Pya, N., & Wood, S. N. (2015). Shape constrained additive models. Statistics and Computing, 25, 543-559. DOI: 10.1007/s11222-013-9448-7

Stat Comput (2015) 25:543–559
DOI 10.1007/s11222-013-9448-7

Shape constrained additive models

Natalya Pya · Simon N. Wood

Received: 6 March 2013 / Accepted: 27 December 2013 / Published online: 25 February 2014

Abstract A framework is presented for generalized additive modelling under shape constraints on the component functions of the linear predictor of the GAM. We represent shape constrained model components by mildly non-linear extensions of P-splines. Models can contain multiple shape constrained and unconstrained terms as well as shape constrained multi-dimensional smooths. The constraints considered are on the sign of the first or/and the second derivatives of the smooth terms. A key advantage of the approach is that it facilitates efficient estimation of smoothing parameters as an integral part of model estimation, via GCV or AIC, and numerically robust algorithms for this are presented. We also derive simulation free approximate Bayesian confidence intervals for the smooth components, which are shown to achieve close to nominal coverage probabilities. Applications are presented using real data examples including the risk of disease in relation to proximity to municipal incinerators and the association between air pollution and health.

Keywords Monotonic smoothing · Convex smoothing · Generalized additive model · P-splines

Electronic supplementary material The online version of this article (doi:10.1007/s11222-013-9448-7) contains supplementary material, which is available to authorized users.

N. Pya (✉) · S. N. Wood
Mathematical Sciences, University of Bath,
Bath BA2 7AY, UK
e-mail: n.y.pya@bath.ac.uk

S. N. Wood
e-mail: s.wood@bath.ac.uk

1 Introduction

This paper is about estimation and inference with the model

$$g(\mu_i) = \mathbf{A}\boldsymbol{\theta} + \sum_j f_j(z_{ji}) + \sum_k m_k(x_{ki}), \quad Y_i \sim \mathrm{EF}(\mu_i, \phi), \tag{1}$$

where Y_i is a univariate response variable with mean μ_i arising from an exponential family distribution with scale parameter ϕ (or at least with mean variance relationship known to within a scale parameter), g is a known smooth monotonic link function, $\mathbf{A}$ is a model matrix, $\boldsymbol{\theta}$ is a vector of unknown parameters, f_j is an unknown smooth function of predictor variable z_j and m_k is an unknown *shape constrained* smooth function of predictor variable x_k. The predictors x_j and z_k may be vector valued.

It is the shape constraints on the m_k that differentiate this model from a standard generalized additive model (GAM). In many studies it is natural to assume that the relationship between a response variable and one or more predictors obeys certain shape restrictions. For example, the growth of children over time and dose-response curves in medicine are known to be monotonic. The relationships between daily mortality and air pollution concentration, between body mass index and incidence of heart disease are other examples requiring shape restrictions. Unconstrained models might be too flexible and give implausible or un-interpretable results.

Here we develop a general framework for shape constrained generalized additive models (SCAM), covering estimation, smoothness selection, interval estimation and also allowing for model comparison. The aim is to make SCAMs as routine to use as conventional unconstrained GAMs. To do this we build on the established framework for generalized

additive modelling covered, for example, in Wood (2006a). Model smooth terms are represented using spline type penalized basis function expansions; given smoothing parameter values, model coefficients are estimated by maximum penalized likelihood, achieved by an inner iteratively reweighted least squares type algorithm; smoothing parameters are estimated by the outer optimization of a GCV or AIC criterion. Interval estimation is achieved by taking a Bayesian view of the smoothing process, and model comparison can be achieved using AIC, for example.

This paper supplies the novel components required to make this basic strategy work, namely

1. We propose shape constrained P-splines (SCOP-splines), based on a novel mildly non linear extension of the P-splines of Eilers and Marx (1996), with novel discrete penalties. These allow a variety of shape constraints for one and multidimensional smooths. From a computational viewpoint, they ensure that the penalized likelihood and the GCV/AIC scores are smooth with respect to the model coefficients and smoothing parameters, allowing the development of efficient and stable model estimation methods.
2. We develop stable computational schemes for estimating the model coefficients and smoothing parameters, able to deal with the ill-conditioning that can affect even unconstrained GAM fits (Wood 2004, 2008), while retaining computational efficiency. The extra non-linearity induced by the use of SCOP-splines does not allow the unconstrained GAM methods to be re-used or simply modified. Substantially new algorithms are required instead.
3. We provide simulation free approximate Bayesian confidence intervals for the SCOP-spline model components in this setting.

The bulk of this paper concentrates on these new developments, covering standard results on unconstrained GAMs only tersely. We refer the reader to Wood (2006a) for a more complete coverage of this background. Technical details and extensive comparative testing are provided in online supplementary material.

To understand the motivation for our approach, note that it is not difficult to construct shape constrained spline like smoothers, by subjecting the spline coefficients to linear inequality constraints (Ramsay 1988; Wood 1994; Zhang 2004; Kelly and Rice 1990; Meyer 2012). However, this approach leads to methodological problems in estimating the smoothing parameters of the spline. The use of linear inequality constraints makes it difficult to optimize standard smoothness selection criteria, such as AIC and GCV with respect to multiple smoothing parameters. The difficulty arises because the derivatives of these criteria change discontinuously as constraints enter or leave the set of active constraints. This leads to failure of the derivative based optimization schemes which are essential for efficient computation when there are many smoothing parameters to optimize. SCOP-splines circumvent this problem.

Other procedures based on B-splines were proposed by He and Shi (1998), Bollaerts et al. (2006), Rousson (2008), Wang and Meyer (2011). Meyer (2012) presented a cone projection method for estimating penalized B-splines with monotonicity or convexity constraints and proposed a GCV based test for checking the shape constrained assumptions. Monotonic regression within the Bayesian framework has been considered by Lang and Brezger (2004), Holmes and Heard (2003), Dunson and Neelon (2003), and Dunson (2005). In spite of their diversity these existing approaches also lack the ability to efficiently compute the smoothing parameter in a multiple smooth context. In addition, to our knowledge except for the bivariate constrained P-spline introduced by Bollaerts et al. (2006), multi-dimensional smooths under shape constraints on either all or a selection of the covariates have not yet been presented in the literature.

The remainder of the paper is structured as follows. The next section introduces SCOP-splines. Section 3.1 shows how SCAMs can be represented for estimation. A penalized likelihood maximization method for SCAM coefficient estimation is discussed in Sect. 3.2. Section 3.3 investigates the selection of multiple smoothing parameters. Interval estimation of the component smooth functions of the model is considered in Sect. 3.4. A simulation study is presented in Sect. 4 while Sect. 5 demonstrates applications of SCAM to two epidemiological examples.

2 SCOP-splines

2.1 B-spline background

In the smoothing literature B-splines are a common choice for the basis functions because of their smooth interpolation property, flexibility, and local support. The B-splines properties are thoroughly discussed in De Boor (1978). Eilers and Marx (1996) combined B-spline basis functions with discrete penalties in the basis coefficients to produce the popular 'P-spline' smoothers. Li and Ruppert (2008) established the corresponding asymptotic theory. Specifically that the rate of convergence of the penalized spline to a smooth function depends on an order of the difference penalty but not on a degree of B-spline basis and number of knots, given that the number of knots grows with the number of data and assuming the function is twice continuously differentiable. Ruppert (2002) and Li and Ruppert (2008) showed that the choice of the basis dimension is not critical but should be above some minimal level which depends on the spline degree. Asymptotic properties of P-splines were also studied in

Kauermann et al. (2009) and Claeskens et al. (2009). Here we propose to build on the P-spline idea to produce SCOP-splines.

2.2 One-dimensional case

The basic idea is most easily introduced by considering the construction of a monotonically increasing smooth, m, using a B-spline basis. Specifically let

$$m(x) = \sum_{j=1}^{q} \gamma_j B_j(x),$$

where q is the number of basis function, the B_j are B-spline basis functions of at least second order for representing smooth functions over interval $[a, b]$, based on equally spaced knots, and the γ_j are the spline coefficients.

It is well known that a sufficient condition for $m'(x) \geq 0$ over $[a, b]$ is that $\gamma_j \geq \gamma_{j-1} \forall\, j$ (see Supplementary material, S.1, for details). In the case of quadratic splines this condition is necessary. It is easy to see that this condition could be imposed by re-parameterizing, so that

$$\boldsymbol{\gamma} = \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}},$$

where $\boldsymbol{\beta} = \left[\beta_1, \beta_2, \ldots, \beta_q\right]^{\mathrm{T}}$ and $\tilde{\boldsymbol{\beta}} = \left[\beta_1, \exp(\beta_2), \ldots, \exp(\beta_q)\right]^{\mathrm{T}}$, while $\Sigma_{ij} = 0$ if $i < j$ and $\Sigma_{ij} = 1$ if $i \geq j$.

So if $\mathbf{m} = [m(x_1), m(x_2), \ldots, m(x_n)]^{\mathrm{T}}$ is the vector of m values at the observed points x_i, and $\mathbf{X}$ is the matrix such that $X_{ij} = B_j(x_i)$, then we have

$$\mathbf{m} = \mathbf{X}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}.$$

2.2.1 Smoothing

In a smoothing context we would also like to have a penalty on the $m(x)$ which can be used to control its 'wiggliness'. Eilers and Marx (1996) introduced the notion of directly penalizing differences in the basis coefficients of a B-spline basis, which is used with a relatively large q to avoid underfitting. We can adapt this idea here. For $j > 2$ our β_j are log differences in γ_j. We therefore propose penalizing the squared differences between adjacent β_j, starting from β_2, using the penalty $\|\mathbf{D}\boldsymbol{\beta}\|^2$ where $\mathbf{D}$ is the $(q-2) \times q$ matrix that is all zero except that $D_{i,i+1} = -D_{i,i+2} = 1$ for $i = 1, \ldots, q-2$. The penalty is zeroed when all the β_j after β_1 are equal, so that the γ_j form a uniformly increasing sequence and $m(x)$ is an increasing straight line (see Fig. 1). As a result our penalty shares with a second order P-spline penalty, the basic feature of 'smoothing towards a straight line', but in manner that is computationally convenient for constrained smoothing.

It might be asked whether penalization is necessary at all, given the restrictions imposed by the shape constraints?

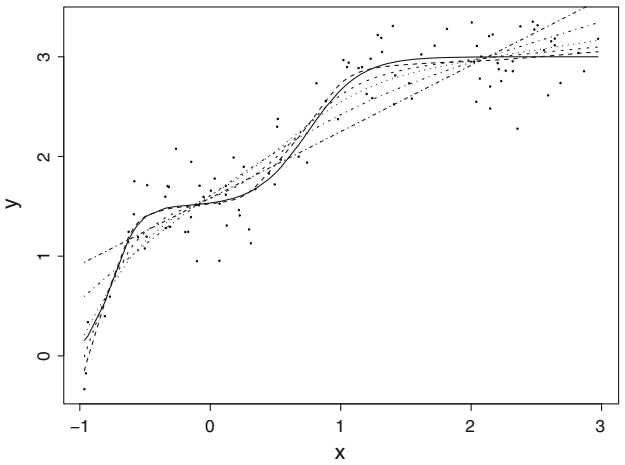

Fig. 1 Illustration of the SCOP-splines for five values of the smoothing parameter: $\lambda_1 = 10^{-4}$ (*long dashed curve*), $\lambda_2 = 0.005$ (*short dashed curve*), $\lambda_3 = 0.01$ (*dotted curve*), $\lambda_4 = 0.1$ (*dot-dashed curve*), and $\lambda_5 = 100$ (*two dashed curve*). The true curve is represented as a *solid line* and *dots* show the simulated data. Twenty five B-spline basis functions of the third order were used

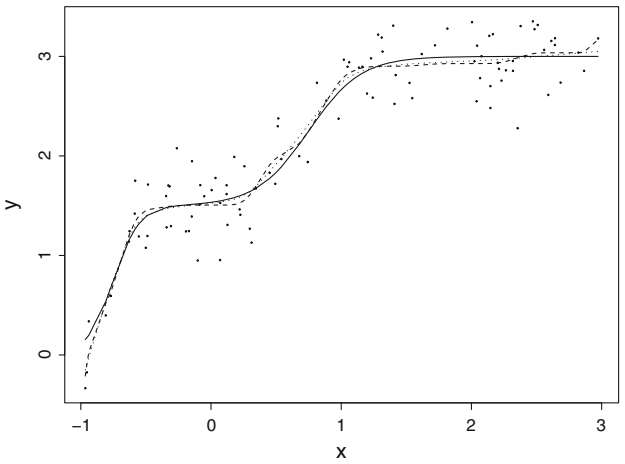

Fig. 2 Illustration of the SCOP-splines: un-penalized (*long dashed curve*, $\lambda = 0$), penalized (*dotted curve*, $\lambda = 10^{-4}$), and the true curve (*solid line*). Despite a monotonicity constraint, the un-penalized curve shows spurious detail that the penalty can remove

Figure 2 provides an illustration of what the penalty achieves. Even with shape constraint, the unpenalized estimated curve shows a good deal of spurious variation that the penalty removes.

2.2.2 Identifiability, basis dimension

If we were interested solely in smoothing one-dimensional Gaussian data then $\boldsymbol{\beta}$ would be chosen to minimize

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}\|^2 + \lambda\|\mathbf{D}\boldsymbol{\beta}\|^2,$$

where λ is a smoothing parameter controlling the trade-off between smoothness and fidelity to the response data $\mathbf{y}$. Here,

we are interested in the basis and penalty in order to be able to embed the shape constrained smooth $m(x)$ in a larger model. This requires an additional constraint on $m(x)$ in order to achieve identifiability to avoid confounding with the intercept of the model in which it is embedded. A convenient way to do this is to use centering constraints on the model matrix columns, i.e. the sum of the values of the smooth is set to be zero $\sum_{i=1}^{n} m(x_i) = 0$ or equivalently $\mathbf{1}^T \mathbf{X} \boldsymbol{\Sigma} \boldsymbol{\beta} = 0$.

As with any penalized regression spline approaches, the choice of the basis dimension, q, is not crucial but should be generous enough to avoid oversmoothing/underfitting (Ruppert 2002; Li and Ruppert 2008). Ruppert (2002) suggested algorithms for the basis dimension selection by minimizing GCV over a set of specified values of q, while Kauermann and Opsomer (2011) proposed an equivalent likelihood based scheme.

This simple monotonically increasing smooth can be extended to a variety of monotonic functions, including decreasing, convex/concave, increasing/decreasing and concave, increasing/ decreasing and convex, the difference between alternative shape constraints being the form of the matrices $\boldsymbol{\Sigma}$ and $\mathbf{D}$. Table 1 details eight possibilities, while Supplementary material, S.2, provides the corresponding derivations.

2.3 Multi-dimensional SCOP-splines

Using the concept of tensor product spline bases it is possible to build up smooths of multiple covariates under the monotonicity constraint, where monotonicity may be assumed on either all or a selection of the covariates. In this section the construction of a multivariable smooth, $m(x_1, x_2, \ldots, x_p)$, with multiple monotonically increasing constraints along all covariates is first considered, followed by a discussion of single monotonicity along a single direction.

2.3.1 Tensor product basis

Consider p B-spline bases of dimensions q_1, q_2, and q_p for representing marginal smooth functions, each of a single covariate

$$f_1(x_1) = \sum_{k_1=1}^{q_1} \alpha_{k_1}^1 B_{k_1}(x_1), \quad f_2(x_2) = \sum_{k_2=1}^{q_2} \alpha_{k_2}^2 B_{k_2}(x_2), \ldots,$$

$$f_p(x_p) = \sum_{k_p=1}^{q_p} \alpha_{k_p}^p B_{k_p}(x_p),$$

Table 1 Univariate shape constrained smooths

Shape constraints	$\boldsymbol{\Sigma}$	$\mathbf{D}$
Monotone increasing	$\Sigma_{ij} = \begin{cases} 0, & \text{if } i < j \\ 1, & \text{if } i \geq j \end{cases}$	$D_{i,i+1} = -D_{i,i+2} = 1$, $i = 1, \ldots, q-2$; $D_{ij} = 0$, otherwise
Monotone decreasing	$\Sigma_{ij} = \begin{cases} 0, & \text{if } i < j \\ 1, & \text{if } j = 1, \ i \geq 1 \\ -1, & \text{if } j \geq 2, \ i \geq j \end{cases}$	$D_{i,i+1} = -D_{i,i+2} = 1$, $i = 1, \ldots, q-2$; $D_{ij} = 0$, otherwise
Convex	$\Sigma_{ij} = \begin{cases} 0, & \text{if } i < j \\ 1, & \text{if } j = 1, \ i \geq 1 \\ -(i-1), & \text{if } j = 2, \ i \geq j \\ i-j+1, & \text{if } j \geq 3, \ i \geq j \end{cases}$	$D_{i,i+2} = -D_{i,i+3} = 1$, $i = 1, \ldots, q-3$; $D_{ij} = 0$, otherwise
Concave	$\Sigma_{ij} = \begin{cases} 0, & \text{if } i < j \\ 1, & \text{if } j = 1, \ i \geq 1 \\ i-1, & \text{if } j = 2, \ i \geq j \\ -(i-j+1), & \text{if } j \geq 3, \ i \geq j \end{cases}$	As above
increasing and convex	$\Sigma_{ij} = \begin{cases} 0, & \text{if } i < j \\ 1, & \text{if } j = 1, \ i \geq 1 \\ i-j+1, & \text{if } j \geq 2, \ i \geq j \end{cases}$	As above
Increasing and concave	$\Sigma_{ij} = \begin{cases} 0, & \text{if } i = 1, \ j \geq 2 \\ 1, & \text{if } j = 1, \ i \geq 1 \\ i-1, & \text{if } i \geq 2, \ j = 2, \ldots, q-1+2 \\ q-j+1, & \text{if } i \geq 2, \ j = q-i+3, \ldots, q \end{cases}$	As above
Decreasing and convex	$\Sigma_{ij} = \begin{cases} 0, & \text{if } i = 1, \ j \geq 2 \\ 1, & \text{if } j = 1, \ i \geq 1 \\ -(i-1), & \text{if } i \geq 2, \ j = 2, \ldots, q-1+2 \\ -(q-j+1), & \text{if } i \geq 2, \ j = q-i+3, \ldots, q \end{cases}$	As above
Decreasing and concave	$\Sigma_{ij} = \begin{cases} 0, & \text{if } i < j \\ 1, & \text{if } j = 1, \ i \geq 1 \\ -(i-j+1), & \text{if } j \geq 2, \ i \geq j \end{cases}$	As above

where $B_{k_j}(x_j)$, $j = 1, \ldots, p$, are B-spline basis functions, and $\alpha^j_{k_j}$ are spline coefficients. Then, following Wood (2006a) the multivariate smooth can be represented by expressing spline coefficients of the marginal smooths as the B-spline of the following covariate, starting from the first marginal smooth. By denoting $B_{k_1 \ldots k_p}(x_1, \ldots, x_p) = B_{k_1}(x_1) \cdot \ldots \cdot B_{k_p}(x_p)$, the smooth of p covariates may be written as follows

$$m(x_1, \ldots, x_p) = \sum_{k_1=1}^{q_1} \ldots \sum_{k_p=1}^{q_p} B_{k_1 \ldots k_p}(x_1, \ldots, x_p) \gamma_{k_1 \ldots k_p},$$

where $\gamma_{k_1 \ldots k_p}$ are unknown coefficients.

So if $\mathbf{X}$ is the matrix such that its ith row is $\mathbf{X}_i = \mathbf{X}_{1i} \otimes \mathbf{X}_{2i} \otimes \cdots \otimes \mathbf{X}_{pi}$, where $\otimes$ denotes a Kronecker product, and $\boldsymbol{\gamma} = (\gamma_{11\ldots1}, \ldots, \gamma_{k_1 k_2 \ldots k_p}, \ldots, \gamma_{q_1 q_2 \ldots q_p})^T$, then

$$\mathbf{m} = \mathbf{X}\boldsymbol{\gamma}.$$

2.3.2 Constraints

By extending the univariate case one can see that a sufficient condition for $\partial f(x_1, \ldots, x_p)/\partial x_j \geq 0$ is $\gamma_{k_1 \ldots k_j \ldots k_p} \geq \gamma_{k_1 \ldots (k_j - 1) \ldots k_p}$. To impose these conditions the re-parametrization $\boldsymbol{\gamma} = \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}$ is proposed, where

$$\tilde{\boldsymbol{\beta}} = \left[\beta_{11\ldots1}, \exp(\beta_{11\ldots2}), \ldots, \exp(\beta_{k_1 \ldots k_p}), \ldots, \exp(\beta_{q_1 \ldots q_p})\right]^{\mathrm{T}},$$

and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 \otimes \boldsymbol{\Sigma}_2 \otimes \cdots \otimes \boldsymbol{\Sigma}_p$. The elements of $\boldsymbol{\Sigma}_j$ are the same as for the univariate monotonically increasing smooth (see Table 1). For the multiple monotonically decreasing multivariate function $\boldsymbol{\Sigma} = \left[\mathbf{1} : \boldsymbol{\Sigma}'_{(,-1)}\right]$, where $\boldsymbol{\Sigma}' = -\boldsymbol{\Sigma}_1 \otimes \boldsymbol{\Sigma}_2 \otimes \cdots \otimes \boldsymbol{\Sigma}_p$, that is $\boldsymbol{\Sigma}$ is a matrix $\boldsymbol{\Sigma}'$ with the first column replaced by the column of one's.

To satisfy conditions for a monotonically increasing or decreasing smooth with respect to only one covariate the following re-parameterizations are suggested:

1. For the single monotonically increasing constraint along the x_j direction:
   Let $\boldsymbol{\Sigma}_j$ be defined as previously while $\mathbf{I}_s$ is an identity matrix of size q_s, $s \neq j$, then

   $$\boldsymbol{\Sigma} = \mathbf{I}_1 \otimes \cdots \otimes \boldsymbol{\Sigma}_j \otimes \cdots \otimes \mathbf{I}_p,$$

   and $\boldsymbol{\gamma} = \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}$, where $\tilde{\boldsymbol{\beta}}$ is a vector containing a mixture of un-exponentiated and exponentiated coefficients with $\tilde{\beta}_{k_1 \ldots k_j \ldots k_p} = \exp(\beta_{k_1 \ldots k_j \ldots k_p})$ when $k_j \neq 1$.
2. For the single monotonically decreasing constraint along the x_j direction:
   The re-parametrization is the same as above except for the representation of the matrix $\boldsymbol{\Sigma}_j$ which is as for univariate smooth with monotonically decreasing constraint (see Table 1).

By analogy it is not difficult to construct tensor products with monotonicity constraints along any number of covariates.

2.3.3 Penalties

For controlling the level of smoothing, the penalty introduced in Sect. 2 can be extended. For multiple monotonicity the penalties may be written as

$$\mathscr{P} = \lambda_1 \boldsymbol{\beta}^T \mathbf{S}_1 \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^T \mathbf{S}_2 \boldsymbol{\beta} + \cdots + \lambda_p \boldsymbol{\beta}^T \mathbf{S}_p \boldsymbol{\beta},$$

where $\mathbf{S}_j = \mathbf{D}_j^{\mathrm{T}} \mathbf{D}_j$ and $\mathbf{D}_j = \mathbf{I}_1 \otimes \mathbf{I}_2 \otimes \cdots \otimes \mathbf{D}_{mj} \otimes \cdots \otimes \mathbf{I}_p$. $\mathbf{D}_{mj}$ is as $\mathbf{D}$ in Table 1 for a monotone smooth. Penalties for single monotonicity along x_j are

$$\mathscr{P} = \lambda_1 \boldsymbol{\beta}^T \tilde{\mathbf{S}}_1 \boldsymbol{\beta} + \cdots + \lambda_j \boldsymbol{\beta}^T \mathbf{S}_j \boldsymbol{\beta} + \lambda_p \boldsymbol{\beta}^T \tilde{\mathbf{S}}_p \boldsymbol{\beta},$$

where $\mathbf{S}_j$ is defined as above. The penalty matrices $\tilde{\mathbf{S}}_i$, $i \neq j$, in the unconstrained directions can be constructed using the marginal penalty approach described in Wood (2006a). The degree of smoothness in the unconstrained directions can be controlled by the second-order difference penalties applied to the non-exponentiated coefficients, and by the first-order difference penalties for the exponentiated coefficients. As in the univariate case, these penalties keep the parameter estimates close to each other, resulting in similar increments in the coefficients of marginal smooths. When $\lambda_j \to \infty$ such penalization results in straight lines for marginal curves.

3 SCAM

3.1 SCAM representation

To represent (1) for computation we now choose basis expansions, penalties and identifiability constraints for all the unconstrained f_j, as described in detail in Wood (2006a), for example. This allows $\sum_j f_j(z_{ji})$ to be replaced by $\mathbf{F}_i \boldsymbol{\gamma}$, where $\mathbf{F}$ is a model matrix determined by the basis functions and the constraints, and $\boldsymbol{\gamma}$ is a vector of coefficients to be estimated. The penalties on the f_j are quadratic in $\boldsymbol{\gamma}$.

Each shape constrained term m_k is represented by a model matrix of the form $\mathbf{X}\boldsymbol{\Sigma}$ and corresponding coefficient vector. Identifiability constraints are absorbed by the column centering constraints. The model matrices for all the m_k are then combined so that we can write

$$\sum_k m_k(x_{ki}) = \mathbf{M}_i \tilde{\boldsymbol{\beta}},$$

where $\mathbf{M}$ is a model matrix and $\tilde{\boldsymbol{\beta}}$ is a vector containing a mixture of model coefficients (β_i) and exponentiated model

coefficients ($\exp(\beta_i)$). The penalties in this case are quadratic in the coefficients $\boldsymbol{\beta}$ (not in the $\tilde{\boldsymbol{\beta}}$).

So (1) becomes

$$g(\mu_i) = \mathbf{A}_i\boldsymbol{\theta} + \mathbf{F}_i\boldsymbol{\gamma} + \mathbf{M}_i\tilde{\boldsymbol{\beta}}, \quad Y_i \sim \mathrm{EF}(\mu_i, \phi).$$

For fitting purposes we may as well combine the model matrices column-wise into one model matrix $\mathbf{X}$, and write the model as

$$g(\mu_i) = \mathbf{X}_i\tilde{\boldsymbol{\beta}}, \tag{2}$$

where $\tilde{\boldsymbol{\beta}}$ has been enlarged to now contain $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ and the original $\tilde{\boldsymbol{\beta}}$. Similarly there is a corresponding expanded model coefficient vector $\boldsymbol{\beta}$ containing $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ and the original $\boldsymbol{\beta}$. The penalties on the terms have the general form $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}_\lambda\boldsymbol{\beta}$ where $\mathbf{S}_\lambda = \sum_k \lambda_k\mathbf{S}_k$, and the $\mathbf{S}_k$ are the original penalty matrices expanded with zeros everywhere except for the elements which correspond to the coefficients of the kth smooth.

3.2 SCAM coefficient estimation

Now consider the estimation of $\boldsymbol{\beta}$ given values for the smoothing parameters $\boldsymbol{\lambda}$. The exponential family chosen determines the form of the log likelihood $l(\boldsymbol{\beta})$ of the model, and to control the degree of model smoothness we seek to maximize its penalized version

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}_\lambda\boldsymbol{\beta}/2.$$

However, the non-linear dependence of $\mathbf{X}\tilde{\boldsymbol{\beta}}$ on $\boldsymbol{\beta}$ makes this more difficult than in the case of unconstrained GAMs. In particular we found that optimization via Fisher scoring caused convergence problems for some models, and we therefore use a full Newton approach. The special structure of the model means that it is possible to work entirely in terms of a matrix square root of the Hessian of l, when applying Newton's method, thereby improving the numerical stability of computations, so we also adopt this refinement. Also, since SCAM is very much within GAM theory, the same convergence issues might arise as in the case of GAM/GLM fitting (Wood 2006a). In particular, the likelihood might not be uni-modal and the process may converge to different estimates depending on the starting values of the fitting process. However, if the initial values are reasonably selected then it is unlikely that there will be major convergence issues. The following algorithm suggests such initial values.

Let $V(\mu)$ be the variance function for the model's exponential family distribution, and define

$$\alpha(\mu_i) = 1 + (y_i - \mu_i)\left\{\frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)}\right\}.$$

Penalized likelihood maximization is then achieved as follows:

1. To obtain an initial estimate of $\boldsymbol{\beta}$, minimize $\|g(\mathbf{y}) - \mathbf{X}\tilde{\boldsymbol{\beta}}\|^2 + \tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{S}_\lambda\tilde{\boldsymbol{\beta}}$ w.r.t. $\tilde{\boldsymbol{\beta}}$, *subject to linear inequality constraints* ensuring that $\tilde{\beta}_j > 0$ whenever $\tilde{\beta}_j = \exp(\beta_j)$. This is a standard quadratic programming (QP) problem. (If necessary $\mathbf{y}$ is adjusted slightly to avoid infinite $g(\mathbf{y})$.)
2. Set k = 0 and repeat the steps 3–11 to convergence…
3. Evaluate $z_i = (y_i - \mu_i)g'(\mu_i)/\alpha(\mu_i)$ and $w_i = \omega_i\alpha(\mu_i)/\{V(\mu_i)g'^2(\mu_i)\}$, using the current estimate of μ_i.
4. Evaluate vectors $\tilde{\mathbf{w}} = |\mathbf{w}|$ and $\tilde{\mathbf{z}}$ where $\tilde{z}_i = \mathrm{sign}(w_i)z_i$.
5. Evaluate the diagonal matrix $\mathbf{C}$ such that $C_{jj} = 1$ if $\tilde{\beta}_j = \beta_j$, and $C_{jj} = \exp(\beta_j)$ otherwise.
6. Evaluate the diagonal matrix $\mathbf{E}$ such that $E_{jj} = 0$ if $\tilde{\beta}_j = \beta_j$, and $E_{jj} = \sum_i^n w_i g'(\mu_i)[\mathbf{XC}]_{ij}(y_i - \mu_i)/\alpha(\mu_i)$ otherwise.
7. Let $\mathbf{I}^-$ be the diagonal matrix such that $I^-_{ii} = 1$ if $w_i < 0$ and $I^-_{ii} = 0$ otherwise.
8. Letting $\tilde{\mathbf{W}}$ denote $\mathrm{diag}(\tilde{\mathbf{w}})$, form the QR decomposition

$$\begin{bmatrix}\sqrt{\tilde{\mathbf{W}}}\mathbf{XC}\\ \mathbf{B}\end{bmatrix} = \mathbf{QR},$$

   where $\mathbf{B}$ is any matrix square root such that $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{S}_\lambda$.
9. Letting $\mathbf{Q}_1$ denote the first n rows of $\mathbf{Q}$, form the symmetric eigen-decomposition

$$\mathbf{Q}_1^{\mathrm{T}}\mathbf{I}^-\mathbf{Q}_1 + \mathbf{R}^{-\mathrm{T}}\mathbf{E}\mathbf{R}^{-1} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}.$$

10. Hence define $\mathbf{P} = \mathbf{R}^{-1}\mathbf{U}(\mathbf{I} - \boldsymbol{\Lambda})^{-1/2}$ and $\mathbf{K} = \mathbf{Q}_1\mathbf{U}(\mathbf{I} - \boldsymbol{\Lambda})^{-1/2}$.
11. Update the estimate of $\boldsymbol{\beta}$ as $\boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} + \mathbf{PK}^{\mathrm{T}}\sqrt{\tilde{\mathbf{W}}}\tilde{\mathbf{z}} - \mathbf{PP}^{\mathrm{T}}\mathbf{S}_\lambda\boldsymbol{\beta}^{[k]}$ and increment k.

The algorithm is derived in Appendix 1 which shows several similarities to a standard penalized IRLS scheme for penalized GLM estimation. However, the more complicated structure results from the need to use full Newton, rather than Fisher scoring, while at the same time avoiding computation of the full Hessian, which would approximately square the condition number of the update computations.

Two refinements of the basic iteration may be required.

1. If the Hessian of the log likelihood is indefinite then step 10 will fail, because some Λ_{ii} will exceed 1. In this case a Fisher update step must be substituted, by setting $\alpha(\mu_i) = 1$.
2. There is considerable scope for identifiability issues to hamper computation. In common with unconstrained GAMs, flexible SCAMs with highly correlated covariates can display co-linearity problems between model coefficients, which require careful handling numerically, in order to ensure numerical stability of the estimation

algorithms. An additional issue is that the non-linear constraints mean that parameters can be poorly identified on flat sections of a fitted curve, where β is simply 'very negative', but the data contain no information on how negative. So steps must be taken to deal with unidentifiable parameters. One approach is to work directly with the QR decomposition to calculate which coefficients are unidentifiable at each iteration and to drop these, but a simpler strategy substitutes a singular value decomposition for the R factor at step 8 if it is rank deficient, so that

$$\mathbf{R} = \mathscr{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}.$$

Then we set $\mathscr{Q} = \mathbf{Q}\mathscr{U}$, $\mathscr{R} = \mathbf{D}\mathbf{V}^{T}$, and $\mathbf{Q}_1$ is the first n rows of $\mathscr{Q}$, and everything proceeds as before, except for the inversion of $\mathbf{R}$. We now substitute the pseudoinverse $\mathbf{R}^{-} = \mathbf{V}\mathbf{D}^{-}$, where the diagonal matrix $\mathbf{D}^{-}$ is such that $D^{-}_{jj} = 0$ if the singular value D_{jj} is 'too small', but otherwise $D^{-}_{jj} = 1/D_{jj}$. 'Too small' is judged relative to the largest singular value D_{11} multiplied by some power (in the range .5 to 1) of the machine precision. If all parameters are numerically identifiable then the pseudoinverse is just the inverse.

3.3 SCAM smoothing parameter estimation

We propose to estimate the smoothing parameter vector $\boldsymbol{\lambda}$ by optimizing a prediction error criterion such as AIC (Akaike 1973) or GCV (Craven and Wahba 1979). The model deviance is defined in the standard way as

$$D(\hat{\boldsymbol{\beta}}) = 2\{l_{\max} - l(\hat{\boldsymbol{\beta}})\}\phi,$$

where $l_{\max}$ is the saturated log likelihood. When the scale parameter is known we find $\boldsymbol{\lambda}$ which minimizes $\mathscr{V}_u = D(\hat{\boldsymbol{\beta}}) + 2\phi\gamma\tau$, where τ is the effective degrees of freedom (edf) of the model. γ is a parameter that in most cases has the value of 1, but is sometimes increased above 1 to obtain smoother models [see Kim and Gu (2004)]. When the scale parameter is unknown we find $\boldsymbol{\lambda}$ minimizing the GCV score, $\mathscr{V}_g = nD(\hat{\boldsymbol{\beta}})/(n - \gamma\tau)^2$. For both criteria the dependence on $\boldsymbol{\lambda}$ is via the dependence of τ and $\hat{\boldsymbol{\beta}}$ on $\boldsymbol{\lambda}$ [see Hastie and Tibshirani (1990) and Wood (2008) for further details].

The edf can be found, following Meyer and Woodroofe (2000) as

$$\tau = \sum_{i=1}^{n} \frac{\partial \hat{\mu}_i}{\partial y_i} = \mathrm{tr}(\mathbf{K}\mathbf{K}^{\mathrm{T}}\mathbf{L}^{+}), \tag{3}$$

where $\mathbf{L}^{+}$ is the diagonal matrix such that

$$L^{+}_{ii} = \begin{cases} \alpha(\mu_i)^{-1}, & \text{if } w_i \geq 0 \\ -\alpha(\mu_i)^{-1}, & \text{otherwise.} \end{cases}$$

Details are provided in Appendix 2.

Optimization of the $\mathscr{V}_*$ w.r.t. $\boldsymbol{\rho} = \log(\boldsymbol{\lambda})$ can be achieved by a quasi-Newton method. Each trial $\boldsymbol{\rho}$ vector will require a Sect. 3.2 iteration to find the corresponding $\hat{\boldsymbol{\beta}}$ so that the criterion can be evaluated. In addition the first derivative vector of $\mathscr{V}_*$ w.r.t. $\boldsymbol{\rho}$ will be required, which in turn requires $\partial\hat{\boldsymbol{\beta}}/\partial\boldsymbol{\rho}$ and $\partial\tau/\partial\rho$.

As demonstrated in Supplementary material, S.3, implicit differentiation can be used to obtain

$$\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\rho_k} = -\lambda_k\mathbf{P}\mathbf{P}^{\mathrm{T}}\mathbf{S}_k\hat{\boldsymbol{\beta}}.$$

The derivatives of D and τ then follow, as S.4 (Supplementary material), shows in tedious detail.

3.4 Interval estimation

Having obtained estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\lambda}}$, we have point estimates for the component smooth functions of the model, but it is usually desirable to obtain interval estimates for these functions as well. To facilitate the computation of such intervals we seek distributional results for the $\tilde{\boldsymbol{\beta}}$, i.e. for the coefficients on which the estimated functions depend linearly.

Here we adopt the Bayesian approach to interval estimation pioneered in Wahba (1983), but following Silverman's (1985) formulation. Such intervals are appealing following Nychka's (1988) analysis showing that they have good frequentist properties by virtue of accounting for both sampling variability and smoothing bias. Specifically, we view the smoothness penalty as equivalent to an improper prior distribution on the model coefficients

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{S}^{-}_{\lambda}/(2\phi)),$$

where $\mathbf{S}^{-}_{\lambda}$ is the Moore–Penrose pseudoinverse of $\mathbf{S}_{\lambda} = \sum_k \lambda_k\mathbf{S}_k$. In conjunction with the model likelihood, Bayes theorem then leads to the *approximate* result

$$\tilde{\boldsymbol{\beta}}|\mathbf{y} \sim N(\hat{\tilde{\boldsymbol{\beta}}}, \mathbf{V}_{\tilde{\boldsymbol{\beta}}}), \tag{4}$$

where $\mathbf{V}_{\tilde{\boldsymbol{\beta}}} = \mathbf{C}(\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\mathbf{C} + \mathbf{S}_{\lambda})^{-1}\mathbf{C}\phi$, and $\mathbf{W}$ is the diagonal matrix of w_i calculated with $\alpha(\mu_i) = 1$. Supplementary material, S.5, derives this result. The deviance or Pearson statistic divided by the effective residual degrees of freedom provides an estimate of ϕ, if required. To use the result we condition on the smoothing parameter estimates: the intervals display surprisingly good coverage properties despite this (Marra and Wood (2012), provide a theoretical analysis which partly explains this).

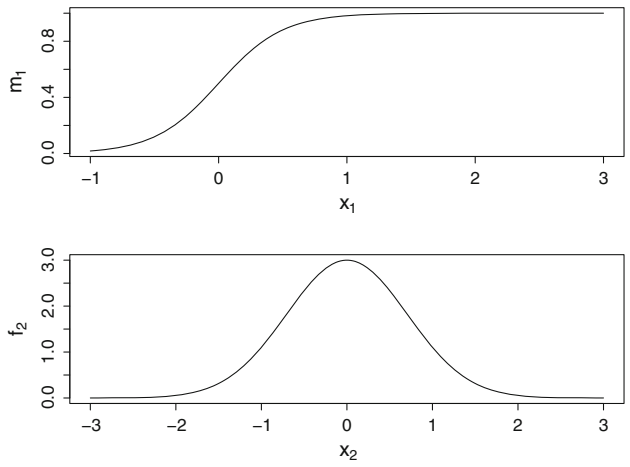

Fig. 3 Shape of the functions used for the simulation study

4 Simulated examples

4.1 Simulations: comparison with alternative methods

In this section performance comparison with unconstrained GAM and the QP approach to shape preserving smoothing (Wood 1994) is illustrated on a simulated example of an additive model with a mixture of monotone and unconstrained smooth terms. All simulation studies and data applications were performed using R packages `scam`, which implements the proposed SCAM approach, and `mgcv` for GAM and QP implementations. A more extensive simulation study is given in Supplementary material, S.6. Particularly, the first subsection of S.6 shows comparative study with the constrained P-spline regression (Bollaerts et al. 2006), monotone piecewise quadratic splines of Meyer (2008), and shape-restricted penalized B-splines of Meyer (2012) on simulated example on univariate single smooth term models. Since there was no mean square error advantage of these approaches over the SCAM for the univariate model, and moreover, the direct grid search for multiple optimal smoothing parameter is computational expensive, (and to the authors' knowledge, R routines for the implementation of these methods are not freely available) the comparison for multivariate and additive examples were performed only with the unconstrained GAM and QP approach.

The following additive model is considered:

$$g(\mu_i) = m_1(x_{1i}) + f_2(x_{2i}), \quad \mathrm{E}(Y_i) = \mu_i, \tag{5}$$

where $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$ or $\mathrm{Poi}(\mu_i)$ distribution. Figure 3 illustrates the graphs of the true functions used for this study. Their analytical expressions are given in Supplementary material, S.4.

The covariate values, x_{1i} and x_{2i}, were simulated from the uniform distribution on $[-1, 3]$ and $[-3, 3]$ respectively. For the Gaussian data the values of σ were 0.05, 0.1, 0.2, which gave the signal to noise ratios of about 0.97, 0.88, and 0.65. For the Poisson model the noise level was controlled by multiplying $g(\mu_i)$ by d, taking values 0.5, 0.7, 1.2, which resulted in the signal to noise ratios of about 0.58, 0.84, and 0.99. For the SCAM implementation a cubic SCOP-spline of the dimension 30 was used to represent the first monotonic smooth term and a cubic P-spline with $q = 15$ for the second unconstrained term. For an unconstrained GAM, P-splines with the same basis dimensions were used for both model components. The models were fitted by penalized likelihood maximization with the smoothing parameter selected using $\mathcal{V}_g$ in the Gaussian case and $\mathcal{V}_u$ for the Poisson case.

For implementing the QP approach to monotonicity preserving constraint, we approximated the necessary and sufficient condition $f'(x) \geq 0$, via the standard technique (Villalobos and Wahba 1987) of using a fine grid of linear constraints $(f'(x_i^*) \geq 0, i = 1, \ldots, n)$, where x_i^* are spread evenly through the range of x (strictly such constraints are necessary, but only sufficient as $n \to \infty$, but in practice we observed no violations of monotonicity). Cubic regression spline bases were used here together with the integrated squared second order derivative of the smooth as the penalty. The model fit is obtained by setting the QP problem within a penalized IRLS loop given $\boldsymbol{\lambda}$ chosen via GCV/UBRE from unconstrained model fit. Cubic regression splines tend to have slightly better MSE performance than P-splines (Wood 2006a) and moreover, the conditions built on finite differences are not only sufficient but also necessary for monotonicity. So this is a challenging test for SCAM. Three hundred replicates were produced for Gaussian and Poisson distributions at each of three levels of noise and for two sample sizes, 100, 200, for the three alternative approaches.

The simulation results for the Gaussian data are illustrated in Fig. 4. The results show that SCAM works better than the other two alternative methods in the sense of MSE performance. Note that the performance of GAM was better than the performance of the QP approach in this case, but the difference in MSE between SCAM and GAM is much less than that in the one-dimensional simulation studies shown in Supplementary material, S.6. Also it is noticeable that GAM reconstructed the truth better than the QP method. The explanation may be due to there being only one monotonic term, and both GAM and SCAM gave similar fits for the unconstrained term, f_2. At lower noise levels GAM might also be able to reconstruct the monotone shape of m_1 for some replicates. The results also suggest that the SCAM works better than GAM for greater levels of noise which seems to be natural since at lower noise levels the shapes of constrained terms can be captured by the unconstrained GAM. The reduction in performance of the QP compared to GAM was due to the smoothing parameter estimation from the unconstrained fit which sometimes resulted in less smooth

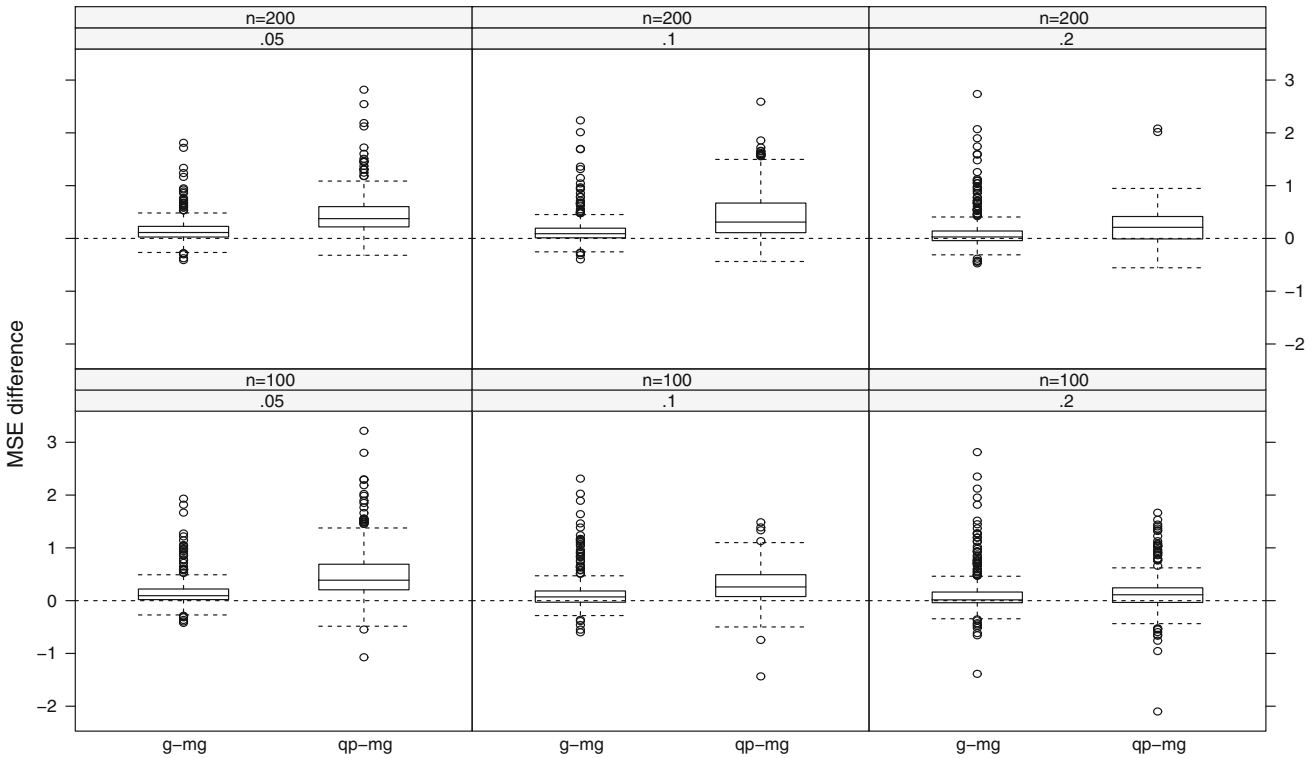

Fig. 4 MSE comparisons between SCAM (mg), GAM (g), and quadratic programming (qp) approaches for the Gaussian distribution for each of three noise levels. The upper panel illustrates the results for $n = 200$, the lower for $n = 100$. *Boxplots* show the distributions of differences in relative MSE between each alternative method and SCAM. 300 replicates were used. Relative MSE was calculated by dividing the MSE value by the average MSE of SCAM for the given case

tails of the smooth term than those of the unconstrained GAM. For the Poisson data of the samples size $n = 100$ all three methods worked similarly, but with an increase in sample size SCAM outperformed the other two approaches (plots are not shown). As in the Gaussian case the unconstrained GAM worked better than QP.

The simulation studies show that SCAM may have practical advantages over the alternative methods considered. It is computationally slower than GAM and QP approaches, however, obviously GAM cannot impose monotonicity, and the selection of the smoothing parameter for SCAM is well founded, in contrast to the ad hoc method used with QP of choosing λ from an unconstrained fit, and then refitting subject to constraint. Finally, the practical MSE performance of SCAM seems to be better than that of the alternatives considered here.

4.2 Coverage probabilities

The proposed Bayesian approach for confidence intervals construction makes a number of key assumptions; (i) it uses linear approximation of the exponentiated parameters, and in the case of non-Gaussian models adopts large sample inference; (ii) the smoothing parameters are treated as fixed. The simulation example of the previous subsection is used in order to examine how these restrictions affect the performance of the confidence intervals. The realized coverage probabilities is taken as a measure of their performance. Supplementary material, S.7, demonstrates two other examples for more thorough confidence interval performance presentation.

The simulation study of confidence interval performance is conducted in an analogous manner to Wood (2006b). Samples of sizes $n = 200$ and 500 were generated from (5) for Gaussian and Poisson distributions. 500 replicates were produced for both distributions at each of three levels of noise and for two sample sizes. For each replicate the realized coverage proportions were calculated as the proportions of the values of the true functions (at each of the covariate values) falling within the constructed confidence interval. Three confidence levels were considered 90, 95, 99 %. An overall mean coverage probability and its standard error were obtained from the 500 'across-the-function' coverage proportions. The results of the study are presented in Fig. 5 for the Gaussian and Poisson models. The realized coverage probabilities are near the corresponding nominal values, the larger sample size reduces the standard errors as expected. The results for the Poisson models are quite good with an

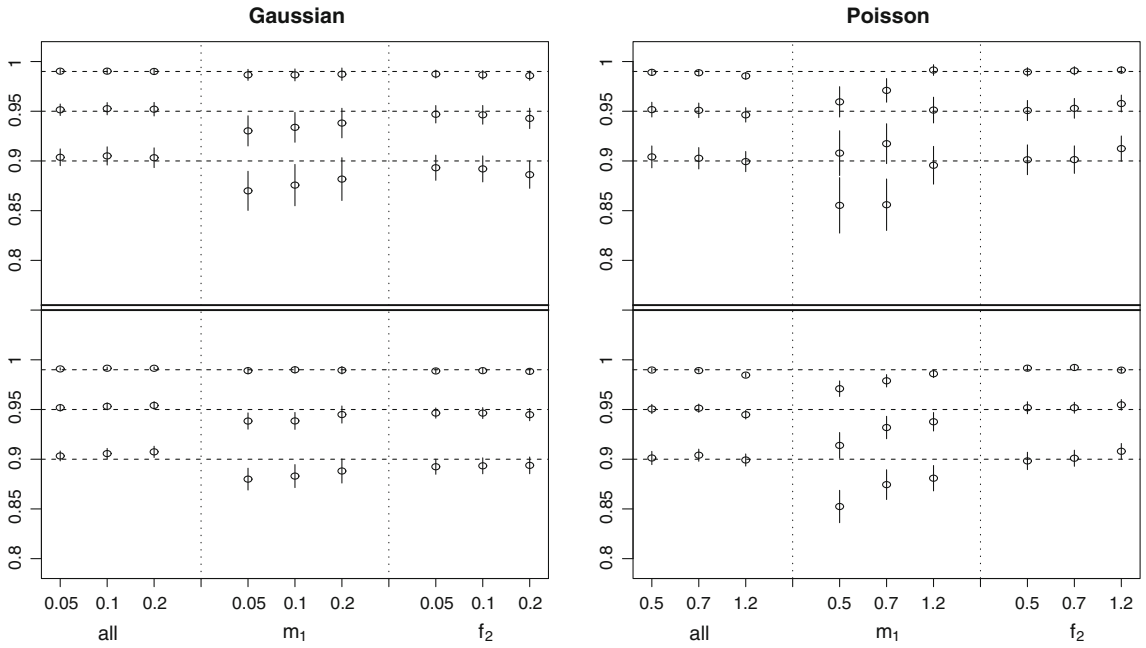

Fig. 5 Realized coverage probabilities for confidence intervals from the SCAM simulation study of the first example, for normal and Poisson data for $n = 200$ (*top panel*) and $n = 500$ (*bottom panel*). Three noise levels are used for each smooth term and for the overall model ("all"). The nominal coverage probabilities of 0.90, 0.95, and 0.99, are shown as *horizontal dashed lines*. '∘' indicates the average realized coverage probabilities over 500 replicate data sets. *Vertical lines* show twice standard error intervals of the mean coverage probabilities

exception for the first monotone smooth, $m_1(x_1)$, for the low signal strength, which may be explained by the fact that the optimal fit inclines toward a straight line model (Marra and Wood 2012).

5 Examples

This section presents the application of SCAM to two different data sets. The purpose of the first application is to investigate whether proximity to municipal incinerators in Great Britain is associated with increased risk of stomach cancer (Elliott et al. 1996; Shaddick et al. 2007). It is hypothesized that the risk of cancer is a decreasing function of distance from an incinerator. The second application uses data from the National Morbidity, Mortality, and Air Pollution Study (Peng and Welty 2004). The relationship between daily counts of mortality and short-term changes in air pollution concentrations is investigated. It is assumed that increases in concentrations of ozone, sulphur dioxide, particular matter will be associated with adverse health effects.

Incinerator data: Elliott et al. (1996) presented a large-scale study to investigate whether proximity to incinerators is associated with an increased risk of cancer. They analyzed data from 72 municipal solid waste incinerators in Great Britain and investigated the possibility of a decline in risk with distance from sources of pollution for a number of cancers. There was significant evidence for such a decline for stomach cancer, among several others. Data from a single incinerator from those 72 sources, located in the northeast of England, are analyzed using the SCAM approach in this section. This incinerator had a significant result indicating a monotone decreasing risk with distance (Elliott et al. 1996).

The data are from 44 enumeration districts (census-defined administrative areas), ED, whose geographical centroids lay within 7.5 km of the incinerator. The response variable, Y_i, are the observed numbers of cases of stomach cancer for each enumeration district. Associated estimates of the expected number of cases, E_i, available for risk determination, $\texttt{risk}_i = Y_i/E_i$, obtained using national rates for the whole of Great Britain, standardized for age and sex, were calculated for each ED. The two covariates are the distance (km), $\texttt{dist}_i$, from the incinerator and a deprivation score, the Carstairs score, $\texttt{cs}_i$.

Under the model, it is assumed that, Y_i are independent Poisson variables, $Y_i \sim \text{Poi}(\mu_i)$, where $\mu_i = \lambda_i E_i$, μ_i is the rate of the Poisson distribution with E_i the expected number of cases (in area i) and λ_i the relative risk. Shaddick et al. (2007) proposed a model under which the effect of a covariate, e.g., distance, on cancer risk was linear through an exponential function, i.e. $\lambda_i = \exp(\beta_0 + \beta_1 \texttt{dist}_i)$. Since the risk of cancer might be expected to decrease with distance from the incinerator, in this paper a smooth monotonically decreasing function, $m(\texttt{dist}_i)$, is suggested for modelling its rela-

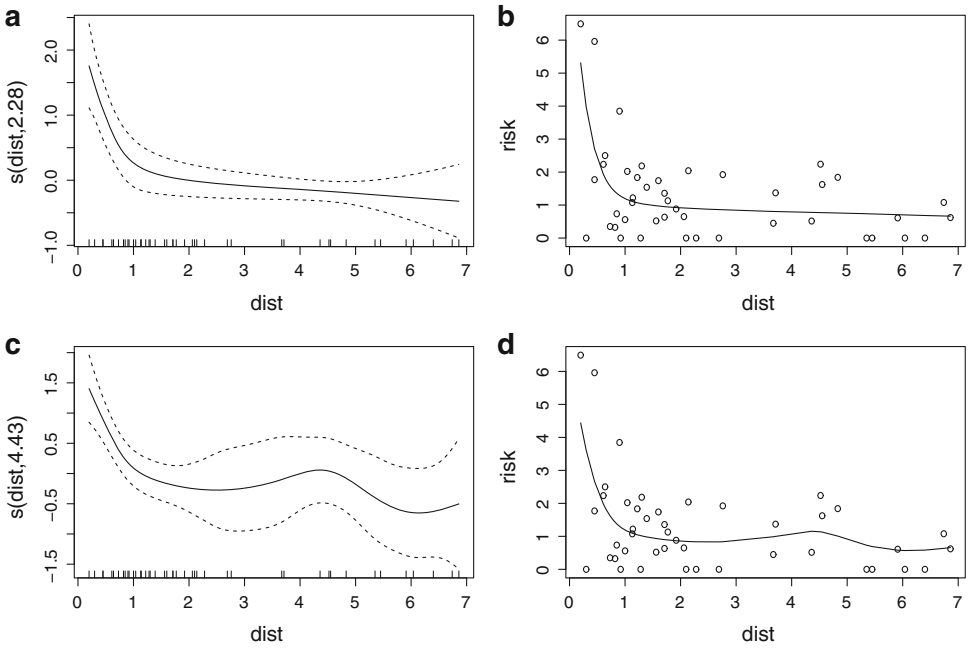

Fig. 6 The estimated smooth and cancer risk function for monotone and unconstrained versions of model 1 (incinerator data). **a** the estimated smooth of SCAM + 95 % confidence interval; **b** the SCAM estimated risk as the function of distance; **c** the GAM estimated smooth + 95 % confidence interval; **d** the GAM estimated risk as the function of distance. *Points* show the observed data. As noted in the text, AIC suggests that the shape constrained model (**a**, **b**) is better than the unconstrained version (**c**, **d**)

tionship with the distance $\lambda_i = \exp\{m(\texttt{dist}_i)\}$. Hence, the model can be represented as the following:

$$\log(\lambda_i) = m(\texttt{dist}_i) \quad \Rightarrow \quad \log(\mu_i/E_i) = m(\texttt{dist}_i) \quad \Rightarrow \quad \log(\mu_i) = \log(E_i) + m(\texttt{dist}_i),$$

which is a single smooth generalized Poisson regression model under monotonicity constraint, where $\log(E_i)$ is treated as an offset (a variable with a coefficient equal to 1). Therefore, the SCAM approach can be applied to fit such a model. Carstairs score is known to be a good predictor of cancer rates (Elliott et al. 1996; Shaddick et al. 2007), so its effect may also be included in the model. The following four models are considered for this application.

Model 1: $\log\{\mathrm{E}(Y_i)\} = \log(E_i) + m_1(\texttt{dist}_i)$, $m_1'(\texttt{dist}_i) < 0$. Model 2 is the same as model 1 but with $m_2(\texttt{cs}_i)$ as its smooth term instead with $m_2'(\texttt{cs}_i) > 0$. Model 3 combines both smooths while model 4 takes a bivariate function $m_3(-\texttt{dist}_i, \texttt{cs}_i)$ subject to double monotone increasing constraint. The univariate smooth terms were represented by the third order SCOP-splines with $q = 15$, while $q_1 = q_2 = 6$ were used for the bivariate SCOP-spline.

Plots for assessing the suitability of model one are given in Supplementary material, S.8. The first model for comparison has been also fitted without constraint. The estimated smooths and risk functions for both methods are illustrated in Fig. 6. The estimate of the cancer risk function was obtained by $\hat{\texttt{risk}}_i = \hat{\mu}_i/E_i = \exp\{\hat{m}_1(\texttt{dist}_i)\}$. Note, that the unconstrained GAM resulted in a non-monotone smooth, which supports the SCAM approach. The AIC score allows us to compare models with and without shape constraints. The AIC values were 152.35 for GAM and 150.57 for SCAM which favoured the shape constrained model.

In model 2 the number of cases of stomach cancer are represented by a smooth function of deprivation score. This function is assumed to be monotonically increasing since it was shown (Elliott et al. 1996) that in general, people living closer to incinerators tend to be less affluent (low Carstairs score). The AIC value for this model was 155.59, whereas the unconstrained version gave AIC = 156.4, both of which were higher than for the previous model. The other three measures of the model performance, $\mathscr{V}_u$, the adjusted r^2, and the deviance explained, also gave slightly worse results than those seen in model 1.

Model 3 incorporates both covariates, `dist` and `cs`, assuming an additive effect on log scale. The estimated edf of $m_2(\texttt{cs})$ was about zero. This smoothing term was insignificant in this model, with all its coefficients near zero. This can be explained by a high correlation between two covariates. Considering a linear effect of Carstairs in place of the smooth function, m_2, as it was proposed in Shaddick et al. (2007), $\log\{\mathrm{E}(Y_i)\} = \log(E_i) + m_1(\texttt{dist}_i) + \beta\texttt{cs}_i$, also resulted in an insignificant value for β.

The bivariate function, $m_3(-\texttt{dist}_i, \texttt{cs}_i)$, is considered in the last model. The perspective plot of the estimated smooth is illustrated in Fig. 7. This plot also supports the previous result, that the Carstairs score does not provide any additional information for modelling cancer risk when distance is included in the model. The graph of the estimated smooth has almost no increasing trend with respect to the second covariate. The measures of the model performance, such as $\mathscr{V}_u$, adjusted r^2, and the percentage of deviance explained were not as good as for the first simple model 1. The equiv-

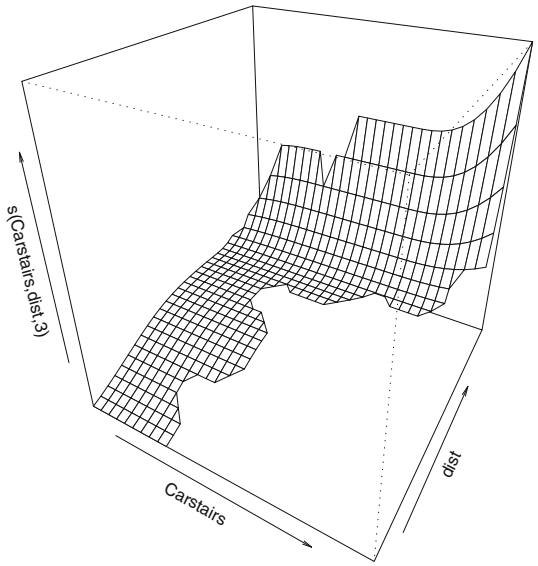

Fig. 7 Perspective plot of the estimated bivariate smooths of model 4 (incinerator data)

alent model without shape constraints resulted in the AIC = 157.35, whereas the AIC score for SCAM was 155.4. Hence, the AIC best selected model is the simple shape constrained model which only includes distance.

Air pollution data: The second application investigates the relationship between non-accidental daily mortality and air pollution. The data were from the National Morbidity, Mortality, and Air Pollution Study (Peng and Welty 2004) which contains 5,114 daily measurements on different variables for 108 cities within the United States. As an example a single city (Chicago) study was examined in Wood (2006a). The response variable was the daily number of deaths in Chicago (death) for the years 1987–1994. Four explanatory variables were considered: average daily temperature (tempd), levels of ozone (o3median), levels of particulate matter (pm10median), and time. Since it might be expected that increased mortality will be associated with increased concentrations of air pollution, modelling with SCAM may prove useful.

The preliminary modelling and examination of the data showed that the mortality rate at a given day could be better predicted if the aggregated air pollution levels and aggregated mean temperature were incorporated into the model, rather than levels of pollution and temperature on the day in question (Wood 2006a). It was proposed that the aggregation should be the sum of each covariate (except time), over the current day and three preceding days. Hence, the three aggregated predictors are as follows

$$\texttt{tmp}_i = \sum_{j=i-3}^{i} \texttt{tempd}_j, \quad \texttt{o3}_i = \sum_{j=i-3}^{i} \texttt{o3median}_j,$$

$$\texttt{pm10}_i = \sum_{j=i-3}^{i} \texttt{pm10median}_j.$$

Assuming that the observed numbers of daily death are independent Poisson random variables, the following additive model structure can be considered

$$\text{Model 1: } \log\{\mathrm{E}(\texttt{death}_i)\} = f_1(\texttt{time}_i) + m_2(\texttt{pm10}_i) + m_3(\texttt{o3}_i) + f_4(\texttt{tmp}_i),$$

where monotonically increasing constraints are assumed on m_2 and m_3, since increased air pollution levels are expected to be associated with increases in mortality. The plots for assessing the suitability of this model together and the plots of the smooth estimates are illustrated in Supplementary material, S.8. This model indicates that though the effect of the ozone level is only with one degree of freedom, it is positive and increasing. The rapid increase in the smooth of aggregated mean temperature can be explained by the four highest daily death rates occurring on four consecutive days of very high temperature, which also experienced high levels of ozone (Wood 2006a).

Since the combination of high temperatures together with high levels of ozone might be expected to result in higher mortality, we consider a bivariate smooth of these predictors. The following model is now considered

$$\text{Model 2: } \log\{\mathrm{E}(\texttt{death}_i)\} = f_1(\texttt{time}_i) + m_2(\texttt{pm10}_i) + m_3(\texttt{o3}_i, \texttt{tmp}_i),$$

where $m_2(\texttt{pm10}_i)$ is a monotone increasing function and $m_3(\texttt{o3}_i, \texttt{tmp}_i)$ is subject to single monotonicity along the first covariate. The diagnostic plots of this model showed a slight improvement in comparison to the first model (Supplementary material, S.8). The estimates of the univariate smooths and perspective plot of the estimated bivariate smooth of model 2 are illustrated in Fig. 8. The second model also has a lower $\mathcal{V}_u$ score which implies that model 2 is a preferable model.

The current approach has been applied to air pollution data for Chicago just for demonstration purpose. It would be of interest to apply the same model to other cities, to see whether the relationship between non-accidental mortality and air pollution can be described by the proposed SCAM in other locations.

6 Discussion

In this paper a framework for generalized additive modelling with a mixture of unconstrained and shape restricted smooth terms, SCAM, has been presented and evaluated on

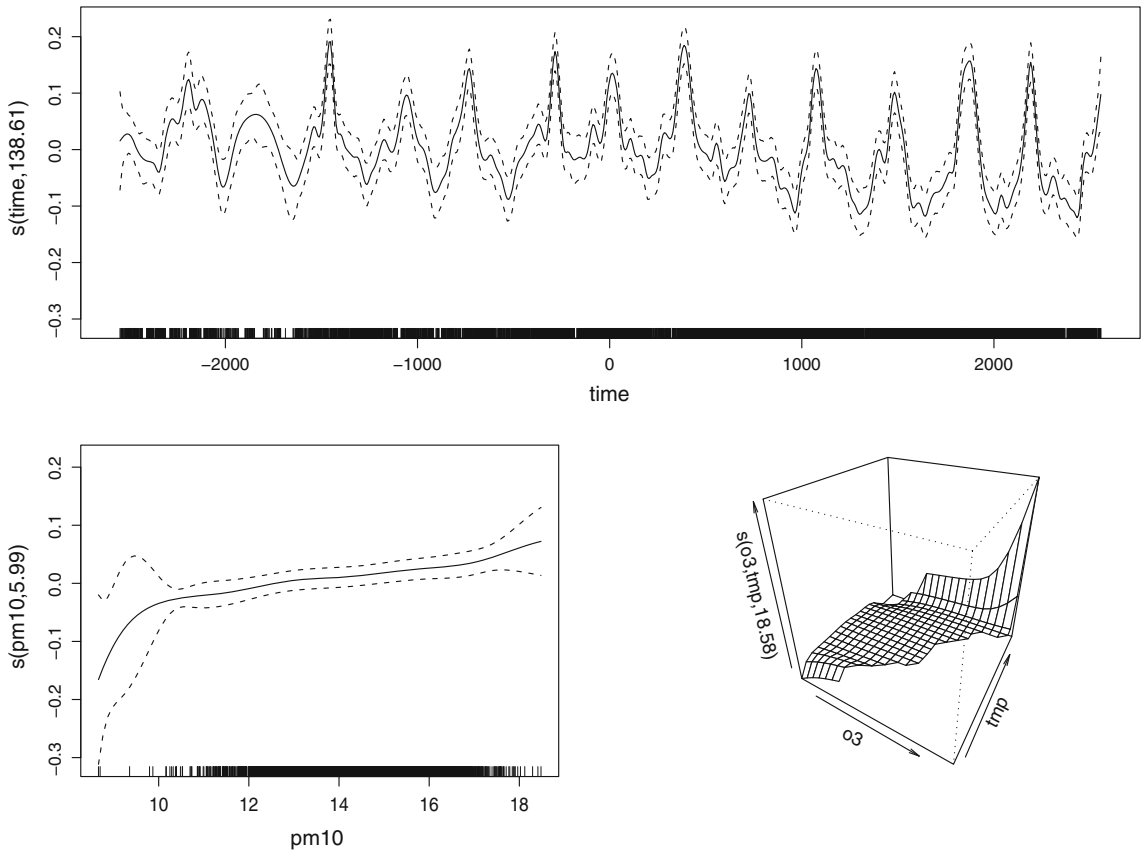

Fig. 8 The estimates of the smooth terms of model 2 (air pollution data). A cubic regression spline was used for f_1 with $q = 200$, SCOP-spline of the third order with $q = 10$ for m_2, and bivariate SCOP-spline with the marginal basis dimensions $q_1 = q_2 = 10$ for m_3

a range of simulated and real data sets. The motivation of this framework is an attempt to develop general methods for estimating SCAMs similar to that of a standard unconstrained GAM. SCAM models allow inclusion of multiple unconstrained and shape constrained smooths of both univariate and multi-dimensional type which are represented by the proposed SCOP-splines. It should be mentioned that the shape constraints were assured by the sufficient but not necessary condition for the cubic and higher order splines. However, this condition for the cubic splines is equivalent to that of Fritsch and Carlson (1980) who showed that the sufficient parameter space constitutes the substantial part of the necessary parameter space (see their Fig. 2, p. 242). Also the sensitivity analysis of Brezger and Steiner (2008) on an empirical application models defends the point that the sufficient condition is not highly restrictive.

Since a major challenge of any flexible regression method is its implementation in a computationally efficient and stable manner, numerically robust algorithms for model estimation have been presented. The main benefit of the procedure is that smoothing parameter selection is incorporated into the SCAM parameter estimation scheme, which also produces interval estimates at no additional cost. The approach has the $O(nq^2)$ computational cost of standard penalized regression spline based GAM estimation, but typically involves 2–4 times as many $O(nq^2)$ steps because of the additional non-linearities required for the monotonic terms, and the need to use Quasi-Newton in place of full Newton optimization. However, in contrast to the ad hoc methods of choosing the smoothing parameter used in other approaches, smoothing parameter selection for SCAMs is well founded. It should also be mentioned that although the simulation free intervals proposed in this paper show good coverage probabilities it might be of interest to see whether Bayesian confidence intervals derived from posterior distribution simulated via MCMC would give better results.

Acknowledgments The incinerator data were provided by the Small Area Health Statistics Unit, a unit jointly funded by the UK Department of Health, the Department of the Environment, Food and Rural Affairs, Environment Agency, Health and Safety Executive, Scottish Executive, National Assembly for Wales, and Northern Ireland Assembly. The authors are grateful to Jianxin Pan and Gavin Shaddick for useful discussions on several aspects of the work. The authors are also grateful for the valuable comments and suggestions of two referees and an associated editor. NP was funded by EPSRC/NERC grant EP/1000917/1.

Appendix

Appendix 1: Newton method for penalized likelihood estimation of SCAM

This appendix describes a full Newton (Newton–Raphson) method for maximizing the penalized likelihood of a SCAM. The penalized log likelihood function to be maximized w.r.t. $\boldsymbol{\beta}$ is

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}_\lambda\boldsymbol{\beta}/2,$$

where the log likelihood of $\boldsymbol{\beta}$ can be written as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\{y_i\theta_i - b_i(\theta_i)\}/a_i(\phi) + c_i(\phi, y_i)\right], \tag{6}$$

where a_i, b_i, and c_i are arbitrary functions, ϕ an arbitrary 'scale' parameter, and θ_i a 'canonical parameter' of the distribution related to the linear predictor via the relationship $\mathrm{E}(Y_i) = b'(\theta_i)$ (Wood 2006a). While the functions a_i, b_i, and c_i may vary with i, the scale parameter ϕ is assumed to be constant for all observations.

The distribution parameters θ_i depend on the model coefficients β_j via the link between the mean of Y_i and θ_i, $\mathrm{E}(Y_i) = b_i'(\theta_i)$. Recall that the smoothing parameter vector $\boldsymbol{\lambda}$ is considered to be fixed while estimating $\boldsymbol{\beta}$. Consider only cases where $a_i(\phi) = \phi/\omega_i$, and ω_i is a known constant, which usually equals 1. Almost all probability distributions of interest from the exponential family are covered by such a limitation. Then

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\omega_i \{y_i\theta_i - b_i(\theta_i)\}/\phi + c_i(\phi, y_i)\right]$$

and the first order derivative of $l(\boldsymbol{\beta})$ w.r.t. β_j is

$$\frac{\partial l_p}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} \omega_i \left\{ y_i \frac{\partial \theta_i}{\partial \beta_j} - b_i'(\theta_i)\frac{\partial \theta_i}{\partial \beta_j} \right\} - \mathbf{S}_{\lambda j}\boldsymbol{\beta},$$

where (for this appendix only) $\mathbf{S}_{\lambda j} = \sum_k \lambda_k \mathbf{S}_{kj}$ while $\mathbf{S}_{kj}$ is the jth row of the matrix $\mathbf{S}_k$, and

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}.$$

Taking the first order derivatives from the both sides of the linking equation $\mathrm{E}(Y_i) = b_i'(\theta_i)$, we get

$$\frac{\partial \mu_i}{\partial \theta_i} = b_i''(\theta_i) \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b_i''(\theta_i)},$$

$$\frac{\partial l_p}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} \frac{\{y_i - b_i'(\theta_i)\}}{b_i''(\theta_i)/\omega_i}\frac{\partial \mu_i}{\partial \beta_j} - \mathbf{S}_{\lambda j}\boldsymbol{\beta}. \tag{7}$$

Since $g(\mu_i) = \mathbf{X}_i\tilde{\boldsymbol{\beta}}$, then

$$g'(\mu_i)\frac{\partial \mu_i}{\partial \beta_j} = [\mathbf{X}]_{ij} \quad \text{if} \quad \tilde{\beta}_j = \beta_j,$$

$$g'(\mu_i)\frac{\partial \mu_i}{\partial \beta_j} = [\mathbf{X}]_{ij}\exp(\beta_j) \quad \text{otherwise.}$$

Hence

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{[\mathbf{X}]_{ij}}{g'(\mu_i)} \quad \text{if} \quad \tilde{\beta}_j = \beta_j,$$

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{[\mathbf{X}]_{ij}\exp(\beta_j)}{g'(\mu_i)} \quad \text{otherwise.}$$

Another key point of the exponential family concerns the variance

$$\mathrm{var}(Y_i) = b_i''(\theta_i)a_i(\phi) = b_i''(\theta_i)\phi/\omega_i,$$

which is represented in the theory of GLMs in terms of μ_i as $\mathrm{var}(Y_i) = V(\mu_i)\phi$, where $V(\mu_i) = b_i''(\theta_i)/\omega_i$.

Let $\mathbf{G}$ and $\mathbf{W}_1$ be $n \times n$ diagonal matrices with the diagonal elements $G_i = g'(\mu_i)$ and

$$w_{1i} = \frac{\omega_i}{V(\mu_i)g'^2(\mu_i)},$$

and let $\mathbf{C}$ be a $q \times q$ diagonal matrix such that

$$C_{jj} = \begin{cases} 1, & \text{if } \tilde{\beta}_j = \beta_j \\ \exp(\beta_j), & \text{otherwise.} \end{cases}$$

Then a penalized score vector may be written as

$$\mathbf{u}_p(\boldsymbol{\beta}) = \frac{\partial l_p}{\partial \boldsymbol{\beta}} = \frac{1}{\phi}\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}_1\mathbf{G}(\mathbf{y} - \boldsymbol{\mu}) - \mathbf{S}_\lambda\boldsymbol{\beta}. \tag{8}$$

To find the working model parameters estimates, $\hat{\boldsymbol{\beta}}$, one needs to solve $\mathbf{u}_p(\boldsymbol{\beta}) = \mathbf{0}$. These equations are non-linear and have no analytical solution, so some numerical methods should be applied. In the case of unconstrained GAM the penalized iteratively reweighed least squares (P-IRLS) scheme based on Fisher scoring is used to solve these equations.

To proceed the Hessian of the log-likelihood function is derived from (8)

$$\mathbf{H}(\boldsymbol{\beta}) = \left[\frac{\partial^2 l_p}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k}\right] = -\frac{1}{\phi}\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\mathbf{C} + \frac{1}{\phi}\mathbf{E} - \mathbf{S}_\lambda, \tag{9}$$

where $\mathbf{W}$ is a diagonal matrix with

$$w_i = \frac{\omega_i \alpha_i}{V(\mu_i) g'^2(\mu_i)}, \quad \text{and}$$
$$\alpha_i = 1 + (y_i - \mu_i)\left\{\frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)}\right\}, \qquad (10)$$

$\mathbf{E}$ is a $q \times q$ diagonal matrix with

$$E_{jj} = \begin{cases} 0, & \text{if } \tilde{\beta}_j = \beta_j \\ \sum_{i=1}^{n} w_i g'(\mu_i)[\mathbf{XC}]_{ij}(y_i - \mu_i)/\alpha(\mu_i), & \text{otherwise.} \end{cases}$$

Note that for the model with a canonical link function, the second term of α_i is equal to zero, since in this case

$$V'(\mu_i)/V(\mu_i) + g''(\mu_i)/g'(\mu_i) = 0.$$

Therefore, $\alpha_i = 1$ and the matrices $\mathbf{W}_1$ and $\mathbf{W}$ are identical.

So, using the Newton method, if $\boldsymbol{\beta}^{[k]}$ is the current estimate of $\boldsymbol{\beta}$, then the next estimate is

$$\boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} + \left\{\mathbf{C}^{[k]\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{[k]}\mathbf{X}\mathbf{C}^{[k]} - \mathbf{E}^{[k]} + \mathbf{S}_\lambda\right\}^{-1} \left\{\mathbf{C}^{[k]\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}_1^{[k]}\mathbf{G}^{[k]}(\mathbf{y} - \boldsymbol{\mu}^{[k]}) - \mathbf{S}_\lambda\boldsymbol{\beta}^{[k]}\right\}, \qquad (11)$$

where the scale parameter ϕ is absorbed into the smoothing parameter λ.

To use (11) directly for $\boldsymbol{\beta}$ estimation is not efficient since explicit formation of the Hessian would square the condition number of the working model matrix, $\sqrt{\mathbf{W}}\mathbf{X}$ (Golub and van Loan 1996). It should be noted that the Hessian matrix also appears in an expression for the edf of the fitted model (Appendix 2). In the case of the unconstrained model (Wood 2006a) a stable solution for $\hat{\boldsymbol{\beta}}$ is based on a QR decomposition of $\sqrt{\mathbf{W}}\mathbf{X}$ augmented with $\mathbf{B}$, where $\mathbf{B}^T\mathbf{B} = \mathbf{S}_\lambda$. The same approach can be applied here for the shape constrained model, i.e. use a QR decomposition of the augmented $\sqrt{\mathbf{W}}\mathbf{X}\mathbf{C}$. However, the values of $\mathbf{W}$ can be negative when a non-canonical link function is assumed, so firstly, the issue with these negative weights has to be handled.

The approach applied here is similar to that given in Sect. 3.3 of Wood (2011). Let $\tilde{\mathbf{W}}$ denote a diagonal matrix with the elements $|w_i|$, and $\mathbf{W}^-$ be a diagonal matrix with

$$w_i^- = \begin{cases} 0, & \text{if } w_i \geq 0 \\ -w_i, & \text{otherwise.} \end{cases}$$

Then

$$\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\mathbf{C} = \mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{W}}\mathbf{X}\mathbf{C} - 2\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-}\mathbf{X}\mathbf{C}.$$

Now the QR decomposition may be used for the augmented matrix,

$$\begin{bmatrix} \sqrt{\tilde{\mathbf{W}}}\mathbf{X}\mathbf{C} \\ \mathbf{B} \end{bmatrix} = \mathbf{QR}, \qquad (12)$$

where $\mathbf{Q}$ is a rectangular matrix with orthogonal columns, and $\mathbf{R}$ is upper triangular. Now let $\mathbf{Q}_1$ be the first n rows of $\mathbf{Q}$, then $\sqrt{\tilde{\mathbf{W}}}\mathbf{X}\mathbf{C} = \mathbf{Q}_1\mathbf{R}$.

Therefore

$$\begin{aligned}\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\mathbf{C} + \mathbf{S}_\lambda - \mathbf{E} &= \mathbf{R}^{\mathrm{T}}\mathbf{R} - 2\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-}\mathbf{X}\mathbf{C} - \mathbf{E} \\ &= \mathbf{R}^{\mathrm{T}}\left(\mathbf{I} - 2\mathbf{R}^{-\mathrm{T}}\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-}\mathbf{X}\mathbf{C}\mathbf{R}^{-1} - \mathbf{R}^{-\mathrm{T}}\mathbf{E}\mathbf{R}^{-1}\right)\mathbf{R} \\ &= \mathbf{R}^{\mathrm{T}}\left(\mathbf{I} - 2\mathbf{Q}_1^{\mathrm{T}}\mathbf{I}^{-}\mathbf{Q}_1 - \mathbf{R}^{-\mathrm{T}}\mathbf{E}\mathbf{R}^{-1}\right)\mathbf{R},\end{aligned}$$

where $\mathbf{I}^-$ is an $n \times n$ diagonal matrix with

$$I_i^- = \begin{cases} 0, \text{ if } & w_i \geq 0 \\ 1, \text{ otherwise.} \end{cases}$$

Note that several near non-identifiability issues can arise here. In order to deal with unidentifiable parameters it is proposed to use a singular value decomposition for the R factor of the QR decomposition if it is rank deficient. This step is described in Sect. 3.2.

The next step is to apply the eigen-decomposition

$$2\mathbf{Q}_1^{\mathrm{T}}\mathbf{I}^{-}\mathbf{Q}_1 + \mathbf{R}^{-\mathrm{T}}\mathbf{E}\mathbf{R}^{-1} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$$

which gives

$$\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\mathbf{C} + \mathbf{S}_\lambda - \mathbf{E} = \mathbf{R}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}\right)\mathbf{R} = \mathbf{R}^{\mathrm{T}}\mathbf{U}\left(\mathbf{I} - \boldsymbol{\Lambda}\right)\mathbf{U}^{\mathrm{T}}\mathbf{R}.$$

Defining a vector $\mathbf{z}$ such that $z_i = (y_i - \mu_i)g'(\mu_i)/\alpha(\mu_i)$ and $\tilde{\mathbf{z}}$ where $\tilde{z}_i = \mathrm{sign}(w_i)z_i$, then

$$\boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} + \mathbf{R}^{-1}\mathbf{U}(\mathbf{I} - \boldsymbol{\Lambda})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{Q}_1^{\mathrm{T}}\sqrt{\tilde{\mathbf{W}}}\tilde{\mathbf{z}} - \mathbf{R}^{-1}\mathbf{U}(\mathbf{I} - \boldsymbol{\Lambda})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{R}^{-\mathrm{T}}\mathbf{S}_\lambda\boldsymbol{\beta}^{[k]}. \qquad (13)$$

By denoting

$$\mathbf{P} = \mathbf{R}^{-1}\mathbf{U}(\mathbf{I} - \boldsymbol{\Lambda})^{-1/2} \quad \text{and} \quad \mathbf{K} = \mathbf{Q}_1\mathbf{U}(\mathbf{I} - \boldsymbol{\Lambda})^{-1/2} \qquad (14)$$

(13) may be written as

$$\boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} + \mathbf{P}\mathbf{K}^{\mathrm{T}}\sqrt{\tilde{\mathbf{W}}}\tilde{\mathbf{z}} - \mathbf{P}\mathbf{P}^{\mathrm{T}}\mathbf{S}_\lambda\boldsymbol{\beta}^{[k]}. \qquad (15)$$

The last expression has roughly the square root of the condition number of (11) for the unpenalized likelihood maximization problem, since the condition number of $\mathbf{R}^{-1}$ equals the condition number of $\sqrt{\tilde{\mathbf{W}}}\mathbf{X}\mathbf{C}$.

Another refinement may be required in the last step. If the Hessian of the log likelihood is indefinite then step in expression (14) will fail because some Λ_{ii} will exceed 1. To avoid indefiniteness problem a Fisher update step must be substituted by setting $\alpha(\mu_i) = 1$ so that $w_i \geq 0$ for any i, then the QR decomposition is used as previously

$$\begin{bmatrix} \sqrt{\mathbf{W}}\mathbf{X}\mathbf{C} \\ \mathbf{B} \end{bmatrix} = \mathbf{QR},$$

then $\boldsymbol{\beta}^{[k+1]} = \mathbf{R}^{-1}\mathbf{Q}_1^T\sqrt{\mathbf{W}}\mathbf{z}$, where $\mathbf{z} = \mathbf{G}(\mathbf{y}-\boldsymbol{\mu})+\mathbf{X}\mathbf{C}\boldsymbol{\beta}^{[k]}$. If there is an identifiability issue then the singular value decomposition step is applied on the QR factor, $\mathbf{R} = \mathscr{U}\mathbf{D}\mathbf{V}^T$, resulting in

$$\boldsymbol{\beta}^{[k+1]} = \mathbf{V}\mathbf{D}^{-}\mathbf{Q}_1^T\sqrt{\mathbf{W}}\mathbf{z},$$

where $\mathbf{Q}_1$ is the first n rows of $\mathbf{Q}\mathscr{U}$.

Note that in case of the canonical link function $\alpha_i = 1$ for any i, and therefore, $\tilde{\mathbf{W}} = \mathbf{W}$.

Appendix 2: SCAM degrees of freedom

An un-penalized model would have as many degrees of freedom as the number of unconstrained model parameters. However, the use of penalties decreases the number of degrees of freedom so that a model with $\lambda \to \infty$ would have the degrees of freedom near 1. Using the concept of the divergence of the maximum likelihood estimator, the edf of the penalized fit can be found as (Meyer and Woodroofe 2000; Wood 2001)

$$\tau = \mathrm{div}(\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n}\frac{\partial}{\partial y_i}\hat{\mu}_i(\mathbf{y}).$$

Substituting (11) (Appendix 1) into the model $g(\mu_i) = \mathbf{X}_i\tilde{\boldsymbol{\beta}}$ and taking first-order derivatives with respect to y_i, we get

$$\frac{\partial\hat{\mu}_i}{\partial y_i} = \left[\mathbf{X}\mathbf{C}\left\{\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\mathbf{C} - \mathbf{E} + \mathbf{S}_\lambda\right\}^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}_1\right]_{ii},$$

where the right-hand-side of this expression is the ith diagonal element of the matrix written in the square brackets.

Therefore,

$$\tau = \mathrm{tr}(\mathbf{F}), \tag{16}$$

where

$$\mathbf{F} = \left\{\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\mathbf{C} - \mathbf{E} + \mathbf{S}_\lambda\right\}^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}_1\mathbf{X}\mathbf{C}$$

and the matrices $\mathbf{W}$, $\mathbf{W}_1$, $\mathbf{C}$, and $\mathbf{E}$ are evaluated at convergence. Note that $\mathbf{F}$ is the expected Hessian of $l(\boldsymbol{\beta})$, premultiplied by the inverse of the Hessian of $l_p(\boldsymbol{\beta})$.

Using the approach and notations of Appendix 1, τ can also be obtained in a stable manner. Introducing $n \times n$ diagonal matrices $\mathbf{L}^+$ such that

$$L_{ii}^+ = \begin{cases}\alpha(\mu_i)^{-1}, & \text{if } w_i \geq 0\\ -\alpha(\mu_i)^{-1}, & \text{otherwise,}\end{cases}$$

then the expression for the edf (16) becomes

$$\begin{aligned}\mathrm{tr}(\mathbf{F}) &= \mathrm{tr}\left(\{\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\mathbf{C} - \mathbf{E} + \mathbf{S}_\lambda\}^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\sqrt{\tilde{\mathbf{W}}}\mathbf{L}^{+}\sqrt{\tilde{\mathbf{W}}}\mathbf{X}\mathbf{C}\right)\\ &= \mathrm{tr}\left(\mathbf{P}\mathbf{K}^{\mathrm{T}}\mathbf{L}^{+}\sqrt{\tilde{\mathbf{W}}}\mathbf{X}\mathbf{C}\right) = \mathrm{tr}\left(\mathbf{K}\mathbf{K}^{\mathrm{T}}\mathbf{L}^{+}\right).\end{aligned} \tag{17}$$

References

Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, B.F. (eds.) Second International Symposium on Information Theory. Academiai Kiado, Budapest (1973)

Bollaerts, K., Eilers, P., van Mechelen, I.: Simple and multiple P-splines regression with shape constraints. Br. J. Math. Stat. Psychol. **59**, 451–469 (2006)

Brezger, A., Steiner, W.: Monotonic regression based on Bayesian P-splines: an application to estimating price response functions from store-level scanner data. J. Bus. Econ. Stat. **26**(1), 90–104 (2008)

Claeskens, G., Krivobokova, T., Opsomer, J.: Asymptotic properties of penalized spline estimators. Biometrica **96**(3), 529–544 (2009)

Craven, P., Wahba, G.: Smoothing noisy data with spline functions. Numer. Math. **31**, 377–403 (1979)

De Boor, C.: A Practical Guide to Splines. Cambridge University Press, Cambridge (1978)

Dunson, D.: Bayesian semiparametric isotonic regression for count data. J. Am. Stat. Assoc. **100**(470), 618–627 (2005)

Dunson, D., Neelon, B.: Bayesian inference on order-constrained parameters in generalized linear models. Biometrics **59**, 286–295 (2003)

Eilers, P., Marx, B.: Flexible smoothing with B-splines and penalties. Stat. Sci. **11**, 89–121 (1996)

Elliott, P., Shaddick, G., Kleinschmidt, I., Jolley, D., Walls, P., Beresford, J., Grundy, C.: Cancer incidence near municipal solid waste incinerators in Great Britain. Br. J. Cancer **73**, 702–710 (1996)

Fritsch, F., Carlson, R.: Monotone piecewise cubic interpolation. SIAM J. Numer. Anal. **17**(2), 238–246 (1980)

Golub, G., van Loan, C.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)

Hastie, T., Tibshirani, R.: Generalized Additive Models. Chapman & Hall, New York (1990)

He, X., Shi, P.: Monotone B-spline smoothing. J. Am. Stat. Assoc. **93**(442), 643–650 (1998)

Holmes, C., Heard, N.: Generalized monotonic regression using random change points. Stat. Med. **22**, 623–638 (2003)

Kauermann, G., Krivobokova, T., Fahrmeir, L.: Some asymptotic results on generalized penalized spline smoothing. J. R. Stat. Soc. B **71**(2), 487–503 (2009)

Kauermann, G., Opsomer, J.: Data-driven selection of the spline dimension in penalized spline regression. Biometrika **98**(1), 225–230 (2011)

Kelly, C., Rice, J.: Monotone smoothing with application to dose-response curves and the assessment of synergism. Biometrics **46**, 1071–1085 (1990)

Kim, Y.-J., Gu, C.: Smoothing spline gaussian regression: more scalable computation via efficient approximation. J. R. Stat. Soc: Ser. B. **66**(2), 37–356 (2004)

Lang, S., Brezger, A.: Bayesian P-splines. J. Comput. Graph. Stat. **13**(1), 183–212 (2004)

Li, Y., Ruppert, D.: On the asymptotics of penalized splines. Biometrika **95**(2), 415–436 (2008)

Marra, G., Wood, S.N.: Coverage properties of confidence intervals for generalized additive model components. Scand. J. Stat. **39**(1), 53–74 (2012)

Meyer, M.: Inference using shape-restricted regression splines. Ann. Appl. Stat. **2**(3), 1013–1033 (2008)

Meyer, M.: Constrained penalized splines. Can. J. Stat. **40**(1), 190–206 (2012)
Meyer, M., Woodroofe, M.: On the degrees of freedom in shape-restricted regression. Ann. Stat. **28**(4), 1083–1104 (2000)
Nychka, D.: Bayesian confidence intervals for smoothing splines. J. Am. Stat. Assoc. **88**, 1134–1143 (1988)
Peng, R., Welty, L.: The NMMAPSdata package. R News **4**(2), 10–14 (2004)
Ramsay, J.: Monotone regression splines in action (with discussion). Stat. Sci. **3**(4), 425–461 (1988)
Rousson, V.: Monotone fitting for developmental variables. J. Appl. Stat. **35**(6), 659–670 (2008)
Ruppert, D.: Selecting the number of knots for penalized splines. J. Comput. Graph. Stat. **11**(4), 735–757 (2002)
Silverman, B.: Some aspects of the spline smoothing approach to non-parametric regression curve fitting. J. R. Stat. Soc.: Ser. B. **47**, 1–52 (1985)
Shaddick, G., Choo, L., Walker, S.: Modelling correlated count data with covariates. J. Stat. Comput. Simul. **77**(11), 945–954 (2007)
Villalobos, M., Wahba, G.: Inequality-constrained multivariate smoothing splines with application to the estimation of posterior probabilities. J. Am. Stat. Assoc. **82**(397), 239–248 (1987)
Wahba, G.: Bayesian confidence intervals for the cross validated smoothing spline. J. R. Stat. Soc: Ser. B. **45**, 133–150 (1983)
Wang, J., Meyer, M.: Testing the monotonicity or convexity of a function using regression splines. Can. J. Stat. **39**(1), 89–107 (2011)
Wood, S.: Monotonic smoothing splines fitted by cross validation. SIAM J. Sci. Comput. **15**(5), 1126–1133 (1994)
Wood, S.: Partially specified ecological models. Ecol. Monogr. **71**(1), 1–25 (2001)
Wood, S.: Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Am. Stat. Assoc. **99**, 673–686 (2004)
Wood, S.: Generalized Additive Models. An Introduction with R. Chapman & Hall, Boca Raton (2006a)
Wood, S.: On confidence intervals for generalized additive models based on penalized regression splines. Aust. N. Z. J. Stat. **48**(4), 445–464 (2006b)
Wood, S.: Fast stable direct fitting and smoothness selection for generalized additive models. J. R. Stat. Soc. B **70**(3), 495–518 (2008)
Wood, S.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. B **73**(1), 1–34 (2011)
Zhang, J.: A simple and efficient monotone smoother using smoothing splines. J. Nonparametr. Stat. **16**(5), 779–796 (2004)