion formulae for the overall fit in terms of sub-models fits are developed. From this an asymptotic approximation of the overall fit is derived, which forms the basis of the risk and degrees of freedom approximations. The results are both useful and revealing. For example, they can be used to provide rough starting values for the smoothing parameters. They also provide some backup for commonly used degrees of freedom approximations.

Section 3 repeats this for more complicated settings: semiparametric models, generalized additive models and additive generalized estimating equations.

2. Additive models

In this section, we will study the penalized regression spline estimators in the standard additive models framework. In these models the response Y_i $(i=1,\dots,n)$ depends in an additive way on the d covariates, $x_{1i},\dots,x_{di}$ through arbitrary univariate functions f_j,

$$Y_i = \beta_0 + \sum_{j=1}^{d} f_j(x_{ji}) + \varepsilon_i. \tag{1}$$

It is assumed that the errors are independent and identically distributed with mean zero, variance σ^2 and are independent of the covariates. Each of the functions f_j will be estimated by a degree p_j penalized spline estimator with smoothing parameter α_j.

Note that, due to identifiability requirements, the f_j in (1) are defined only up to an additive constant. Therefore, they can be replaced by $f_j(x_{ji}) - 1/n \sum_{i=1}^{n} f_j(x_{ji})$. This makes β_0 orthogonal to the f_j's. The estimate of β_0, $\hat{\beta}_0 = \bar{Y}$, is independent of the x_{ji}'s. Since $\bar{Y}$ can be subtracted from the Y_i*'s* without affecting the model fitting we can assume, without loss of generality, that $\beta_0 = 0$. This convention will be made from here onwards.

For any $(n \times m)$ matrix $\mathbf{C}$, denote by $\mathbf{C}^*$ the centered matrix

$$\mathbf{C}^* = (\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathsf{T}}/n)\mathbf{C},$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is an $n \times 1$ column of ones. The notation $\mathbf{0}_{p\times q}$ is shorthand for the $p \times q$ matrix with zero in each position. A similar definition, with one instead of zero, applies to $\mathbf{1}_{p\times q}$. For any real number u, we define $u_+ = \max(0, u)$.

In matrix notation we can write the model as

$$\mathbf{Y} = \sum_{j=1}^{d} \mathbf{f}_j + \boldsymbol{\varepsilon}, \quad \mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}, \tag{2}$$

where $\mathbf{Y} = [Y_1, \ldots, Y_n]^{\mathsf{T}}$, $\mathbf{f}_j = [f_j(x_{j1}), \ldots, f_j(x_{jn})]^{\mathsf{T}}$ and $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]^{\mathsf{T}}$. Let $\mathbf{f} = \sum_{j=1}^{d} \mathbf{f}_j$.

The penalized spline estimate of $\mathbf{f}$ is

$$\hat{\mathbf{f}}_{\boldsymbol{\alpha}} = \mathbf{G}_{\boldsymbol{\alpha}} \mathbf{Y} \quad \text{where} \quad G_{\boldsymbol{\alpha}} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{A}_{\boldsymbol{\alpha}})^{-1}\mathbf{X}^{\mathsf{T}}$$

with centred design matrix

$$\mathbf{X} = [\mathbf{X}_1^* \cdots \mathbf{X}_d^*], \quad \mathbf{A}_{\boldsymbol{\alpha}} = \underset{1 \leqslant j \leqslant d}{\mathrm{blockdiag}}(\alpha_j \mathbf{D}_j),$$

$$\mathbf{X}_j = \begin{bmatrix} x_{j1} & \cdots & x_{j1}^{p_j} & (x_{j1} - \kappa_{j1})_+^{p_j} & \cdots & (x_{j1} - \kappa_{jK_j})_+^{p_j} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{jn} & \cdots & x_{jn}^{p_j} & (x_{jn} - \kappa_{j1})_+^{p_j} & \cdots & (x_{jn} - \kappa_{jK_j})_+^{p_j} \end{bmatrix}, \quad \mathbf{D}_j = \mathrm{diag}(\mathbf{0}_{p_j \times 1}, \mathbf{1}_{K_j \times 1}),$$

$\kappa_{j1}, \ldots, \kappa_{jK_j}$ $(j = 1, \ldots, d)$ is a set of knots in the jth direction and p_j is the degree of the splines used in direction j. Note that $\mathbf{X}_j$ is a basis for the set of piecewise continuous p_jth degree polynomials with knots, or join points, at the κ_{jk} and is sometimes referred to as the truncated polynomial basis. The estimator can be reformulated in terms of other bases such as the B-spline basis (Eilers and Marx, 1996) and the Demmler–Reinsch basis (Nychka and Cummins, 1996). These alternative bases have better numerical properties, so they are preferable for computation. But for formulation and theory the simplicity of the truncated polynomial basis is preferred.

The knots are usually taken to be relatively "dense" among the observations in an attempt to capture the curvature in $\mathbf{f}_j$. A reasonable allocation rule is one knot for every 4–5 observations, up to a maximum of about 40 knots. Ruppert and Carroll (2000) describe an algorithm for choosing the number of knots, and demonstrate its effectiveness through simulation. Subscripts denoting the dependence of matrices on the smoothing vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_d]^{\mathsf{T}}$, the vector of penalty parameters, will be omitted, unless $\boldsymbol{\alpha} = \mathbf{0}$, in which case a subscript $\mathbf{0}$ will be used.

For any $j = 1, \ldots, d$, the full smoother matrix $\mathbf{G}$ can be decomposed as

$$\mathbf{G} = \mathbf{G}_j + \mathbf{G}_{[-j]},$$

where

$$\mathbf{G}_j = [\mathbf{0} \cdots \mathbf{0}\ \mathbf{X}_j\ \mathbf{0} \cdots \mathbf{0}](\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{A})^{-1}\mathbf{X}^{\mathsf{T}}, \tag{3}$$

$$\mathbf{G}_{[-j]} = [\mathbf{X}_1 \cdots \mathbf{X}_{j-1}\ \mathbf{0}\ \mathbf{X}_{j+1} \cdots \mathbf{X}_d](\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{A})^{-1}\mathbf{X}^{\mathsf{T}}. \tag{4}$$

The corresponding additive component fits are

$$\hat{\mathbf{f}}_j = \mathbf{G}_j\mathbf{Y} \quad \text{and} \quad \hat{\mathbf{f}}_{[-j]} = \mathbf{G}_{[-j]}\mathbf{Y}.$$

In the model with only covariate x_j, the regression spline smoother matrix is

$$\mathbf{S}_j = \mathbf{X}_j(\mathbf{X}_j^{\mathsf{T}}\mathbf{X}_j + \mathbf{A}_j)^{-1}\mathbf{X}_j^{\mathsf{T}}. \tag{5}$$

In the model with all covariates except x_j, the regression spline smoother matrix is

$$\mathbf{S}_{[-j]} = \mathbf{X}_{[-j]}\{\mathbf{X}_{[-j]}^{\mathsf{T}}\mathbf{X}_{[-j]} + \mathbf{A}_{[-j]}\}^{-1}\mathbf{X}_{[-j]}^{\mathsf{T}}, \tag{6}$$

where

$$\mathbf{X}_{[-j]} = [\mathbf{X}_1 \cdots \mathbf{X}_{j-1}\mathbf{X}_{j+1} \cdots \mathbf{X}_d] \quad \text{and} \quad \mathbf{A}_{[-j]} = \underset{1 \leqslant i \leqslant d, i \neq j}{\text{blockdiag}}(\alpha_i\mathbf{D}_i).$$

The following recursive formula, which we call Result 1, is an important stepping stone towards obtaining approximations in penalized spline additive models. For any covariate x_j ($j = 1, \ldots, d$), the result allows one to write the full smoother matrix $\mathbf{G}$ in terms of the design matrix $\mathbf{X}_j$ and the smoother matrix in the sub-model corresponding to deletion of the jth covariate.

Result 1. *For any* $j = 1, \ldots, d$,

$$\mathbf{G} = \mathbf{S}_{[-j]} + (\mathbf{I} - \mathbf{S}_{[-j]})\mathbf{G}_j \tag{7}$$

and

$$\begin{aligned} \mathbf{G}_j &= \mathbf{X}_j\{\mathbf{X}_j^{\mathsf{T}}(\mathbf{I} - \mathbf{S}_{[-j]})\mathbf{X}_j + \mathbf{A}_j\}^{-1}\mathbf{X}_j^{\mathsf{T}}(\mathbf{I} - \mathbf{S}_{[-j]}) \\ &= \mathbf{S}_j[\mathbf{I} - \mathbf{X}_{[-j]}\{\mathbf{X}_{[-j]}^{\mathsf{T}}(\mathbf{I} - \mathbf{S}_j)\mathbf{X}_{[-j]} + \mathbf{A}_{[-j]}\}^{-1}\mathbf{X}_{[-j]}^{\mathsf{T}}(\mathbf{I} - \mathbf{S}_j)], \end{aligned} \tag{8}$$

if the inverse matrices exist.

The derivation of this result is provided in the Appendix.

To construct the additive model smoother matrix $\mathbf{G}_j$, it is sufficient to know the smoother matrix $\mathbf{S}_{[-j]}$ and the design matrix $\mathbf{X}_j$, or, using the equivalent expression (8), the univariate smoother matrix $\mathbf{S}_j$ and the design matrix $\mathbf{X}_{[-j]}$ corresponding to the $d-1$ other covariates. Hence, this recursive formula shows how the full smoother matrix $\mathbf{G}$ can be built from lower dimensional pieces. Eq. (7) is comparable to the result of Lemma 2.1 in Opsomer (2000), where, for local polynomial estimators using backfitting, a similar identity can be obtained.

Using formula (7), we obtain:

Result 2. *Assuming that* $\boldsymbol{\alpha} \to \mathbf{0}$ we have

$$\mathbf{G} = \mathbf{G}_{\mathbf{0}} - \sum_{j=1}^{d} \alpha_j\tilde{\mathbf{B}}_j + o\left(\sum_{j=1}^{d} \alpha_j\tilde{\mathbf{B}}_j\right), \tag{9}$$

where

$$\tilde{\mathbf{B}}_j = \tilde{\mathbf{X}}_j(\tilde{\mathbf{X}}_j^{\mathsf{T}}\tilde{\mathbf{X}}_j)^{-1}\mathbf{D}_j(\tilde{\mathbf{X}}_j^{\mathsf{T}}\tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^{\mathsf{T}} \quad \textit{and} \quad \tilde{\mathbf{X}}_j = (\mathbf{I} - \mathbf{S}_{\mathbf{0},[-j]})\mathbf{X}_j,$$

provided that all inverse matrices exist.

The matrix $\mathbf{G_0}$ is the usual "hat" matrix for unpenalized models ($\boldsymbol{\alpha} = \mathbf{0}$). The matrix $\tilde{\mathbf{X}}_j$ is obtained by projecting $\mathbf{X}_j$ orthogonal to the subspace determined by $\mathbf{X}_{[-j]}$. Details on the derivation of Result 2 can be found in the Appendix.

In the next two subsections, we show how approximation (9) can be used to aid practical implementation of additive models.

2.1. Approximation of the risk

Our first application of the results of the preceding section is to approximate the risk in an additive model. Such approximations have the advantage of being simpler to optimize and can, perhaps, aid the practical selection of the smoothing parameters. For convenience, we will work with the mean average squared error (MASE)

$$\text{MASE}(\hat{\mathbf{f}}) = \frac{1}{n} E||\hat{\mathbf{f}} - \mathbf{f}||^2.$$

This can be decomposed into the average variance plus the average squared bias and then simplified to give:

$$\text{MASE}(\hat{\mathbf{f}}) = \frac{\sigma^2}{n}\,\text{tr}(\mathbf{G}^2) + \frac{1}{n}||(\mathbf{G} - \mathbf{I})\mathbf{f}||^2.$$

Result 3 follows from this expression and (9):

Result 3. *The leading terms in the asymptotic expansion of* MASE($\hat{\mathbf{f}}$) *as* $\boldsymbol{\alpha} \to \mathbf{0}$ *are*

$$\text{AMASE}(\hat{\mathbf{f}}) \equiv \frac{1}{n}\left[\sigma^2\left\{\sum_{j=1}^{d}(p_j + K_j) - 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{q}\right\} + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Q}\boldsymbol{\alpha}\right],$$

where the entries of $\mathbf{q}$ ($d \times 1$) *and* $\mathbf{Q}$ ($d \times d$) *are* $\mathbf{q}_j = \text{tr}(\tilde{\mathbf{B}}_j)$, *and*

$$\mathbf{Q}_{jj'} = (\tilde{\mathbf{B}}_j\mathbf{f})^{\mathrm{T}}(\tilde{\mathbf{B}}_{j'}\mathbf{f}) + \sigma^2\,\text{tr}(\tilde{\mathbf{B}}_j\tilde{\mathbf{B}}_{j'}).$$

The AMASE-*optimal smoothing parameters are therefore given by*

$$\boldsymbol{\alpha}_{\text{AMASE}} = \sigma^2\mathbf{Q}^{-1}\mathbf{q}. \tag{10}$$

An application of this result is depicted in Fig. 1. It shows the result of applying (10) to the Californian air pollution data from Breiman and Friedman (1985), and used for illustratory purposes by Hastie (1996). Of course (10) requires knowledge of $\mathbf{f}$ and σ^2, so preliminary estimates of those were plugged in. These were obtained using blockwise quadratic fits as suggested by Härdle and Marron (1995) and Ruppert et al. (1995). As in the latter reference, Mallows' C_p was used to choose among all blockwise quadratic fits with 4 or less blocks. Six knots, equally spaced with the quantiles, were used for each function, and the estimate of $\boldsymbol{\alpha}_{\text{AMASE}}$ was (0.00011, 0.000015, 0.000088).

Comparison of Fig. 1 with Fig. 5 of Hastie (1996) shows the estimated functions for `Dagget pressure gradient` and `Inversion base temperature` being roughly comparable. However, that for `Inversion base height` is quite a bit more wiggly.

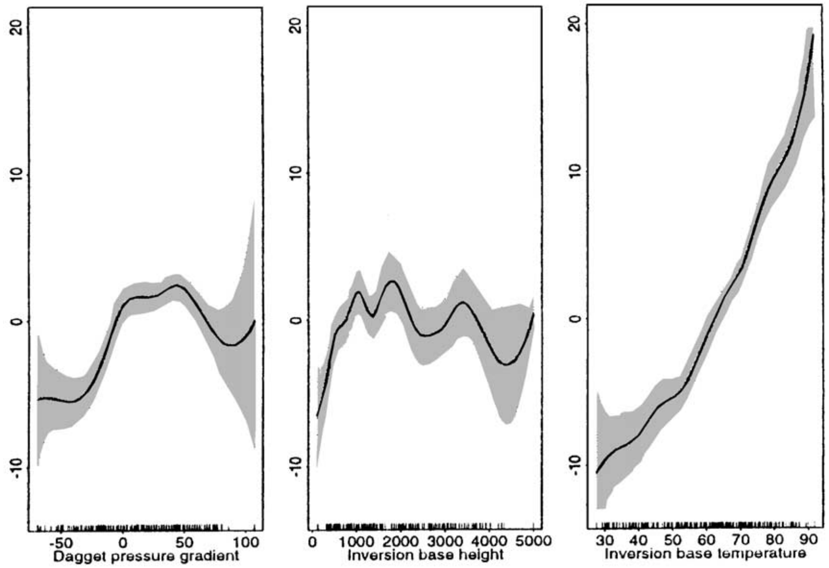

Fig. 1. Penalized spline additive model fit to Californian air pollution data. The amount of smoothing is obtained by application α_{AMASE} with a preliminary estimate of the f_j's obtained via piecewise quadratic fitting and Mallows' C_p. The shaded regions correspond to pointwise 2× standard error bands.

As a check, we fit the data using `S-PLUS`'s `gam()` with default smoothing parameter choice and subtracted off the fitted values for `Dagget pressure gradient` and `Inversion base temperature` from the response. We then applied `smooth.spline()` to resulting the scatterplot, with generalized cross-validation used to select the smoothing parameter. It chose an even larger number of degrees of freedom, so there is some corroboration for the α_{AMASE}-based result obtained here.

While this is just one example, it indicates that α_{AMASE} can be useful for getting an idea of the amount of smoothing required for fitting an additive model.

2.2. Approximation of the degrees of freedom

In an additive model with d components, the *degrees of freedom* for component $\hat{\mathbf{f}}_j$ is defined to be

$$\mathrm{df}_j = \mathrm{tr}(\mathbf{G}_j).$$

While this is straightforward to compute using the full design matrix $\mathbf{X}$, it is often desirable to have an approximation to this quantity that uses only information about component j. In this way, the smoothing parameter corresponding to a particular degrees of freedom value can be specified. A natural candidate for this is

$$\mathrm{tr}(\mathbf{S}_j) = \mathrm{tr}\{(\mathbf{X}_j^{\mathsf{T}}\mathbf{X}_j + \alpha_j\mathbf{D}_j)^{-1}\mathbf{X}_j^{\mathsf{T}}\mathbf{X}_j\},$$

Table 1
Comparison between $\mathrm{tr}(\mathbf{S}_j)$ and $\mathrm{tr}(\mathbf{G}_j)$ for fit to Californian air pollution data

	$j=1$	$j=2$	$j=3$
$\mathrm{tr}(\mathbf{G}_j)$ (exact)	9.036	11.803	9.884
$\mathrm{tr}(\mathbf{S}_j)$ (approx.)	9.136	11.886	9.651
$\mathrm{tr}(\mathbf{G}_j)/\mathrm{tr}(\mathbf{S}_j)$	0.989	0.993	1.024

which has the advantage that it depends only on α_j, and thus allows for easier determination of the smoothing parameter corresponding to the specified degrees of freedom value. This approximation is used, for example, by the function `gam()` in the `S-PLUS` computing package (see e.g. Hastie and Tibshirani, 1990, p. 158).

From (8) we obtain:

Result 4. *For* $\boldsymbol{\alpha}$ *tending to* $\mathbf{0}$ *we have*

$$\frac{\mathrm{tr}(\mathbf{G}_j)}{\mathrm{tr}(\mathbf{S}_j)} - 1 = -\alpha_j \frac{\mathrm{tr}[\mathbf{X}_{[-j]}\{\mathbf{X}_{[-j]}^{\mathsf{T}}(\mathbf{I}-\mathbf{S}_{j0})\mathbf{X}_{[-j]}\}^{-1}\mathbf{X}_{[-j]}^{\mathsf{T}}\mathbf{B}_j]}{p_j + K_j - \alpha_j\,\mathrm{tr}(\mathbf{B}_j)}\{1+\mathrm{o}(1)\}$$

where $\mathbf{B}_j = \mathbf{X}_j(\mathbf{X}_j^{\mathsf{T}}\mathbf{X}_j)^{-1}\mathbf{D}_j(\mathbf{X}_j^{\mathsf{T}}\mathbf{X}_j)^{-1}\mathbf{X}_j^{\mathsf{T}}$.

This result shows that, for $\alpha_j \to 0$, $\mathrm{tr}(\mathbf{S}_j)$ is asymptotic to $\mathrm{tr}(\mathbf{G}_j)$, giving some justification for its use.

The derivation of Result 4 is given in the Appendix.

We tested out the accuracy of this approximation for the fit shown in Fig. 1. The results are given in Table 1. It shows that the accuracy is very reasonable in this case.

3. Extensions

In this section, we consider two important extensions of the classical additive model. The first extension allows one to model some of the covariates in a parametric way, while others are modelled nonparametrically using penalized regression splines. It turns out that the theoretical results for this semiparametric model are very similar to the full nonparametric case. In the second subsection, we will extend the above calculations to the broad class of generalized additive models.

3.1. Semiparametric models

When penalized regression splines are used to model the nonparametric components of a semiparametric model, the same methodology as in fully nonparametric models can be used. If $\mathbf{X}_1,\ldots,\mathbf{X}_q$ $(q<d)$ denote the design matrices of the parametric components, setting $\alpha_1=\cdots=\alpha_q=0$ ensures that these components are not being penalized. With this simple adjustment, we can estimate simultaneously all parametric and nonparametric components.

A partitioning of the design matrix in the following way,

$$\mathbf{X} = [\mathbf{X}_{\text{parm}}\mathbf{X}_{\text{nonp}}], \quad \text{where } \mathbf{X}_{\text{parm}} = [\mathbf{X}_1^* \cdots \mathbf{X}_q^*] \text{ and } \mathbf{X}_{\text{nonp}} = [\mathbf{X}_{q+1}^* \cdots \mathbf{X}_d^*],$$

provides us an easy formula to study separately the parametric and nonparametric parts of the model. Note that in the semiparametric case the parametric design matrices do not contain an intercept term. Using formulas for the inverse of a partitioned matrix, we obtain the following expression for the smoother matrix **G**,

$$\mathbf{G} = \mathbf{X}_{\text{parm}}(\mathbf{X}_{\text{parm}}^{\mathsf{T}}\mathbf{X}_{\text{parm}})^{-1}\mathbf{X}_{\text{parm}}^{\mathsf{T}} + \mathbf{R}(\mathbf{R}^{\mathsf{T}}\mathbf{R} + \mathbf{A}_{\text{nonp}})^{-1}\mathbf{R}^{\mathsf{T}} \tag{11}$$

where

$$\mathbf{R} = (\mathbf{I} - \mathbf{X}_{\text{parm}}(\mathbf{X}_{\text{parm}}^{\mathsf{T}}\mathbf{X}_{\text{parm}})^{-1}\mathbf{X}_{\text{parm}}^{\mathsf{T}})\mathbf{X}_{\text{nonp}} \quad \text{and} \quad \mathbf{A}_{\text{nonp}} = \underset{q+1 \leqslant j \leqslant d}{\text{blockdiag}}(\alpha_j \mathbf{D}_j).$$

This implies that the estimator $\hat{\mathbf{f}}$ can be written as the sum of the regression estimator for the parametric part in $\mathbf{X}_{\text{parm}}$ and the regression spline estimator for the other part, after projection of $\mathbf{X}_{\text{nonp}}$ orthogonal to the $\mathbf{X}_{\text{parm}}$-subspace. Matrix **G** can also be written as the sum of a parametric part

$$\mathbf{G}_{\text{parm}} = \mathbf{X}_{\text{parm}}(\mathbf{X}_{\text{parm}}^{\mathsf{T}}\mathbf{X}_{\text{parm}})^{-1}\mathbf{X}_{\text{parm}}^{\mathsf{T}}\{\mathbf{I} - \mathbf{X}_{\text{nonp}}(\mathbf{R}^{\mathsf{T}}\mathbf{R} + \mathbf{A}_{\text{nonp}})^{-1}\mathbf{R}^{\mathsf{T}}\}$$

and a nonparametric part

$$\mathbf{G}_{\text{nonp}} = \mathbf{X}_{\text{nonp}}(\mathbf{R}^{\mathsf{T}}\mathbf{R} + \mathbf{A}_{\text{nonp}})^{-1}\mathbf{R}^{\mathsf{T}} \tag{12}$$

which can be used to obtain separate estimators $\hat{\mathbf{f}}_{\text{parm}} = \mathbf{G}_{\text{parm}}\mathbf{Y}$ and $\hat{\mathbf{f}}_{\text{nonp}} = \mathbf{G}_{\text{nonp}}\mathbf{Y}$ for, respectively, the parametric and nonparametric components. Note that if there is no parametric part ($\mathbf{X}_{\text{parm}} = \mathbf{0}$), everything reduces to the results of Section 2, whereas for a true semiparametric model, the estimator of the nonparametric part is given by (see Eq. (12))

$$\mathbf{X}_{\text{nonp}}(\mathbf{X}_{\text{nonp}}^{\mathsf{T}}\mathbf{W}\mathbf{X}_{\text{nonp}} + \mathbf{A}_{\text{nonp}})^{-1}\mathbf{X}_{\text{nonp}}^{\mathsf{T}}\mathbf{W}\mathbf{Y},$$

where $\mathbf{W} = \mathbf{I} - \mathbf{X}_{\text{parm}}(\mathbf{X}_{\text{parm}}^{\mathsf{T}}\mathbf{X}_{\text{parm}})^{-1}\mathbf{X}_{\text{parm}}^{\mathsf{T}}$.

Also in a semiparametric model an optimal smoothing parameter can be obtained by minimizing the MASE with respect to $\boldsymbol{\alpha}$. The second term in the right-hand side of Eq. (11) shows that this boils down to replacing **X** by **R** in Section 2.1. Another strategy is to select the smoothing parameters optimally for estimation of the nonparametric part only. In this case we focus on matrix (12), which differs from the second term in (11) by the matrix **W**. Because of this, the expression for the asymptotically optimal smoothing parameter becomes more complicated and resembles the formula which will be given in the next section.

3.2. Generalized additive models

Model (2) is mainly used for normally distributed errors. This Gaussian regression model is a member of the class of generalized linear models (GLM), see, e.g., McCullagh and Nelder (1989). In all these GLM, the likelihood of the response belongs to an exponential family and can be written as follows,

$$\exp\left\{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right\},$$

where θ is the so-called natural parameter, which is related to the mean response in the following way, $\mathrm{d}b(\theta)/\mathrm{d}\theta=E(Y)=\mu$, and ϕ is a dispersion parameter. A GLM is further specified by a known "link" function $g(\cdot)$, and $\eta=g(\mu)$ is called the systematic component.

Instead of modelling η as a linear function of the covariates, which would result in the classical generalized linear model, we can use a nonparametric estimator in an additive models framework. More specifically, as in Hastie and Tibshirani (1990), we assume $\eta(x_1,\dots,x_d)=\eta_1(x_1)+\cdots+\eta_d(x_d)$. We will estimate each of these functions η_j using penalized regression splines of degree p_j. This means that each additive component $\boldsymbol{\eta}_j=[\eta_j(x_{j1}),\dots,\eta_j(x_{jn})]^{\mathsf{T}}$ is modelled as $\mathbf{X}_j\boldsymbol{\beta}_j$ with $\mathbf{X}_j$ as in Section 2. The parameter vector $\boldsymbol{\beta}=[\boldsymbol{\beta}_1^{\mathsf{T}},\dots,\boldsymbol{\beta}_d^{\mathsf{T}}]^{\mathsf{T}}$ can now be estimated by maximizing the following penalized log likelihood function:

$$\sum_{i=1}^{n}[Y_i\theta(\mathbf{x}_1,\dots,\mathbf{x}_n;\boldsymbol{\beta})-b\{\theta(\mathbf{x}_1,\dots,\mathbf{x}_n;\boldsymbol{\beta})\}]/a(\phi)-\frac{1}{2}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{A}\boldsymbol{\beta}.$$

For $\mathbf{U}_\beta$ the vector of first partial derivatives and $\mathbf{J}_\beta$ the matrix of minus second partial derivatives of the log likelihood with respect to $\boldsymbol{\beta}$, we immediately obtain that the $(k+1)$st update in a Newton–Raphson iterative procedure is given by

$$\boldsymbol{\beta}^{(k+1)}=(\mathbf{J}_{\beta^{(k)}}+\mathbf{A})^{-1}(\mathbf{J}_{\beta^{(k)}}\boldsymbol{\beta}^{(k)}+\mathbf{U}_{\beta^{(k)}})$$

or, by the chain rule,

$$\boldsymbol{\beta}^{(k+1)}=(\mathbf{X}^{\mathsf{T}}\mathbf{J}_{\eta^{(k)}}\mathbf{X}+\mathbf{A})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{J}_{\eta^{(k)}}(\boldsymbol{\eta}^{(k)}+\mathbf{J}_{\eta^{(k)}}^{-1}\mathbf{U}_{\eta^{(k)}}), \tag{13}$$

with $\mathbf{A}$ as in Section 2 and with $L(\cdot)$ denoting the likelihood of the data,

$$\mathbf{J}_\eta=-\operatorname{diag}_{1\leqslant i\leqslant n}\left\{\frac{\partial^2\log L(Y_i;\mathbf{x}_i,\boldsymbol{\beta})}{\partial\eta^2}\right\}$$

and

$$\mathbf{U}_\eta=\left\{\frac{\partial\log L(Y_1;\mathbf{x}_1,\boldsymbol{\beta})}{\partial\eta},\dots,\frac{\partial\log L(Y_n;\mathbf{x}_n,\boldsymbol{\beta})}{\partial\eta}\right\}^{\mathsf{T}}.$$

Note that this estimate is a special case of the class of estimators presented in Marx et al. (1992).

If the observed Fisher information matrix is replaced by its expectation $\mathbf{I}_\beta=E(\mathbf{J}_\beta)$, we obtain the iterative solutions of a Fisher scoring procedure. Note that these two algorithms coincide if the canonical link function is used, that is, if $\eta=\theta$.

From Eq. (13) it is immediately clear that an equivalent way of obtaining the estimators $\hat{\boldsymbol{\beta}}$ is via iteratively reweighted ridge regression, where the adjusted dependent variable is defined as $\mathbf{Z}_{\eta^{(k)}} = \boldsymbol{\eta}^{(k)} + \mathbf{J}_{\eta^{(k)}}^{-1}\mathbf{U}_{\eta^{(k)}}$. A first order approximation of the estimator of the coefficient $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{J}_{\eta}\mathbf{X} + \mathbf{A})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{J}_{\eta}\mathbf{Z}_{\eta},$$

which allows us to extend the asymptotic results of Section 2 to generalized additive models. For Gaussian responses, all results shown below simplify to those of Section 2.

3.2.1. Approximation of the risk

The smoothing parameter $\boldsymbol{\alpha}$ will be selected by extending the definition of MASE to the context of generalized additive models. The overall risk is now measured by

$$\text{MASE}(\hat{\boldsymbol{\eta}}) = \frac{1}{n}||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}||^2.$$

This can be rewritten as,

$$\text{MASE}(\hat{\boldsymbol{\eta}}) = \frac{1}{n}\,\text{tr}\{\mathbf{G}\,\text{Var}(\mathbf{U}_{\eta})\mathbf{G}\} + \frac{1}{n}||\mathbf{G}\mathbf{I}_{\eta}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\eta}||^2,$$

where $\mathbf{G} = \mathbf{X}(\mathbf{I}_{\beta} + \mathbf{A})^{-1}\mathbf{X}^{\mathsf{T}}$. The asymptotic approximation to MASE is given in Result 5. Its derivation is similar to that of Result 3 and so is not presented.

Result 5. *For* $\boldsymbol{\alpha}$ *tending to* $\mathbf{0}$,

$$\text{AMASE}(\hat{\boldsymbol{\eta}}) = \frac{1}{n}[\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{Q}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{q} + \text{tr}\{\mathbf{G_0}\,\text{Var}(\mathbf{U}_{\eta})\mathbf{G_0}\}]$$

where $\mathbf{q}$ $(d \times 1)$ *and* $\mathbf{Q}$ $(d \times d)$ *have entries* $\mathbf{q}_j = \text{tr}\{\tilde{\mathbf{B}}_j\,\text{Var}(\mathbf{U}_{\eta})\mathbf{G_0}\}$,

$$\mathbf{Q}_{jj'} = (\tilde{\mathbf{B}}_j\mathbf{I}_{\eta}\mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\tilde{\mathbf{B}}_{j'}\mathbf{I}_{\eta}\mathbf{X}\boldsymbol{\beta}) + \text{tr}\{(\tilde{\mathbf{B}}_j)^{\mathsf{T}}\,\text{Var}(\mathbf{U}_{\eta})\tilde{\mathbf{B}}_{j'}\},$$

$$\tilde{\mathbf{B}}_j = \tilde{\mathbf{X}}_j(\mathbf{X}_j^{\mathsf{T}}\mathbf{I}_{\eta}\tilde{\mathbf{X}}_j)^{-1}\mathbf{D}_j(\mathbf{X}_j^{\mathsf{T}}\mathbf{I}_{\eta}\tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^{\mathsf{T}} \quad \textit{and} \quad \tilde{\mathbf{X}}_j = (\mathbf{I} - \mathbf{S}_{\mathbf{0},[-j]}\mathbf{I}_{\eta})\mathbf{X}_j.$$

The AMASE-optimal smoothing parameters are

$$\boldsymbol{\alpha}_{\text{AMASE}} = \mathbf{Q}^{-1}\mathbf{q}.$$

3.2.2. Approximation of the degrees of freedom

We follow Hastie and Tibshirani (1990) in defining the degrees of freedom for the jth component in a generalized additive model as $\text{d}f_j = \text{tr}(\mathbf{G}_j\mathbf{F})$ where now

$$\mathbf{G}_j = [\mathbf{0}\cdots\mathbf{0}\;\mathbf{X}_j\;\mathbf{0}\cdots\mathbf{0}](\mathbf{X}^{\mathsf{T}}\mathbf{F}\mathbf{X} + \mathbf{A})^{-1}\mathbf{X}^{\mathsf{T}}$$

with $\mathbf{F}$ either the observed or the expected Fisher information matrix with respect to $\boldsymbol{\eta}$.

As demonstrated in the Appendix, since we can show that

$$\frac{\text{tr}(\mathbf{G}_j\mathbf{F})}{\text{tr}(\mathbf{S}_j\mathbf{F})} - 1 = -\,\alpha_j\frac{\text{tr}[\mathbf{X}_{[-j]}\{\mathbf{X}_{[-j]}^{\mathsf{T}}\mathbf{F}(\mathbf{I} - \mathbf{S}_{j0}\mathbf{F})\mathbf{X}_{[-j]}\}^{-1}\mathbf{X}_{[-j]}^{\mathsf{T}}\mathbf{F}\mathbf{B}_j\mathbf{F}]}{p_j + K_j - \alpha_j\,\text{tr}\{(\mathbf{X}_j^{\mathsf{T}}\mathbf{F}\mathbf{X}_j)^{-1}\mathbf{D}_j\}}\{1 + \text{o}(1)\}, \tag{14}$$

where

$$\mathbf{B}_j = \mathbf{X}_j(\mathbf{X}_j^{\mathsf{T}}\mathbf{F}\mathbf{X}_j)^{-1}\mathbf{D}_j(\mathbf{X}_j^{\mathsf{T}}\mathbf{F}\mathbf{X}_j)^{-1}\mathbf{X}_j^{\mathsf{T}} \quad \text{and} \quad \mathbf{S}_j = \mathbf{X}_j(\mathbf{X}_j^{\mathsf{T}}\mathbf{F}\mathbf{X}_j + \mathbf{A}_j)^{-1}\mathbf{X}_j^{\mathsf{T}},$$

the degree of freedom value $\mathrm{d}f_j$ might be approximated by $\mathrm{tr}(\mathbf{S}_j\mathbf{F})$. This has the computational advantage that only that part of the design matrix related to the jth covariate needs to be used.

3.2.3. Semiparametric models

If some of the covariates of a generalized additive model are modelled parametrically and others nonparametrically, the resulting semiparametric model can be handled similarly, as before. Assume that the first q covariates are the components of the parametric part, then we take $\alpha_1 = \cdots = \alpha_q = 0$. All results of Section 3.2 remain valid, and a similar decomposition as in Section 3.1 holds. Using the same partitioning of the design matrix as in Section 3.1, the smoother matrix $\mathbf{G}$ is now obtained as

$$\mathbf{G} = \mathbf{X}_{\mathrm{parm}}(\mathbf{X}_{\mathrm{parm}}^{\mathsf{T}}\mathbf{F}\mathbf{X}_{\mathrm{parm}})^{-1}\mathbf{X}_{\mathrm{parm}}^{\mathsf{T}} + \mathbf{R}(\mathbf{X}_{\mathrm{nonp}}^{\mathsf{T}}\mathbf{F}\mathbf{R} + \mathbf{A}_{\mathrm{nonp}})^{-1}\mathbf{R}^{\mathsf{T}}$$

where $\mathbf{R} = \{\mathbf{I} - \mathbf{X}_{\mathrm{parm}}(\mathbf{X}_{\mathrm{parm}}^{\mathsf{T}}\mathbf{F}\mathbf{X}_{\mathrm{parm}})^{-1}\mathbf{X}_{\mathrm{parm}}^{\mathsf{T}}\mathbf{F}\}\mathbf{X}_{\mathrm{nonp}}$. Separate estimators for the parametric and the nonparametric components can be obtained using the following smoother matrices:

$$\mathbf{G}_{\mathrm{parm}} = \mathbf{X}_{\mathrm{parm}}(\mathbf{X}_{\mathrm{parm}}^{\mathsf{T}}\mathbf{F}\mathbf{X}_{\mathrm{parm}})^{-1}\mathbf{X}_{\mathrm{parm}}^{\mathsf{T}}\{\mathbf{I} - \mathbf{X}_{\mathrm{nonp}}(\mathbf{R}^{\mathsf{T}}\mathbf{F}\mathbf{R} + \mathbf{A}_{\mathrm{nonp}})^{-1}\mathbf{R}^{\mathsf{T}}\}$$

and

$$\mathbf{G}_{\mathrm{nonp}} = \mathbf{X}_{\mathrm{nonp}}(\mathbf{R}^{\mathsf{T}}\mathbf{F}\mathbf{R} + \mathbf{A}_{\mathrm{nonp}})^{-1}\mathbf{R}^{\mathsf{T}}.$$

If we focus on the nonparametric part only, we can write the estimator of the nonparametric part of $\boldsymbol{\eta}$ as follows,

$$\mathbf{X}_{\mathrm{nonp}}(\mathbf{X}_{\mathrm{nonp}}^{\mathsf{T}}\mathbf{W}\mathbf{X}_{\mathrm{nonp}} + \mathbf{A}_{\mathrm{nonp}})^{-1}\mathbf{X}_{\mathrm{nonp}}^{\mathsf{T}}\mathbf{W}\mathbf{Z},$$

where now

$$\mathbf{W} = \mathbf{F}\{\mathbf{I} - \mathbf{X}_{\mathrm{parm}}(\mathbf{X}_{\mathrm{parm}}^{\mathsf{T}}\mathbf{F}\mathbf{X}_{\mathrm{parm}})^{-1}\mathbf{X}_{\mathrm{parm}}^{\mathsf{T}}\mathbf{F}\}. \tag{15}$$

3.3. Multiparameter models and generalized estimating equations

Wild and Yee (1996) and Yee and Wild (1996) introduced the use of vector smoothing splines in the multivariate regression and generalized estimating equations (GEE) context, where the parameter vector of interest is modelled in an additive way. The selection of the smoothing parameters and the approximation of the degrees of freedom can be obtained similarly as in the previous section, after introducing the following notation. In most cases, an estimator for the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ is obtained by solving a set of estimating equations:

$$\sum_{i=1}^{n} \psi_k(\mathbf{Y}_i; \mathbf{x}_i, \boldsymbol{\theta}) = 0, \quad k = 1, \ldots, m. \tag{16}$$

If the likelihood of the data is known, ψ_k might be the partial derivative of the log likelihood with respect to θ_k. For example, for $m = 2$, θ_1 can be the mean and θ_2 the log variance of normally distributed data. For robust estimators, Eq. (16) might lead to M-estimators of $\boldsymbol{\theta}$ (Huber, 1981), or (16) can represent some set of generalized estimating equations (Liang and Zeger, 1986). In the latter case, usually the response vector is multidimensional.

In a regression model, the parameters are modelled as function of the covariates $\mathbf{x}_i = [\mathbf{x}_{1i}^{\mathsf{T}}, \ldots, \mathbf{x}_{mi}^{\mathsf{T}}]^{\mathsf{T}}$. Let

$$\begin{bmatrix} \theta_1(\mathbf{x}_{1i}) \\ \vdots \\ \theta_m(\mathbf{x}_{mi}) \end{bmatrix} = \begin{bmatrix} \theta_1(x_{1i1}, \ldots, x_{1id_1}) \\ \vdots \\ \theta_m(x_{mi1}, \ldots, x_{mid_m}) \end{bmatrix}$$

be the parameter vector of interest. The covariate vectors $\mathbf{x}_{ji}$ $(j = 1, \ldots, m)$ can be the same, different for each parameter θ_j or partly overlapping with another vector $\mathbf{x}_{ki}$ $(k \neq j)$. For example, in toxicity studies, typically resulting in clustered binary data, a given dose might have its influence on both proportion of success (e.g., inverse logit of θ_1) and association between outcomes (represented by θ_2), but, say, individual weight of the subjects might be included only in the success probability parameter, and not in the association part.

In additive models, each of these parameter functions is, for some unknown functions f_{kj} $(k = 1, \ldots, m; j = 1, \ldots, d_k)$

$$\theta_k(\mathbf{x}_{ki}) = \beta_{k0} + \sum_{j=1}^{d_k} f_{kj}(x_{kij}).$$

For identifiability purposes, we subtract the mean of the function values from each of the f_{kj}. By introducing a regression spline design matrix $\mathbf{X}_{kj}$ (defined similarly as the matrix $\mathbf{X}_j$ in Section 2), we define the design matrix for θ_k as $\mathbf{X}_k = [\mathbf{X}_{k1} \cdots \mathbf{X}_{kd_k}]$, such that $\theta_k(\mathbf{x}_{ki}) = \mathbf{X}_k^{\mathsf{T}} \boldsymbol{\beta}_k$. The design matrix $\mathbf{X}$ is now defined as $\mathbf{X} = \mathrm{blockdiag}_{1 \leqslant k \leqslant m}(\mathbf{X}_k)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathsf{T}}, \ldots, \boldsymbol{\beta}_m^{\mathsf{T}})^{\mathsf{T}}$. The (observed) "Fisher Information matrix" is $\mathbf{J}_\beta = \mathbf{X}^{\mathsf{T}} \mathbf{J}_\theta \mathbf{X}$ where $\mathbf{J}_\theta$ is a partitioned matrix with (j,k)th block

$$\mathbf{J}_{\theta,jk} = - \operatorname*{diag}_{1 \leqslant i \leqslant n} \frac{\partial \psi_j}{\partial \theta_k}(\mathbf{Y}_i; \mathbf{X}_i, \boldsymbol{\theta}).$$

From this, the expected matrices $\mathbf{I}_\beta$ and $\mathbf{I}_\theta$ are easily obtained. The "score" vector $\mathbf{U}_\theta$ is the column vector

$$[\psi_1(\mathbf{Y}_1; \mathbf{x}_1, \boldsymbol{\theta}), \ldots, \psi_1(\mathbf{Y}_n; \mathbf{x}_n, \boldsymbol{\theta}), \ldots, \psi_m(\mathbf{Y}_n; \mathbf{X}_n, \boldsymbol{\theta})]^{\mathsf{T}}$$

and $\mathbf{U}_\beta = \mathbf{X}^{\mathsf{T}} \mathbf{U}_\theta$.

With these ingredients we can obtain an estimator for $\boldsymbol{\beta}$ via the iteratively reweighted ridge regression scheme, as presented in Section 3.2. To prove the results of Section 2 for general estimating equations, take $\mathbf{W} = \mathbf{J}_\beta$ in the Appendix. The semiparametric case can be handled by similar adjustments as explained in Sections 3.1 and 3.2.3.

4. Discussion

The explicitness of the penalized spline approach to additive modelling has several advantages. In this paper, we have shown that it allows for theoretical analyses via relatively simple mathematical methods. Closed form expressions for the asymptotically optimal smoothing parameters and the error in common degrees of freedom approximations are useful outcomes of this analysis. It is our hope that this paper lays the foundation for revealing analyses of more complex penalized spline models.

Acknowledgements

The first author is supported by a travel grant of the Fund for Scientific Research Flanders–Belgium. Research of the second author is supported by grant "Research Assistant of the Fund for Scientific Research Flanders–Belgium (FWO)", while at Center for Statistics, Limburgs Universitair Centrum, Diepenbeek. The third author's research was supported by EPA Grant R 824757.

Appendix A.

All calculations will be presented generally, using a symmetric weight matrix $\mathbf{W}$. To obtain the results of Section 2 take $\mathbf{W}=\mathbf{I}$. For the nonparametric part of a semiparametric model, $\mathbf{W}$ is defined in Section 3.1, $\mathbf{S}_{[-j]}$ and $\mathbf{G}_j$ are defined similarly as in (3) and (6), but with the matrices $\mathbf{X}$ and $\mathbf{A}$ replaced by $\mathbf{X}_{\text{nonp}}$ and $\mathbf{A}_{\text{nonp}}$ respectively. For the generalized additive model, $\mathbf{W}$ is the Fisher information matrix. And for the semiparametric generalized additive model, $\mathbf{W}$ is defined in (15). If $\mathbf{W}$ is positive definite or a projection matrix, the following results hold.

A.1. Derivation of result 1

For $j=1,\ldots,d$, we rewrite $\mathbf{G}$ in the following way,

$$\begin{aligned}\mathbf{G}&=\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X}+\mathbf{A})^{-1}\mathbf{X}^{\mathsf{T}}=\mathbf{G}_{[-j]}+\mathbf{G}_j\\&=(\mathbf{X}_{[-j]},\mathbf{X}_j)\begin{pmatrix}(\mathbf{X}_{[-j]})^{\mathsf{T}}\mathbf{W}\mathbf{X}_{[-j]}+\mathbf{A}_{[-j]} & (\mathbf{X}_{[-j]})^{\mathsf{T}}\mathbf{W}\mathbf{X}_j\\ \mathbf{X}_j^{\mathsf{T}}\mathbf{W}\mathbf{X}_{[-j]} & \mathbf{X}_j^{\mathsf{T}}\mathbf{W}\mathbf{X}_j+\mathbf{A}_j\end{pmatrix}^{-1}\mathbf{X}^{\mathsf{T}}.\end{aligned}$$

Using formulae for the inverse of a partitioned matrix, (see, e.g., Searle, 1982, p. 260),

$$\mathbf{G}_{[-j]}=\mathbf{S}_{[-j]}[\mathbf{I}-\mathbf{W}\mathbf{X}_j\{\mathbf{X}_j^{\mathsf{T}}\mathbf{W}(\mathbf{I}-\mathbf{S}_{[-j]}\mathbf{W})\mathbf{X}_j+\mathbf{A}_j\}^{-1}\mathbf{X}_j^{\mathsf{T}}(\mathbf{I}-\mathbf{W}\mathbf{S}_{[-j]})] \tag{17}$$

$$\mathbf{G}_j=\mathbf{X}_j\{\mathbf{X}_j^{\mathsf{T}}\mathbf{W}(\mathbf{I}-\mathbf{S}_{[-j]}\mathbf{W})\mathbf{X}_j+\mathbf{A}_j\}^{-1}\mathbf{X}_j^{\mathsf{T}}(\mathbf{I}-\mathbf{W}\mathbf{S}_{[-j]}). \tag{18}$$

Note that for $\mathbf{W}=\mathbf{I}$, the expression for $\mathbf{G}_j$ reduces to (8). From (17) and (18), the recursive formula

$$\mathbf{G}=\mathbf{S}_{[-j]}+(\mathbf{I}-\mathbf{S}_{[-j]}\mathbf{W})\mathbf{G}_j \tag{19}$$

is easily obtained.

A.2. Derivation of result 2

If $d=1$, the result is shown by extending the result of Wand (1999b) to allow for a general weight matrix $\mathbf{W}$. Suppose that (9) is true for $j=d-1$, that is,

$$\mathbf{S}_{[-d]}=\mathbf{S}_{\mathbf{0},[-d]}-\sum_{j=1}^{d-1}\alpha_j\tilde{\mathbf{B}}_j^{[-d]}+o\left(\sum_{j=1}^{d-1}\alpha_j\tilde{\mathbf{B}}_j^{[-d]}\right), \tag{20}$$

where $\tilde{B}_j^{[-d]}$ is defined similar to $\tilde{B}_j$, but now in the model with all covariates except x_d.

Starting from (18) and using (20), we can approximate $\mathbf{G}_d$ by

$$\begin{aligned}\mathbf{G}_d &\approx \mathbf{X}_d\left\{\mathbf{X}_d^{\mathsf{T}}\mathbf{W}\mathbf{X}_d-\mathbf{X}_d^{\mathsf{T}}\mathbf{W}\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W}+\sum_{j=1}^{d-1}\alpha_j\mathbf{X}_d^{\mathsf{T}}\mathbf{W}\tilde{\mathbf{B}}_j^{[-d]}\mathbf{W}\mathbf{X}_d+\alpha_d\mathbf{D}_d\right\}^{-1}\\ &\quad\times\mathbf{X}_d^{\mathsf{T}}(\mathbf{I}-\mathbf{W}\mathbf{S}_{[-d]})\\ &=\mathbf{X}_d\{\mathbf{X}_d^{\mathsf{T}}\mathbf{W}(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d+\mathbf{L}_{\boldsymbol{\alpha}}\}^{-1}\mathbf{X}_d^{\mathsf{T}}(\mathbf{I}-\mathbf{W}\mathbf{S}_{[-d]})\\ &=\mathbf{X}_d\{\mathbf{I}+(\mathbf{X}_d^{\mathsf{T}}\mathbf{W}(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\}^{-1}\mathbf{L}_{\boldsymbol{\alpha}}\}^{-1}(\mathbf{X}^{\mathsf{T}}\mathbf{W}\{\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\}^{-1}\\ &\quad\times\mathbf{X}_d^{\mathsf{T}}(\mathbf{I}-\mathbf{W}\mathbf{S}_{[-d]}),\end{aligned}$$

where

$$\mathbf{L}_{\boldsymbol{\alpha}}=\alpha_d\mathbf{D}_d+\sum_{k=1}^{d-1}\alpha_k\mathbf{X}_d^{\mathsf{T}}\mathbf{W}\tilde{\mathbf{B}}_k^{[-d]}\mathbf{W}\mathbf{X}_d.$$

Then, (19) leads to the following approximation,

$$\begin{aligned}\mathbf{G}\approx\mathbf{G}_\mathbf{0}-\sum_{j=1}^{d-1}\alpha_j\{\tilde{\mathbf{B}}_j^{[-d]}&+[(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\{\mathbf{X}_d^{\mathsf{T}}\mathbf{W}(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\}^{-1}\\ &\times\mathbf{X}_d^{\mathsf{T}}\mathbf{W}-\mathbf{I}]\tilde{\mathbf{B}}_j^{[-d]}\mathbf{W}\mathbf{X}_d\{\mathbf{X}_d^{\mathsf{T}}\mathbf{W}(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\}^{-1}\mathbf{X}_d^{\mathsf{T}}(\mathbf{I}-\mathbf{W}\mathbf{S}_{\mathbf{0},[-d]})\\ &-(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\{\mathbf{X}_d^{\mathsf{T}}\mathbf{W}(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\}^{-1}\mathbf{X}_d^{\mathsf{T}}\mathbf{W}\tilde{\mathbf{B}}_j^{[-d]}\}-\alpha_d\tilde{\mathbf{B}}_d,\end{aligned}$$

where, from (18) and (19) with $\boldsymbol{\alpha}=\mathbf{0}$,

$$\begin{aligned}\mathbf{G}_\mathbf{0}&=\mathbf{S}_{\mathbf{0},[-d]}+(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\{\mathbf{X}_d^{\mathsf{T}}\mathbf{W}(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\}^{-1}\\ &\quad\times\mathbf{X}_d^{\mathsf{T}}(\mathbf{I}-\mathbf{W}\mathbf{S}_{\mathbf{0},[-d]}).\end{aligned}$$

Since

$$\begin{aligned}\tilde{\mathbf{B}}_j&=(\mathbf{I}-(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\{\mathbf{X}_d^{\mathsf{T}}\mathbf{W}(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\}^{-1}\mathbf{X}_d^{\mathsf{T}}\mathbf{W})\tilde{\mathbf{B}}_j^{[-d]}\\ &\quad\times(\mathbf{I}-(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\{\mathbf{X}_d^{\mathsf{T}}\mathbf{W}(\mathbf{I}-\mathbf{S}_{\mathbf{0},[-d]}\mathbf{W})\mathbf{X}_d\}^{-1}\mathbf{X}_d^{\mathsf{T}}\mathbf{W})^{\mathsf{T}},\end{aligned}$$

the result is proven.

A.3. Derivation of result 4

Using the equivalent definition of the inverse of $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{A}$, we obtain, similarly as (17) and (18), that

$$\mathbf{G}_j = \mathbf{S}_j\{\mathbf{I}-\mathbf{W}\mathbf{X}_{[-j]}\{\mathbf{X}_{[-j]}^\mathsf{T}\mathbf{W}(\mathbf{I}-\mathbf{S}_j\mathbf{W})\mathbf{X}_{[-j]}+\mathbf{A}_{[-j]}\}^{-1}\mathbf{X}_{[-j]}^\mathsf{T}(\mathbf{I}-\mathbf{W}\mathbf{S}_j)\}.$$

Then,

$$\begin{aligned}\operatorname{tr}(\mathbf{S}_j\mathbf{W}-\mathbf{G}_j\mathbf{W}) = {} & \operatorname{tr}(\mathbf{X}_{[-j]}\{(\mathbf{X}_{[-j]})^\mathsf{T}\mathbf{W}(\mathbf{I}-\mathbf{S}_j\mathbf{W})\mathbf{X}_{[-j]}+\mathbf{A}_{[-j]}\}^{-1} \\ & \times(\mathbf{X}_{[-j]})^\mathsf{T}(\mathbf{I}-\mathbf{W}\mathbf{S}_j)\mathbf{W}\mathbf{S}_j\mathbf{W}).\end{aligned}$$

For $\boldsymbol{\alpha}\to\mathbf{0}$, by (9), we have the approximation

$$(\mathbf{I}-\mathbf{W}\mathbf{S}_j)\mathbf{W}\mathbf{S}_j\mathbf{W}\approx\alpha_j\mathbf{B}_j\mathbf{W},$$

from which, by (9), the numerator of (14) is easily obtained.

References

Beck, N., Jackman, S., 1998. Beyond linearity by default: generalized additive models. Amer. J. Polit. Sci. 42, 596–627.

Breiman, L., Friedman, J., 1985. Estimating optimal transformation for multiple regression and correlation (with discussion). J. Amer. Statist. Assoc. 80, 580–619.

Chambers, J.M., Hastie, T.J., 1991. Statistical Models in S. Wadsworth/Brooks Cole, Pacific Grove, CA.

Claeskens, G., Aerts, M., 2000. On local estimating equations in additive multiparameter models, Statist. Prob. Letters 49, 139–148.

Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties (with discussion). Statist. Sci. 89, 89–121.

Fan, J., Hardle, W., Mammen, E., 1998. Direct estimation of low dimensional components in additive models. Ann. Statist. 26, 943–971.

Härdle, W., Marron, J.S., 1995. Fast and simple scatterplot smoothing. Comput. Statist. Data Anal. 20, 1–17.

Hastie, T.J., 1996. Pseudosplines. J. Roy Statist. Soc. Ser. B 58, 379–396.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized additive models. Chapman & Hall, London.

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

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

Linton, O.B., Härdle, W., 1996. Estimation of additive regression models with known links. Biometrika 83, 529–540.

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

Marx, B.D., Eilers, P.H.C., 1998. Direct generalized additive modeling with penalized likelihood. Comput. Statist. Data Anal. 28, 193–209.

Marx, B.D., Eilers, P.H.C, Smith, E.P., 1992. Ridge likelihood estimation for generalized linear regression. In: van der Heijden, Jansen, Francis, Seeber (Eds.), Statistical Modelling. North-Holland, Amsterdam, pp. 227–237.

McCullagh, P., Nelder, J.A., 1989. Generalized linear models, 2nd Edition. Chapman & Hall, London.

Nychka, D., Cummins, D., 1996. Comment on Flexible smoothing with B-splines and penalties by P.H.C. Eilers and B.D. Marx. Statist. Sci. 89, 104–105.

Opsomer, J.D., 2000. Asymptotic properties of backfitting estimators. J. Mult. Anal. 73, 166–179.

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

Ruppert, D., Carroll, R.J., 2000. Spatially-adaptive penalties for spline fitting. Austral. NZ J. Statist. 42, 205–224.

Ruppert, D., Sheather, S.J., Wand, M.P., 1995. An effective bandwidth selector for local least squares regression. J. Amer. Statist. Assoc. 90, 1257–1270.

Schwartz, J., 1994. Nonparametric smoothing in the analysis of air pollution and respiratory illness. Canad. J. Statist. 22, 471–487.

Searle, S.R., 1982. Matrix Algebra useful for Statistics. Wiley, New York.

Wand, M.P., 1999a. Central Limit Theorems for Local Polynomial Backfitting Estimators. J. Mult. Anal. 70, 57–65.

Wand, M.P., 1999b. On the optimal amount of smoothing in penalized spline regression. Biometrika 86, 936–940.

Wild, C.J., Yee, T.W., 1996. Additive extensions to generalized estimating equation models. J. Roy. Statist. Soc. B 58, 711–725.

Yee, T.W., Wild, C.J., 1996. Vector generalized additive models. J. Roy. Statist. Soc. B 58, 481–493.