Computational Statistics and Data Analysis 55 (2011) 2372–2387

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Practical variable selection for generalized additive models

Giampiero Marra[a,*], Simon N. Wood[b]

[a] Department of Statistical Science, University College London, London WC1E 6BT, UK
[b] Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK

ARTICLE INFO

Article history:
Received 18 March 2010
Received in revised form 2 February 2011
Accepted 2 February 2011
Available online 17 February 2011

Keywords:
Generalized additive model
Nonnegative garrote estimator
Penalized thin plate regression spline
Practical variable selection
Shrinkage smoother

ABSTRACT

The problem of variable selection within the class of generalized additive models, when there are many covariates to choose from but the number of predictors is still somewhat smaller than the number of observations, is considered. Two very simple but effective shrinkage methods and an extension of the nonnegative garrote estimator are introduced. The proposals avoid having to use nonparametric testing methods for which there is no general reliable distributional theory. Moreover, component selection is carried out in one single step as opposed to many selection procedures which involve an exhaustive search of all possible models. The empirical performance of the proposed methods is compared to that of some available techniques via an extensive simulation study. The results show under which conditions one method can be preferred over another, hence providing applied researchers with some practical guidelines. The procedures are also illustrated analysing data on plasma beta-carotene levels from a cross-sectional study conducted in the United States.

1. Introduction

A generalized additive model (GAM; Hastie and Tibshirani, 1990; Wood, 2006) can be thought of as a generalized linear model (GLM; McCullagh and Nelder, 1989) with a linear predictor involving smooth functions of covariates

$$g\{\mathbb{E}(Y_i)\} = \mathbf{X}_i^*\boldsymbol{\theta}^* + \sum_j f_j(x_{ji}), \tag{1}$$

where $g(\cdot)$ is a specified link function, the response Y_i follows an exponential family distribution, $\mathbf{X}_i^*$ is the ith row of $\mathbf{X}^*$ and strictly contains parametric (linear) model components, with corresponding parameter vector $\boldsymbol{\theta}^*$, and the f_j are smooth functions of the covariates x_j, which may be vector covariates, subject to identifiability constraints such as $\sum_i f_j(x_{ji}) = 0\ \forall j$. The f_j are represented via regression spline bases, with associated measures of function roughness which can be expressed as quadratic forms in the basis coefficients (e.g. Wood, 2006). Given such bases, model (1) can be estimated as a GLM, but to avoid overfitting it is necessary to estimate such a model by penalized maximum likelihood estimation, in which roughness measures are used to control overfit. In practice, the penalized likelihood is maximized by penalized iteratively reweighted least squares (P-IRLS), where the GAM is fitted by iterative minimization of the problem

$$\|\sqrt{\mathbf{W}^{[k]}}(\mathbf{z}^{[k]} - \mathbf{X}\boldsymbol{\beta})\|^2 + \sum_j \lambda_j \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_j\boldsymbol{\beta} \quad \text{w.r.t. } \boldsymbol{\beta}.$$

[*] Corresponding author. Tel.: +44 0 20 7679 1864; fax: +44 0 20 3108 3105.
E-mail address: giampiero@stats.ucl.ac.uk (G. Marra).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.02.004

k is the iteration index, $\mathbf{z}^{[k]} = \mathbf{X}\boldsymbol{\beta}^{[k]} + \mathbf{G}^{[k]}(\mathbf{y} - \boldsymbol{\mu}^{[k]})$, $\mu_i^{[k]}$ is the current model estimate of $\mathbb{E}(Y_i)$, $\mathbf{G}^{[k]}$ is a diagonal matrix such that $G_{ii}^{[k]} = g'(\mu_i^{[k]})$, $\mathbf{W}^{[k]}$ is a diagonal matrix given by $W_{ii}^{[k]} = [G_{ii}^{[k]2} V(\mu_i^{[k]})]^{-1}$ where $V(\mu_i^{[k]})$ gives the variance of Y_i to within a response distribution scale parameter, ϕ, $\mathbf{X}$ includes the columns of $\mathbf{X}^*$ and columns representing the spline bases for the f_j, while $\boldsymbol{\beta}$ contains $\boldsymbol{\theta}^*$ and all the smooth coefficient vectors, $\boldsymbol{\beta}_j$. The $\mathbf{S}_j$ are matrices of known coefficients such that the terms in the summation measure the roughness of the smooth functions. The λ_j are smoothing parameters that control the trade-off between fit and smoothness, and can be selected by minimization of the generalized cross validation (GCV) score, if a scale parameter has to be estimated, the generalized Akaike's information criterion (AIC), and restricted maximum likelihood (REML) estimation, to name a few. The computational methods of Wood (2006, 2008, 2011) are available to estimate the λ_j using the criteria mentioned above.

Variable selection is an important area of research. From a pragmatic point of view, it aims at determining which covariates have the strongest effects on the response of interest, whereas from a statistical perspective it represents a means to achieve a balance between goodness of fit and parsimony. In other words, by effectively identifying a subset of important covariates, variable selection can both enhance model interpretability and improve prediction accuracy. Methods such as subset selection, stepwise procedures and shrinkage methods can be employed (see Guisan et al. (2002) for an overview). Subset selection chooses a model containing a subset of predictors according to some criterion, but all possible subset models have to be explored and hence it can become computationally expensive as the number of predictors increases. Stepwise procedures do not make use of all possible models, therefore reducing computational cost, but they might be inconsistent given the dependence on the path chosen through the variable space. The additional drawback of these procedures is that if we perform variable selection and then hypothesis testing using the selected model, the p-values associated with the model terms will not be strictly correct since they neglect variable selection uncertainty. Shrinkage methods are becoming popular in the statistical literature. In fact, they have proved to be a valid alternative to the procedures above in terms of stability and prediction. Moreover, shrinkage procedures are continuous processes since variable selection is carried out in one single step as opposed to subset selection and stepwise algorithms (Hesterberg et al., 2008).

For the additive model case, subset selection and stepwise procedures can be carried out using, for instance, the Akaike Information Criteria (e.g. Greven and Kneib, 2009; Wager et al., 2007). A number of hypothesis testing approaches have also been proposed, which do model selection in terms of either choosing between linear and more general smooth term alternatives or dropping unimportant components from the model (Cantoni and Hastie, 2002; Hastie and Tibshirani, 1990; Kauermann et al., 2009; Kauermann and Tutz, 2001; Scheipl et al., 2008; Wood, 2006). Despite the fact that some testing methods have been introduced in the GAM context (Hastie and Tibshirani, 1990; Wood, 2006), a *general* reliable distributional theory for the smooth terms of a GAM has not been developed to date. Shrinkage methods for linear models and GLMs, which simultaneously address estimation and variable selection, have been proposed (e.g. Breiman, 1995; Daye and Jeng, 2009; Efron et al., 2004; Nott and Leng, 2010; Similä and Tikka, 2007; Tibshirani, 1996; Tutz and Binder, 2007; Yuan and Lin, 2006; Zou, 2006). Some algorithms have also been introduced to achieve component selection within additive models (Avalos et al., 2007; Belitz and Lang, 2008; Bühlmann and Yu, 2003; Cantoni et al., 2011; Lin and Zhang, 2006; Xue, 2009) and GAMs (see Zhang and Lin (2006) and references therein). However, for the GAM case, the boosting technique of Tutz and Binder (2006) and a generalization of the approach of Belitz and Lang (2008) seem to be the only fitting procedures available for public use.

In this paper, we focus on smooth component selection when dealing with GAMs by pursuing a shrinkage approach. As mentioned earlier on, this approach is appealing since it has the properties of stability and prediction, and variable selection can be carried out in one single step. Furthermore, it avoids having to use testing methods for which there is no general distributional theory. We propose two effective shrinkage methods and extend the nonnegative garrote estimator to achieve component selection within GAMs. Their empirical performance is compared to that of some available methods via an extensive simulation study. The implementation is in R (R Development Core Team, 2010). The procedures are also illustrated by analysing data on plasma beta-carotene levels from a cross-sectional study conducted in the United States. Note that we concentrate throughout on the case in which we need to select from a small to moderate sized set of potential predictors. In part this is due to method constraints. However, we also believe that in practice it is not very common that the modeller *does not* know which of a very large number of predictors is important, but *does* know that an additive structure gives an appropriate model.

2. Methods

Smoothing parameter estimation can select between models of different complexity, but it does not usually remove a smooth term from the model altogether. This is because the usual penalty of a spline basis does not allow for the shrinkage of the functions that are in the penalty null space (and for the most useful smoothers the null space has a dimension greater than zero). The proposals in Sections 2.1 and 2.2 are based on the idea that the space of a spline basis can be decomposed in the sum of two components, one associated with the functions in the penalty null space and the other with the penalty range space. The smoothing penalty shrinks functions in the range space (to zero if the smoothing penalty is high enough), but leaves the function component in the null space untouched. So to have the possibility of shrinking the whole spline term to zero, it is necessary to penalize the null space. As an alternative approach, the method introduced in Section 2.4 does not require the use of such a decomposition, and is based on the idea of shrinking the smooth function estimates obtained from a standard fitted GAM. The proposed methods have the properties of subset selection, but with the advantage that variable selection can be achieved in one single step. Section 2.5 presents some of the available alternatives, whereas Section 2.6 briefly discusses multiple smoothing parameter estimation which is crucial for the variable selection methods to work well.

2.1. Double penalty approach

The generic smoothing penalty matrix $\mathbf{S}_j$ associated with a smooth term of a GAM can be decomposed as

$$\mathbf{U}_j \Lambda_j \mathbf{U}_j^{\mathsf{T}}, \tag{2}$$

where $\mathbf{U}_j$ is an eigenvector matrix associated with the jth smooth function, and Λ_j the corresponding diagonal eigenvalue matrix. The fact that a part of the spline basis space deals with the penalty null space implies that Λ_j contains zero eigenvalues. This may be problematic if variable selection has to be carried out. For instance, let us assume that the jth smooth component is a nuisance function, and that we use a penalty matrix as defined above during the model fitting process. Even if λ_j goes to infinity there will not be any guarantee that the smooth term will be suppressed completely (i.e. estimated as zero).

In order to circumvent this difficulty, we can produce an extra penalty which penalizes only functions in the null space of the penalty, so that a smooth component can be completely removed. Specifically, let us consider decomposition (2). An extra penalty can be formed as follows

$$\mathbf{S}_j^* = \mathbf{U}_j^* \mathbf{U}_j^{*\top},$$

where $\mathbf{U}_j^*$ is the matrix of eigenvectors corresponding to the zero eigenvalues of Λ_j. So a GAM can be fitted subjecting each component function to a double penalty of the form

$$\lambda_j \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_j \boldsymbol{\beta} + \lambda_j^* \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_j^* \boldsymbol{\beta}, \tag{3}$$

where both λ_j and λ_j^* will now have to be estimated. By introducing a penalty for the null space, smoothing parameter estimation (that is part of GAM fitting) can completely remove terms from the model.

To reiterate the basic idea, any spline type smoother can be decomposed into two component functions: a component in the null space of the penalty, and a component in the range space of the penalty. The first term in (3) penalizes only function components in the range space, but can shrink these to zero, while the second term in (3) penalizes only function components in the null space, but can shrink these too to zero. For example, in the case of the usual cubic spline penalty, the second term in (3) would penalize straight line components to zero, while the first term would penalize (towards zero) function components representing departure from straight line behaviour.

This approach can be employed by setting the argument `select=TRUE` in the `gam` function of the `R` package `mgcv`.

2.2. Shrinkage approach

As an alternative approach which avoids doubling the number of smoothing parameters to estimate, we can replace the smoothing penalty matrix $\mathbf{S}_j$ with

$$\tilde{\mathbf{S}}_j = \mathbf{U}_j \tilde{\Lambda}_j \mathbf{U}_j^{\mathsf{T}},$$

where $\tilde{\Lambda}_j$ is the same as Λ_j except for the zero eigenvalues which are set to ϵ, a small proportion of the smallest strictly positive eigenvalues of $\mathbf{S}_j$. This is exactly equivalent to fixing $\lambda_j^* = \epsilon\lambda_j$ in (3) and forces the eigenvalues of $\tilde{\mathbf{S}}_j$ associated with the penalty null space to be different from zero: hence smoothing parameter selection can remove a smooth component from the model altogether. We choose ϵ to be a small proportion of the smallest strictly positive eigenvalues of $\mathbf{S}_j$ to ensure that $\boldsymbol{\beta}^{\mathsf{T}}\tilde{\mathbf{S}}_j\boldsymbol{\beta} \approx \boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_j\boldsymbol{\beta}$ for all the regression spline coefficients except those in or "close to" the null space of $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_j\boldsymbol{\beta}$. A value for ϵ equal to 1/10 yields good results in terms of goodness of fit and shrinkage (see Section 3).

This approach can be employed specifying the GAM formula of `mgcv` as a function of shrinkage smoothers. Two classes are implemented: `cs` and `ts`, based on cubic and thin plate regression spline smoothers, respectively.

2.3. Shrinkage penalty interpretation

Estimation by penalized likelihood with GCV or REML type smoothing parameter estimation can be viewed as an empirical Bayes procedure, with the penalties corresponding to (usually improper) Gaussian priors on the spline coefficients (the basic idea goes back to Kimeldorf and Wahba (1970)). In this case the $\mathbf{S}_j$ are viewed as prior precision matrices. It is the lack of full rank in the $\mathbf{S}_j$ that makes the prior improper, and this impropriety is a consequence of having a null space of functions that are treated as 'completely smooth' according to the penalty.

The proposals in Sections 2.1 and 2.2 both remove the impropriety from the prior, since both $\lambda_j\mathbf{S}_j + \lambda_j^*\mathbf{S}_j^*$ and $\tilde{\mathbf{S}}_j$ are full rank. The double penalty approach of Section 2.1 makes no prior assumption about the how much to penalize the null space relative to the range space for a term, but allows the smoothing parameter estimation to determine this from the data.

On the other hand, the single penalty Section 2.2 approach assumes that the null space should be penalized less than the range space. This is a natural approach to take in some cases. In a cubic spline case for example, this would mean that as the smoothing parameter increases we would first penalize towards a straight line, and then shrink the line towards zero. However there is an inevitable arbitrariness about exactly how to weight the penalization of the two components, and if the data suggest penalizing the null space more heavily than the range space there is often no compelling reason for not doing so (perhaps the data contain no overall trend, for example).

In the work reported here we have employed the simplest null space penalties that will remove all impropriety from the priors and hence allow terms to be completely removed from the model. We have not considered whether some null space components should be penalized more than others. If the null space itself allows moderately complicated functions (e.g. the null space of a multi-dimensional thin plate spline penalty based on moderately high derivatives) then the modeller might want to impose some hierarchy within the null space basis coefficients penalizing some more than others. However it seems likely that the improvements achievable by doing this would be rather modest, and we will not pursue this further here. In any case, for one-dimensional smooths using a cubic spline penalty, the null space is only one-dimensional after the imposition of identifiability constraints on the GAM components, so the issue does not arise.

Finally, it is worth noting that after the imposition of standard identifiability constraints some smoothers have a zero-dimensional null space corresponding to a *proper* prior for the coefficients, and can therefore be selected out of the model without the methods of Sections 2.1 and 2.2. The obvious example is a spline, $f(x)$, based on the penalty functional $\int f'(x)^2 \mathrm{d}x$. The null space of this penalty is the space of constant functions, which is eliminated from the space of the estimates by the identifiability constraint on f. Hence within the space of the identifiability constraint the penalty has full rank, and the corresponding prior of f is proper. Such terms can be used in R package mgcv, and priors of this sort have been used in a fully Bayesian context also (e.g. Chib and Greenberg (2007) use a random walk prior which shrinks towards the constant functions, and to zero after constraint). The difficulty with such terms is that the required low order penalization typically results in poor mean square error performance, in part because of undesirable properties such as tending to a constant function at the boundaries of the data.

2.4. *Nonnegative garrote component selection*

In order to identify the important smooth components of an additive model, Cantoni et al. (2011) and Yuan (2007) suggest employing the nonnegative garrote estimator, first proposed by Breiman (1995) in the linear model context, which has the properties of shrinkage and stability. The idea behind this is as follows. In a first step we obtain the original regression coefficient or smooth function estimates, depending on whether we are in a parametric or nonparametric context. We then shrink the model components by solving a constrained optimization problem.

The method presented here generalizes the nonnegative garrote estimator for additive models proposed by the authors above to the GAM context. First, we obtain some initial estimates for the smooth components of a model, $[\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \ldots]$. Second, we solve the problem

$$\text{minimize } D(\boldsymbol{\eta}) \text{ w.r.t. } \mathbf{d} \quad \text{subject to } \mathbf{d} \geq \mathbf{0} \text{ and } \mathbf{1}^{\mathsf{T}}\mathbf{d} = \gamma, \tag{4}$$

where $\boldsymbol{\eta} = \hat{\mathbf{F}}\mathbf{d}$, and $\hat{\mathbf{F}} = [\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \ldots]$. The parameter vector $\mathbf{d}$ contains the shrinking coefficients, and γ is a tuning parameter. D is the usual model deviance defined as $2\phi\,\{l_{\text{sat}} - l(\boldsymbol{\eta})\}$, where $l(\boldsymbol{\eta})$ is the log-likelihood of the model with linear predictor $\boldsymbol{\eta}$ and l_{sat} the maximum value for the log-likelihood of the model with one parameter per datum.

For a given $\hat{\mathbf{F}}$ and γ, the estimated shrinking coefficients allow us to do variable selection. That is, if $\hat{d}_j = 0$ then the jth component is viewed as uninformative and hence removed from the model. The shrinking coefficients also give information about the importance of each component in the model since some terms can be shrunk by some proportion $\hat{d}_j$, left unchanged (if $\hat{d}_j = 1$) or magnified (if $\hat{d}_j > 1$). The jth final smooth component estimate is given by $\hat{\mathbf{f}}_j^{*} = \hat{\mathbf{f}}_j\hat{\mathbf{d}}$.

A small value for γ shrinks the d_j to zero and vice versa, hence affecting the final estimates. In fact this parameter has to be selected with a certain degree of accuracy. As suggested by Cantoni et al. (2011) and Yuan (2007), a 5-fold cross validation gives satisfactory results in terms of achieving a good balance between bias and variance, and it can be implemented using the following practical algorithm:

1. Split the data into subsets denoted by $I_1, \ldots, I_b, \ldots, I_B$, where b represents the subset considered and B the maximum number of subsets used for cross validation. In this case, $B = 5$.
2. Choose an equally spaced grid of values for γ in the interval $[0, n_c]$, where n_c indicates the total number of covariates used to fit the model.
3. For each value γ in the interval $[0, n_c]$.
   (a) For each value of b.
      i. Fit a standard GAM (employing mgcv or any other available smoothing package) using the sample containing all the observations except those in I_b. Then store the resulting smooth function estimates in $\hat{\mathbf{F}}^{[-I_b]}$.
      ii. Using the subset of observations as in *i*, solve (4) via iterative minimization of the problem

$$\|\sqrt{\mathbf{W}^{[k]}}(\mathbf{z}^{[k]} - \boldsymbol{\eta})\|^2 \quad \text{subject to } \mathbf{d} \geq \mathbf{0} \text{ and } \mathbf{1}^{\mathsf{T}}\mathbf{d} = \gamma,$$

where k is the iteration index, $z_i^{[k]} = \eta_i^{[k]} + g'(\mu_i^{[k]})(y_i - \mu_i^{[k]})$, $\boldsymbol{\eta}^{[k]} = \hat{\mathbf{F}}^{[-I_b]}\mathbf{d}^{[k]}$, $\mu_i^{[k]} = g^{-1}(\eta^{[k]})$ and $W_{ii}^{[k]} = g'(\mu_i^{[k]})^{-2}V(\mu_i^{[k]})^{-1}$. In practice, this can be achieved by replacing, in the inner loop of `glm.fit` in R, the function `fit` with the function `pcls` for quadratic programming available in `mgcv`.

iii. For each observation i in I_b, obtain $D_i(\hat{\eta}_i^{[-I_b]})$ where $\hat{\boldsymbol{\eta}}^{[-I_b]} = \hat{\mathbf{F}}\hat{\mathbf{d}}$, with parameter vector $\hat{\mathbf{d}}$ obtained in the previous two steps, and D_i is the contribution to the "full data" deviance that is associated with the ith datum.

(b) Calculate the cross validation predictive deviance

$$V_B(\gamma) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n_b}\sum_{i\in I_b} D_i(\hat{\eta}_i^{[-I_b]}),$$

where n_b represents the sample size for the subset I_b.

4. Obtain final smooth component estimates by repeating steps i. and ii. but using the whole sample, with value for γ selected to minimize $V_B(\gamma)$.

2.5. Some available alternatives

For the sake of comparison in our simulation study, we briefly describe some of the alternative available methods for carrying out component selection in a regression spline context.

2.5.1. Backward selection

A classic backward selection procedure for variable selection within GAMs can be employed. In order to implement the procedure we need to use some p-value definition. Here, we follow the approach of Wood (2006). In the large sample limit

$$\hat{\boldsymbol{\beta}} \backsim N(\mathbb{E}(\hat{\boldsymbol{\beta}}), \mathbf{V}_{\hat{\boldsymbol{\beta}}}),$$

where $\hat{\boldsymbol{\beta}}$ is the maximum penalized likelihood estimate of $\boldsymbol{\beta}$, which is of the form $(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{z}$, and $\mathbf{S} = \sum_j \lambda_j \mathbf{S}_j$. $\mathbb{E}(\hat{\boldsymbol{\beta}}) \neq \boldsymbol{\beta}$ because of penalty-induced bias, and the most frequentist covariance matrix is given by

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\phi.$$

The dispersion parameter ϕ can be estimated by the Pearson estimator $\hat{\phi} = \|\sqrt{\mathbf{W}}(\mathbf{y} - \mathbf{A}\mathbf{y})\|^2/\{n - \mathrm{tr}(\mathbf{A})\}$, with hat matrix $\mathbf{A} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}$ and number of observations n. The trace of $\mathbf{A}$ represents the estimated degrees of freedom (edf) of the fitted model. Given that $\mathbb{E}(\hat{\boldsymbol{\beta}}_j) \neq \boldsymbol{\beta}_j$, the usual distributional results for GLMs cannot be employed for hypothesis testing. However, when the goal of the analysis is testing that a smooth term of a GAM is equal to zero, we have that if $\boldsymbol{\beta}_j = \mathbf{0}$ then $\mathbb{E}(\hat{\boldsymbol{\beta}}_j) \approx \mathbf{0}$ (Wood, 2006). It follows that, under the null hypothesis that the coefficients of a smooth function are equal to zero,

$$\hat{\boldsymbol{\beta}}_j^{\mathsf{T}}\mathbf{V}_{\hat{\boldsymbol{\beta}}_j}^{r-}\hat{\boldsymbol{\beta}}_j \dot{\backsim} \chi_r^2,$$

where r denotes the rank of the covariance matrix of $\hat{\boldsymbol{\beta}}_j$, and $\mathbf{V}_{\hat{\boldsymbol{\beta}}_j}^{r-}$ is the rank r generalized pseudoinverse of $\mathbf{V}_{\hat{\boldsymbol{\beta}}_j}$ that has to be employed to overcome possible matrix rank deficiencies deriving from the fact that the smoothing penalty may suppress some dimensions of the parameter space. r is determined heuristically as follows. It is the minimum value between the maximum edf value allowed for the jth smooth term (which is also the number of basis functions used for the term) and the smallest integer not less than the quantity calculated as $2 * \mathrm{edf}_j$. If ϕ is unknown, then the null hypothesis can be tested using the following result

$$\frac{\hat{\boldsymbol{\beta}}_j^{\mathsf{T}}\mathbf{V}_{\hat{\boldsymbol{\beta}}_j}^{r-}\hat{\boldsymbol{\beta}}_j/r}{\hat{\phi}/\phi} = \hat{\boldsymbol{\beta}}_j^{\mathsf{T}}\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_j}^{r-}\hat{\boldsymbol{\beta}}_j/r \dot{\backsim} F_{r,n-\mathrm{edf}},$$

since $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_j}$ is based on $\hat{\phi}$.

A backward selection procedure using the p-value definition discussed in this section can be implemented by extracting the p-values for the smooth components of a GAM from the function `summary.gam` in `mgcv`.

2.5.2. GAM boosting

Binder and Tutz (2008) found that when a subset of a large number of predictors has to be selected and the degree of smoothness for the smooth components has to be chosen, generalized additive modelling by likelihood based boosting can achieve these two goals simultaneously (Tutz and Binder, 2006). They also provide simulation evidence that GAM boosting can be much better than alternative methods in very data poor settings, with many spurious covariates. This procedure iteratively fits a GAM by applying a 'weak learner' on the residuals of smooth components. The number of boosting steps is determined by a stopping rule such as cross-validation or an information criterion.

The R package `GAMBoost` can be used for fitting GAMs by likelihood based boosting, using 2nd degree B-splines with 1st order difference penalty as the default settings suggest. The function `optimGAMBoostPenalty` can be employed to select the optimal number of boosting steps.

2.5.3. Modified backfitting

Belitz and Lang (2008) developed an elegantly simple method for simultaneously estimating a model and selecting which components to include, based on a modification of backfitting, with computationally efficient sparse smoothers. As with backfitting, smooths are estimated by iteratively smoothing partial residuals, but at each step, rather than using a single fixed degree of freedom smoother (as in classical backfitting), Belitz and Lang compute a number of alternative smoothers, corresponding to different degrees of freedom *plus* the null function corresponding to dropping the term altogether. To choose between these alternatives they compare a whole model GCV or AIC score for each alternative (using the current best estimate for the rest of the linear predictor). The method gains efficiency by using sparse smoothers (B-splines + discrete penalties), and by using an additive approximation for the effective degrees of freedom for the whole model, required by the GCV or AIC score. The latter approximation is perhaps the method's main potential weakness: the approximation will deteriorate as covariate correlation increases, which has the potential to cause method performance to suffer.

Belitz and Lang (2008) only present the method in the additive context, but as they point out the extension to *generalized* additive models is straightforward, and is available in BayesX (www.statistik.lmu.de/~bayesx/), the command line version of which can be called from within R.

2.5.4. Parsimonious additive models

This approach, introduced by Avalos et al. (2007) for additive models, consists of separating the parametric and nonparametric parts of the smooth functions, and then fitting the parametric bit using a LASSO regression (Tibshirani, 1996) and the nonparametric part by solving a penalized least squares problem. A modified version of this approach can be implemented as follows:

1. Using thin plate regression splines (Wood, 2003, 2006), set up a matrix $\mathbf{X}^*$ containing the terms of the smooth functions which deal with the penalty null space.
2. Store the coefficients ($\hat{\boldsymbol{\alpha}}_{\text{ols}}$) obtained by fitting a linear regression of $\mathbf{y}$ on $\mathbf{X}^*$, where $\mathbf{y}$ represents the response vector.
3. By using the library `lars`, compute the lasso coefficients by minimization of the problem

$$\|\mathbf{y}-\mathbf{X}^*\boldsymbol{\alpha}\|^2+\theta\sum_j^p|\alpha_j| \quad \text{w.r.t. } \boldsymbol{\alpha}.$$

   The tuning parameter θ is selected by K-fold cross validation using the function `cv.lars` in `lars`. Default settings suggest setting $K = 10$.
4. Compute the adjusted variable $\mathbf{y}^* = \mathbf{y} - \mathbf{X}^*\hat{\boldsymbol{\alpha}}_{\text{ols}}$, in order to ensure orthogonality between linear and nonlinear fits.
5. Set up a matrix $\mathbf{X}^+$ containing the terms of the smooth functions that deal with the penalty range space. Then, by using for instance `mgcv`, solve the penalized least squares problem

$$\|\mathbf{y}^*-\mathbf{X}^+\boldsymbol{\beta}_+\|^2+\sum_j\lambda_j\boldsymbol{\beta}_+^{\mathsf{T}}\mathbf{S}_j^+\boldsymbol{\beta}_+ \quad \text{w.r.t. } \boldsymbol{\beta}_+,$$

   where the $\mathbf{S}_j^+$ are the smoothing penalty matrices associated with the penalty range space.
6. Combine the results obtained in steps 3 and 5 and work out the final smooth function estimates.

This approach may look appealing since the LASSO regression can yield, via the use of a l_1 penalty, $\hat{\boldsymbol{\alpha}} = \mathbf{0}$. This means that, provided the smoothing parameters associated with the nuisance functions go to infinity, such a procedure can produce parsimonious additive models. However, as discussed in Hesterberg et al. (2008), the main drawback is that a linear term may be shrunk to zero while keeping the corresponding higher order components.

2.6. Smoothness selection

In order to implement the methods discussed in the previous sections, some multiple smoothing parameter selection procedure is needed. Importantly, for the methods to perform well it is crucial to use some stable and reliable computational method.

Multiple smoothing parameter estimation can be achieved by minimization of a prediction error estimate, such as the generalized cross validation (GCV) score, if a dispersion parameter has to be estimated, or the generalized Akaike's information criterion (AIC). Following Wood (2008), smoothing parameter selection via the GCV score consists of minimizing

$$V_g(\boldsymbol{\lambda})=\frac{nD(\hat{\boldsymbol{\beta}})}{\{n-\text{tr}(\mathbf{A})\}^2} \quad \text{w.r.t. } \boldsymbol{\lambda},$$

where tr($\mathbf{A}$) is defined in Section 2.5.1. $D(\hat{\boldsymbol{\beta}})$, the model deviance, is defined as $2\phi(\hat{l}_{\text{sat}}-\hat{l})$, $\hat{l}$ is the log-likelihood of the fitted model and $\hat{l}_{\text{sat}}$ the maximum value for the log-likelihood of the model with one parameter per datum. In case ϕ is known, the following generalized AIC is minimized instead

$$V_a(\boldsymbol{\lambda})=D(\hat{\boldsymbol{\beta}})+2\text{tr}(\mathbf{A})\phi.$$

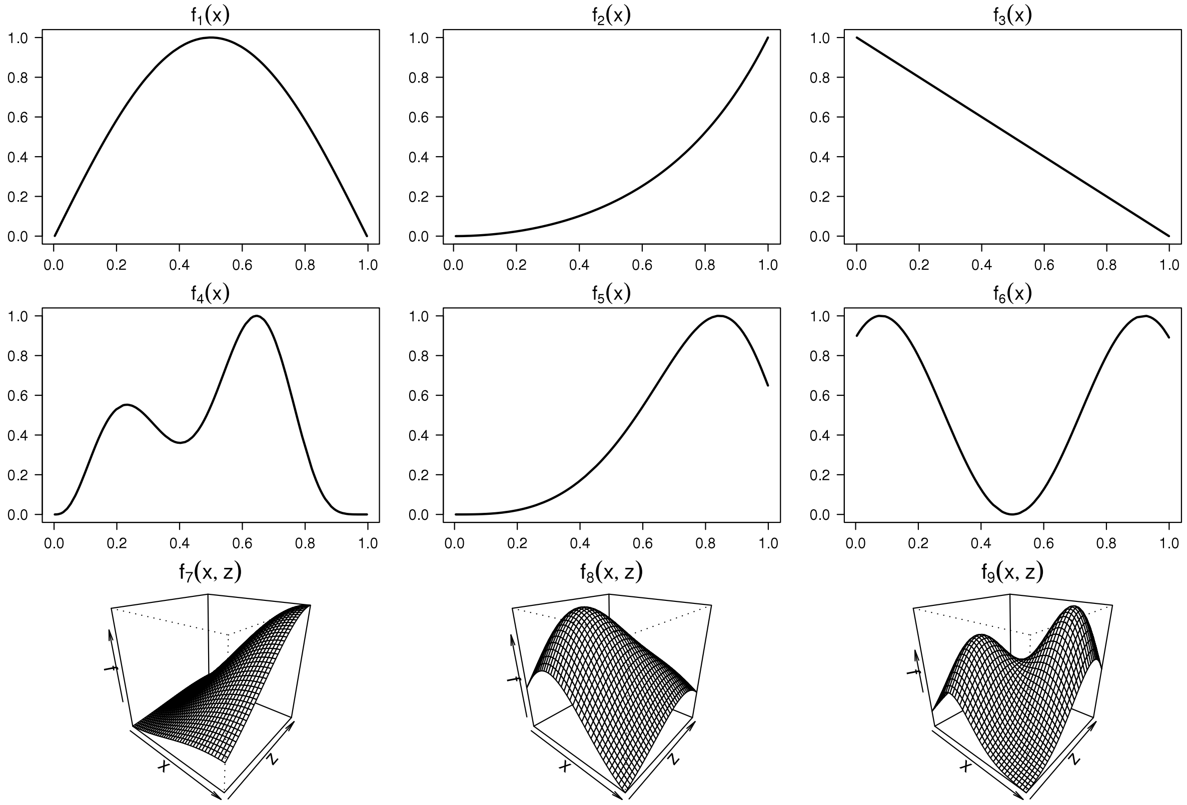

Fig. 1. The test functions used to generate the datasets.

As an alternative, REML can be employed. Within this framework, the penalized likelihood estimates, $\hat{\boldsymbol{\beta}}$, can be seen as the posterior modes of the distribution of $\boldsymbol{\beta}|\mathbf{y}$ if $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{S}^{-}\phi)$, where $\mathbf{S}^{-}$ is an appropriate generalized inverse. Viewing the spline parameters as random effects allows for the possibility to estimate the λ_i via REML (Wahba, 1985). Wahba (1985) showed that asymptotically prediction error criteria are better in a mean square error sense, even though Härdle et al. (1988) pointed out that these criteria give slow convergence to the optimal smoothing parameters. The recent work by Reiss and Ogden (2009) shows that at finite sample sizes GCV or AIC is prone to undersmoothing and is more likely to develop multiple minima than REML (see Figure 1 in Wood, 2011). So, it would appear that REML should be preferred over GCV/AIC especially when the primary purpose of the analysis is to carry out smooth component selection.

The computational methods of Wood (2006, 2008, 2011) implemented in mgcv can be used for reliable multiple smoothing parameter selection according to the criteria above. The simulation study in the next section will shed light on which criteria yield the best results.

3. Simulation study

A simulation study was conducted to compare the practical performance of the methods discussed in the previous section. Under a wide variety of settings, and employing a number of test functions, the procedures were compared in terms of shrinkage properties and a measure of fit.

3.1. Design and model fitting settings

The three linear predictors used for the simulation study are defined as

$$\eta_{1i} = \sum_{j=1}^{6} f_j(x_{ji}), \qquad \eta_{2i} = f_7(x_{7i}, x_{8i}) + f_8(x_{9i}, x_{10i}) + f_9(x_{11i}, x_{12i})$$

$$\text{and} \quad \eta_{3i} = f_1(x_{1i}) + f_3(x_{3i}) + f_4(x_{4i}).$$

The functions are displayed in Fig. 1 and defined in Table 1. Uniform covariates on $(0, 1)$ with equal correlations were obtained using the algorithm from Gentle (2003). For example, using R, three uniform variables with correlations approximately equal

Table 1
Test function definitions. $f_1 - f_9$ are plotted in Fig. 1.

$f_1(x) = 2\sin(\pi x)$
$f_2(x) = \mathrm{e}^{x^2}$
$f_3(x) = -x$
$f_4(x) = x^{11}\{10(1-x)\}^6 + 10(10x)^3(1-x)^{10}$
$f_5(x) = 0.5\{x^3 + \sin(\pi x^3)\}$
$f_6(x) = \cos(2\pi x) + \sin(\pi x)$
$f_7(x,z) = 0.7\mathrm{e}^{-\{(-3x+3)^2+0.7(3z-3)^2\}/5}$
$f_8(x,z) = 0.39\mathrm{e}^{\left\{-\frac{(x-0.3)^2}{0.25}-\frac{(z-0.3)^2}{0.25}\right\}} + 0.20\mathrm{e}^{\left\{-\frac{(x-0.8)^2}{0.25}-\frac{(z-0.8)^2}{0.25}\right\}}$
$f_9(x,z) = 0.16\mathrm{e}^{\left\{-\frac{(x-0.3)^2}{0.25}-\frac{(z-0.3)^2}{0.25}\right\}} + 0.20\mathrm{e}^{\left\{-\frac{(x-0.8)^2}{0.25}-\frac{(z-0.8)^2}{0.25}\right\}}$

Table 2
Observations were generated from the appropriate distribution with true response means, laying in the specified range, obtained by transforming the linear predictors by the inverse of the chosen link function. l, u and s/n stand for lower bound, upper bound and signal to noise ratio parameter, respectively. The linear predictor for the binomial case was scaled to produce probabilities in the range [0.02, 0.98]; observations were then simulated from binomial distributions with denominator n_{bin}. In the gamma case the linear predictor was scaled to have range [0.2, 3] and three levels of ϕ used. For the Gaussian case normal random deviates with mean 0 and standard deviation σ were added to the true expected values, which were then scaled to lay in [0, 1]. The linear predictor of the Poisson case was scaled to yield true means in the interval [0.2, pmax].

	Binomial	Gamma	Gaussian	Poisson
$g(\mu)$	logit	log	identity	log
$l \le \eta \le u$	[0.02, 0.98]	[0.2, 3]	[0, 1]	[0.2, pmax]
s/n	$n_{\text{bin}} = 1, 3, 5$	$\phi = 0.6, 0.4, 0.2$	$\sigma = 0.4, 0.2, 0.1$	pmax = 3, 6, 9

to ρ can be obtained in the following way

```
library(mvtnorm)
cor <- array(c(1,rho,rho,rho,1,rho,rho,rho,1),dim=c(3,3))
var <- pnorm(rmvnorm(n,sigma=cor))
x1 <- var[,1]; x2 <- var[,2]; x3 <- var[,3]
```

This procedure was employed to obtain correlation among all covariates involved in the linear predictor. The cases in which ρ was set to 0 and 0.9 were considered. The functions were scaled to have the same range and then summed. Data were simulated under the four error model—link function combinations detailed in Table 2, at each of three signal to noise ratio levels. The three signal to noise ratio parameters were chosen so that the squared correlation coefficient between μ_i and y_i was about 0.4, 0.55, and 0.7 respectively (see Table 2 for further details). 100 replicate data sets were then generated at each distribution and error level combination.

To maintain computational feasibility and because of limitations applying to some methods, the simulation study did not employ a completely factorial design. Instead it was conducted in the following four phases. In each phase 100 replicates of each combination of conditions were used, 3 noise levels were considered at each of $\rho = 0$ and 0.9:

1. Gaussian identity link models were compared for all methods, for η_1 with 6 nuisance covariates. Both REML and GCV smoothness selection were compared, and the sample size was 200. This phase suggested eliminating the Lasso & Splines method and the Belitz & Lang approach from the subsequent phases (the published versions of these do not treat the generalized case, although for Belitz and Lang, this is not a serious problem).
2. The other three distribution-link models were compared for η_1 with 6 nuisance covariates using all remaining methods. Again REML and GCV were compared where appropriate, and the sample size was 200. This phase suggested that GAM boosting is not competitive, at least for low numbers of nuisance variables. The combination of phases 1 and 2 suggested dropping GCV selection.
3. All distribution link models were compared for η_2 plus 6 nuisance covariates, using all remaining methods except GAM boosting. Smoothness selection was by REML for those methods were there is a choice. The sample size was 200.
4. All distribution link models were compared for all remaining methods including GAM boosting using η_3 and either 11 or 27 nuisance covariates. Sample sizes were 200 for 11 nuisance and 400 for 27. GAM boosting was re-considered here as this situation is the one were it is expected to be competitive.

For phases 1, 2 and 4, all procedures, except for GAM boosting and the Belitz & Lang approach, were implemented using TPRSs (Wood, 2003) based on second-order derivatives and with basis dimensions equal to 10. For phase 3 TPRSs with basis dimensions equal to 20, 20 and 50 were used.

The methods were compared in terms of shrinkage, and mean squared error (MSE) in predicting the linear predictors. To assess the shrinkage properties we used the false negative rate (i.e. rates at which influential covariates are not selected) for the variables in the linear predictors, and false positive rate (i.e. rates at which spurious terms are selected) for non influential covariates. The rates were calculated according to the MSEs rounded up to 7 digits. Notice that using as a criterion edf ≈ 0 led to the same results. Backward selection was carried out at the 5% significance level.

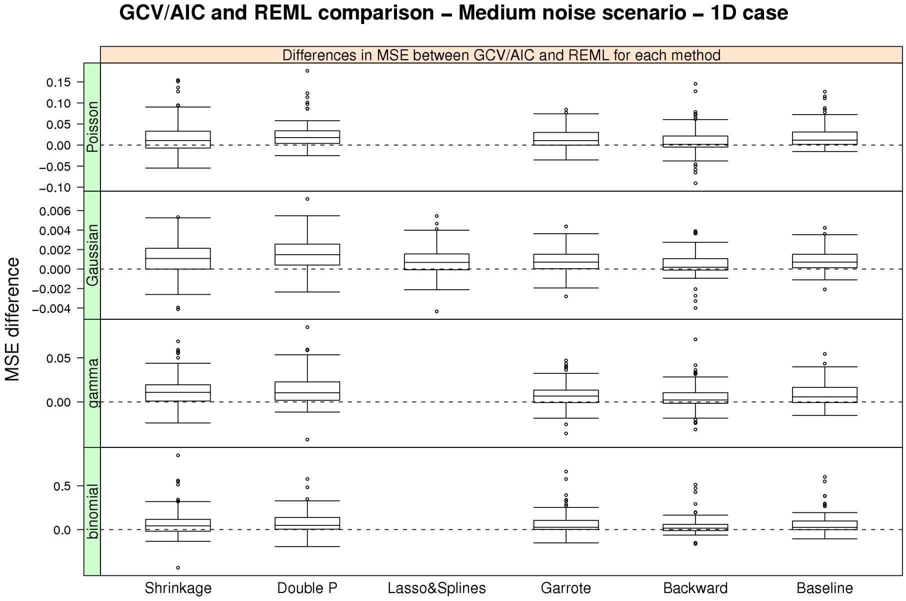

Fig. 2. MSE comparisons between GCV/AIC and REML for four error distributions and methods discussed in Section 2, when using linear predictor η_1. Covariate correlation is 0 and the signal to ratio level is medium. Baseline indicates that no shrinkage smoother is used during the model fitting process. Further simulation details are given in Section 3.1. Boxplots show the distributions of differences in mean squared error between GCV/AIC and REML. In all cases a Wilcoxon signed rank test indicates the REML has lower MSE than GCV/AIC (p-value $< 10^{-2}$).

3.2. Results

To save space, only some of the most important examples are shown. The displayed plots have been chosen to be representative and to convey enough information to draw some general conclusions. Additional plots are available upon request.

Fig. 2 shows the difference in MSE between the same models estimated by GCV/AIC and by REML, for each error model and method combination, from phases 1 and 2 of the simulation study (employing linear predictor η_1). The covariate correlation is 0. Missing box plots within the figures are because the method described in Section 2.5.4 only deals with additive models (and was anyway not competitive in phase 1). REML outperforms GCV/AIC smoothness selection; this suggests that REML allows for better smoothing parameter estimation, hence smooth term estimates are more accurate than when using GCV/AIC. The plots for cases in which $\rho = 0.9$ are omitted since they lead to the same conclusions. In the subsequent plots, we only report the results obtained when using REML.

Fig. 3 compares the MSE performance of all the methods discussed in the paper, relative to the double penalty approach of Section 2.1, from phases 1 and 2 of the study (above the zero line indicates performance worse than the double penalty method). Notice that `GAMBoost` supports only canonical link functions, hence MSE results for the gamma case are not available. Our results indicate that, overall, the double penalty approach performs significantly better than the competing methods in terms of MSE.

Figs. 4 and 5 show the false positive rates for the methods considered here and the four error models, at each signal to noise ratio level. GAM boosting is not competitive. This result is in agreement with the findings of Cantoni et al. (2011). Shrinkage and double penalty are competitive as compared to the alternatives. The nonnegative garrote estimator is also competitive but not for the binomial case. As covariate correlation increases, the nonnegative garrote performance worsens. Backward selection yields the best results, but false negative rates are about 0.4, 0.3, and 0.1. These increase by about 0.1 point when the covariate correlation is 0.9. So, if the data have high information content then Backward selection may be preferred over the competing methods, otherwise our proposals yield the most reliable results.

The Belitz & Lang approach yields good false positive rates. However, false negative rates (plots not shown here) indicate that this method eliminates influential covariates with rates about 0.25, 0.18, and 0.09 for the high, medium and low noise cases, respectively. This also explains its MSE performance as compared to the other approaches (see Fig. 3). False negative rates are about 0.60, 0.29 and 0.17 when the covariate correlation is 0.9. The combination of high false negative rates and relatively high MSE led us to drop the Belitz & Lang method after phase 1 of the study.

The poor false positive rate performance of `GAMBoost` is because the procedure typically retains predictors whose estimated curves are close to the zero line and that have been selected in a small number of boosting steps. These two facts could be combined into a procedure to improve false positive rates. However this presents us with some difficulty in obtaining fair false positive rate criteria for comparison with other methods, so is not pursued here. Alternative boosting

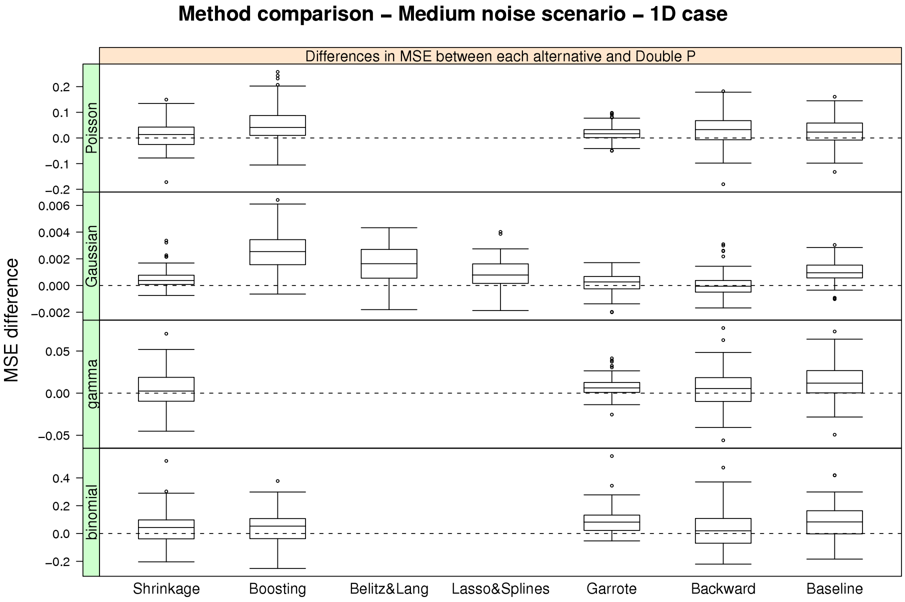

Fig. 3. MSE results between the methods discussed in Section 2 and the double penalty approach for four error distributions and linear predictor η_1. REML estimation is employed for all methods except for GAM boosting and the Belitz & Lang approach. Covariate correlation is 0 and the signal to ratio level is medium. The Baseline indicates that no shrinkage smoother is used during the model fitting process. Further simulation details are given in Section 3.1. Boxplots show the distributions of differences in mean squared error between each method and the double penalty approach. In all cases a Wilcoxon signed rank test indicates that the double penalty has lower MSE than the competing methods (p-value $< 10^{-6}$), except for the Backward method in the Gaussian and Binomial cases where there is no significant difference (p-value > 0.10).

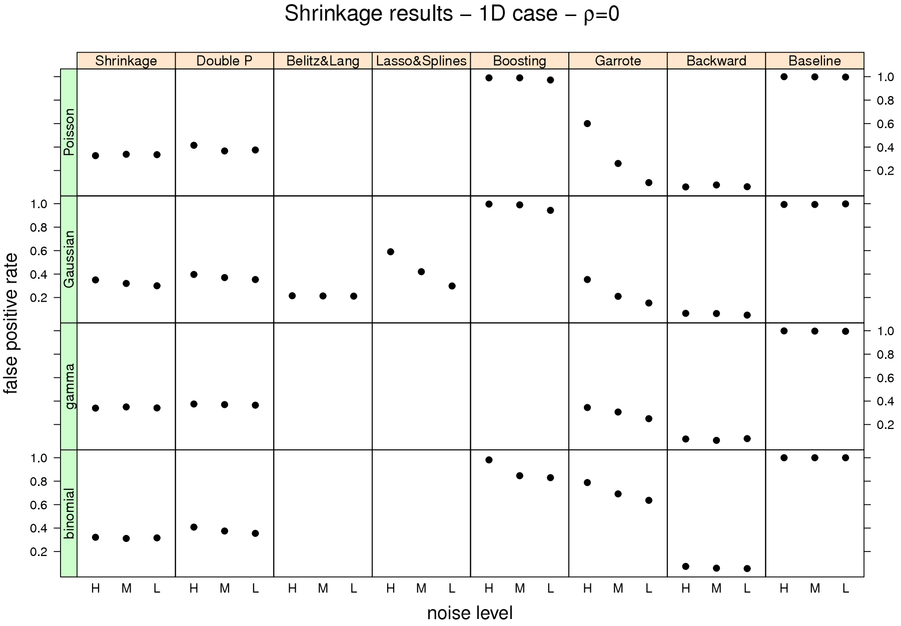

Fig. 4. Shrinkage results for the methods discussed in Section 2, for four error distributions and linear predictor η_1. REML estimation is employed for all methods except for GAM boosting and the Belitz & Lang approach. Covariate correlation is 0 and H, M and L stand for high, medium and low signal level. The Baseline indicates that no shrinkage smoother is used during the model fitting process. Further simulation details are given in Section 3.1. False positive rates give the proportion of times spurious terms are selected. Vertical lines show ±2 standard error bands.

procedures such as those documented in Bühlmann and Hothorn (2010) and Shafik and Tutz (2009) should also improve performance, but in the absence of public domain software we do not pursue these approaches here.

Fig. 6 compares the MSE of the methods considered in phase 3 (linear predictor η_2) to the MSE of the double penalty approach. Results are again given by error model and method. The results confirm the finding that the double penalty approach yields overall the smallest MSEs. Similar conclusions were obtained when $\rho = 0.9$. Fig. 7 indicates that shrinkage

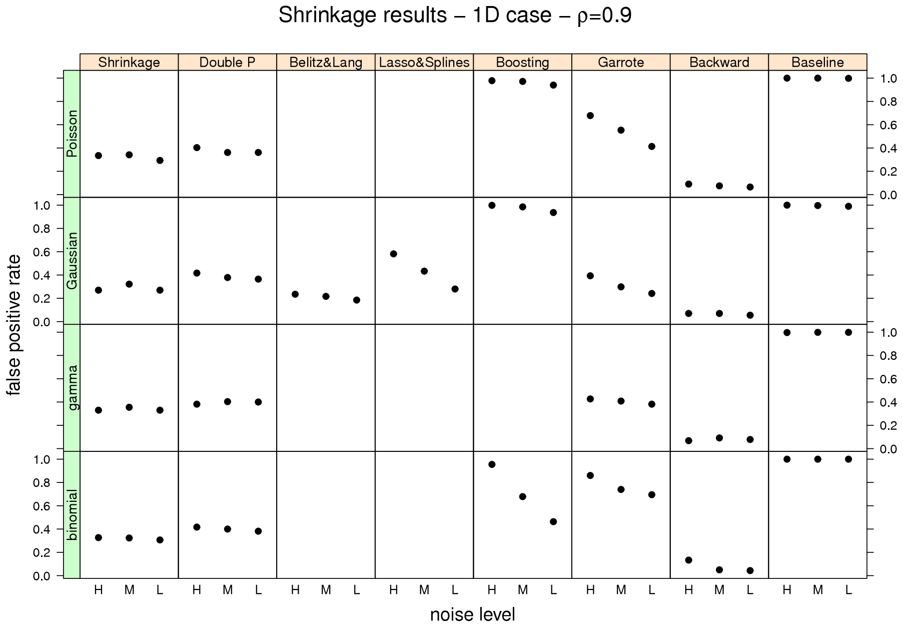

Fig. 5. Shrinkage results for the methods discussed in Section 2, for four error distributions and linear predictor η_1. REML estimation is employed for all methods except for GAM boosting and the Belitz & Lang approach. The covariate correlation is 0.9. Further details are given in the caption of Fig. 4.

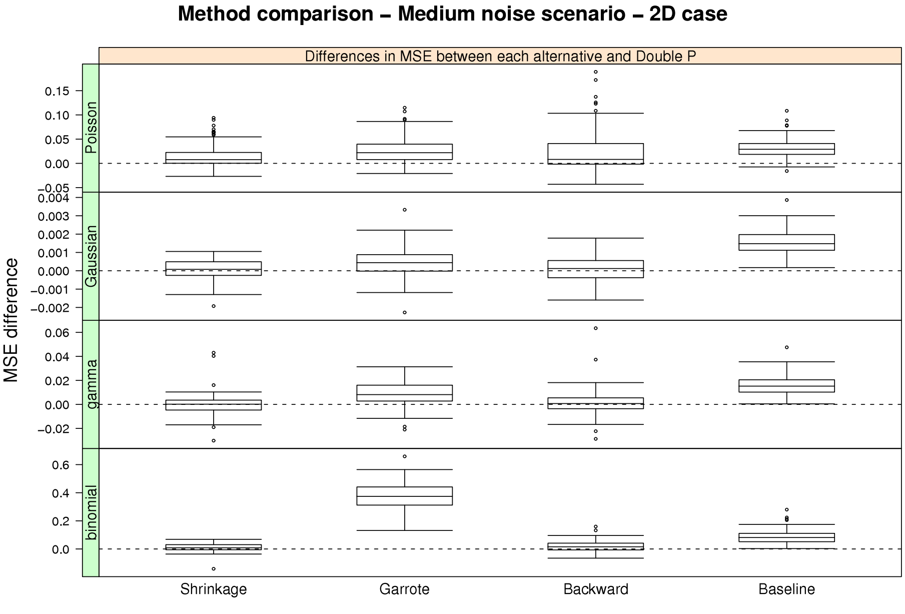

Fig. 6. MSE comparisons between some of the methods discussed in Section 2 and the double penalty approach for four error distributions, when REML estimation and linear predictor η_2 are employed. The covariate correlation is 0 and the signal to ratio level is medium. The Baseline indicates that no shrinkage smoother is used during the model fitting process. Further simulation details are given in Section 3.1 and in the caption of Fig. 3. In all cases a Wilcoxon signed rank test indicates that double penalty has lower MSE than the competing methods (p-value $< 10^{-6}$).

and double penalty yield, overall, reasonable false positive rate results. The nonnegative garrote estimator also produces reasonable results but not for the binomial case. As before, false negative rates (not reported here) indicate the Backward selection should be preferred over the other methods if the data have high information content.

Finally, Figs. 8 and 9 show MSE comparisons by error model and method for linear predictor η_3 from phase 4 of the simulations, when models are fitted using fourteen and thirty covariates, respectively, most of which are nuisance variables. The results confirm the findings of this section, with the only difference that double penalty does not outperform the shrinkage approach. The overall shrinkage performance of the methods for the two scenarios was in line with that reported

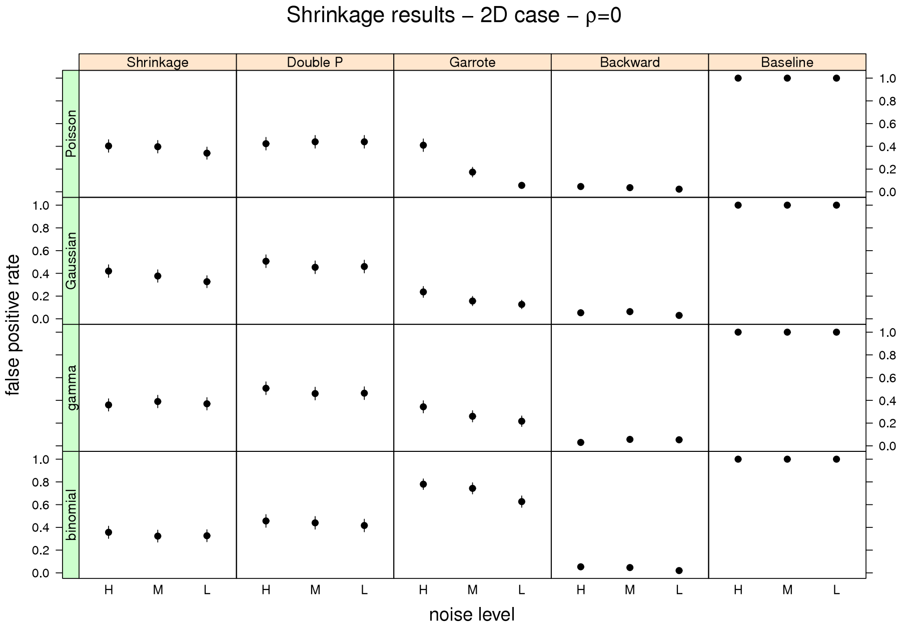

Fig. 7. Shrinkage results for some of the methods discussed in Section 2 and four error distributions, when REML estimation and linear predictor η_2 are employed. The covariate correlation is 0. Further simulation details are given in Section 3.1 and in the caption of Fig. 4.

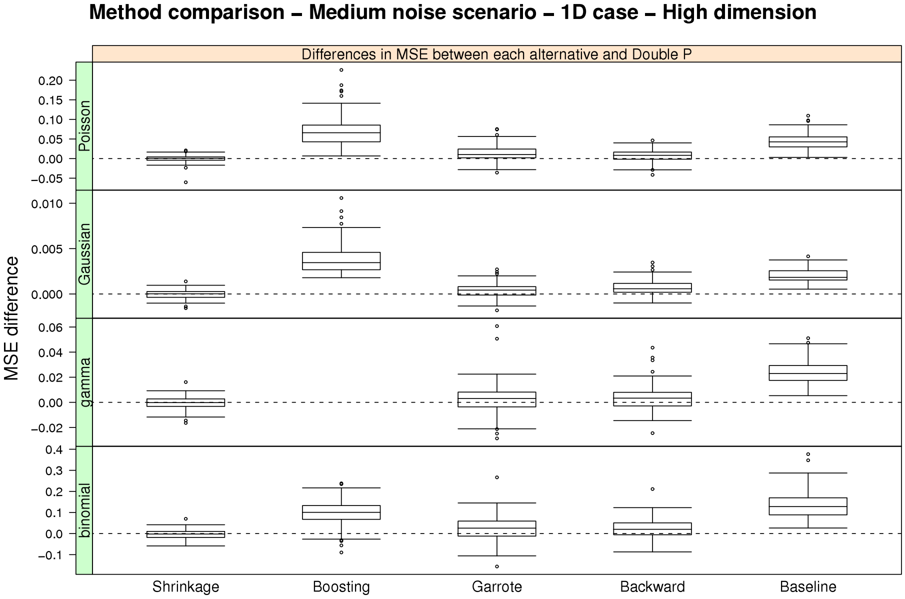

Fig. 8. MSE comparisons between some of the methods discussed in Section 2 and the double penalty approach for four error distributions and linear predictor η_3. REML estimation is employed for all methods except for GAM boosting. Models are fitted using fourteen covariates, eleven of which are not influential. The covariate correlation is 0 and the signal to ratio level is medium. Further simulation details are given in Section 3.1 and in the caption of Fig. 3. In all cases a Wilcoxon signed rank test indicates that double penalty has lower MSE than the competing methods (p-value $< 10^{-5}$), except for the Shrinkage approach where there is no significant difference (p-value > 0.29).

in the previous plots (overall false positive rates about 0.33, 0.37, 0.84, 0.51, 0.09, and 1 for Shrinkage, Double penalty, Boosting, Garrote, Backward and Baseline, respectively).

4. Real data example

In this section, we show the results obtained by applying the methods discussed in the paper to a real dataset on plasma beta-carotene levels.

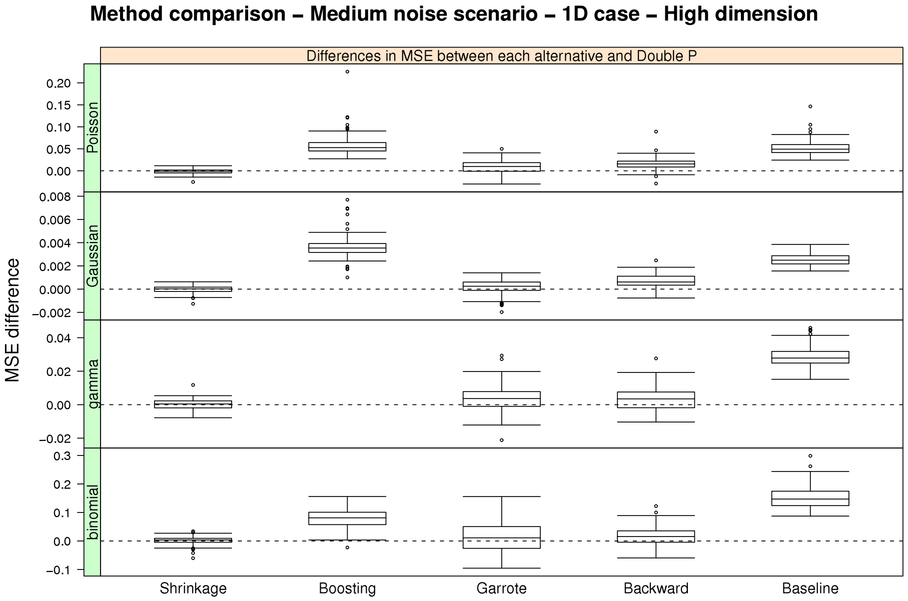

Fig. 9. MSE comparisons between some of the methods discussed in Section 2 and the double penalty approach for four error distributions and linear predictor η_3. REML estimation is employed for all methods except for GAM boosting. Models are fitted using thirty covariates, twenty-seven of which are spurious. Covariate correlation is 0 and the signal to ratio level is medium. Further simulation details are given in Section 3.1 and in the caption of Fig. 3. In all cases a Wilcoxon signed rank test indicates that double penalty has lower MSE than the competing methods (p-value $< 10^{-6}$), except for the Shrinkage approach where there is no significant difference (p-value > 0.33).

4.1. Beta-carotene data

The data are from a cross-sectional study conducted in the United States. The aim of the analysis was to investigate the relationship between beta-carotene plasma concentrations and personal characteristics as well as dietary variables of subjects who had a biopsy examination or removed lesions of the lung, colon, breast, skin, ovary or uterus that were not found to be cancerous (Nierenberg et al., 1989; Marra and Radice, 2010). The dataset was obtained from the StatLib-Datasets Archive website (http://lib.stat.cmu.edu/datasets/Plasma_Retinol) and is made up of 315 individuals. The dataset contains a number of continuous variables.

The covariates considered were *age* (in years), *Quetelet index* (which is a measure of obesity defined as weight divided by the square of height), number of *calories* consumed per day, *plasma beta-carotene* (ng/ml), grams of *fat* consumed per day, grams of *fiber* consumed per day, *cholesterol* (mg per day), and *dietary beta-carotene* (mcg per day).

4.2. Results

The aim was to fit a nonparametric model and perform variable selection. As pointed out in Marra and Radice (2010), plasma BC levels strongly exhibit a positively skewed distribution. Therefore, a gamma distribution with a log link function between the linear predictor and the mean was employed. Notice that GAM boosting, the Lasso & Splines approach and Belitz & Lang method were not applied on this dataset for the reasons given in the previous section. We applied the remaining methods using each of GCV and REML with model fitting settings as discussed in Section 3.1. Pearson correlations among the covariates were in the range $[0.05, 0.9]$, and the squared correlation coefficient between μ_i (calculated using a standard GAM) and y_i was about 0.25, suggesting that the noise in this dataset is high.

According to the shrinkage, double penalty and nonnegative garrote approaches the variables *calories* and *fat* were not influential, hence removed from the model. Backward selection removed *calories, fat* and *fiber*, suggesting the elimination of an important covariate. In fact, this was consistent with our simulation study which showed that if the data do not have high information content then Backward selection eliminates influential predictors.

Fig. 10 shows the smooth function estimates obtained by applying the double penalty approach on the plasma beta-carotene dataset. Similar results were obtained by using the shrinkage and nonnegative garrote methods (plots not reported here). The estimated functions reveal the presence of non-linear relationships between the outcome and the selected regressors. This allows the researcher to gain more insights into the phenomenon of plasma beta-carotene in comparison to using a fully parametric approach. The smooths of *dietary beta-carotene* and *fiber* exhibit a linear behaviour, hence these terms can enter the model in a parametric manner. Finally, we repeated 5-fold cross validation 100 times, and then calculated prediction risk estimates. The results, displayed in Fig. 11, are consistent with the findings of our simulation study; overall, REML outperforms GCV and the double penalty approach performs significantly better than the competing methods in terms of prediction.

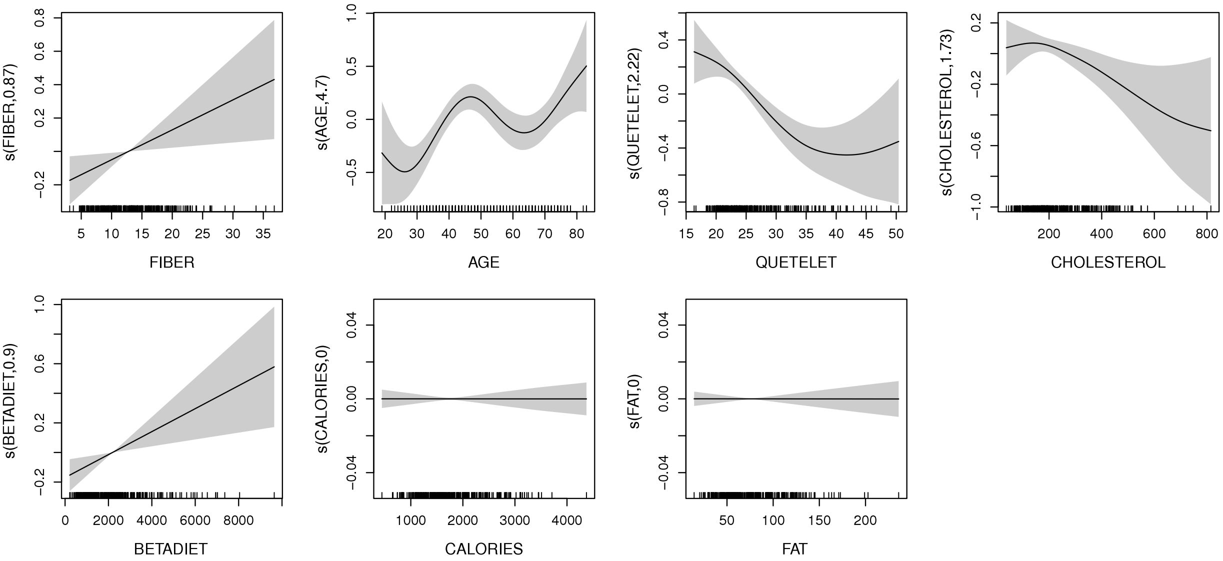

Fig. 10. Smooth function estimates obtained applying the double penalty approach with REML estimation on the plasma beta-carotene dataset described in Section 4.1. The results are reported on the scale of the linear predictor. The numbers in brackets in the y-axis captions are the edf of the smooth curves. The 'rug plot', at the bottom of each graph, shows the covariate values. The shaded regions represent the usual 95% Bayesian 'confidence' intervals (e.g. Wood, 2006). Due to the identifiability constraints, the estimated curves pass through zero; consequently, there is no uncertainty about this point.

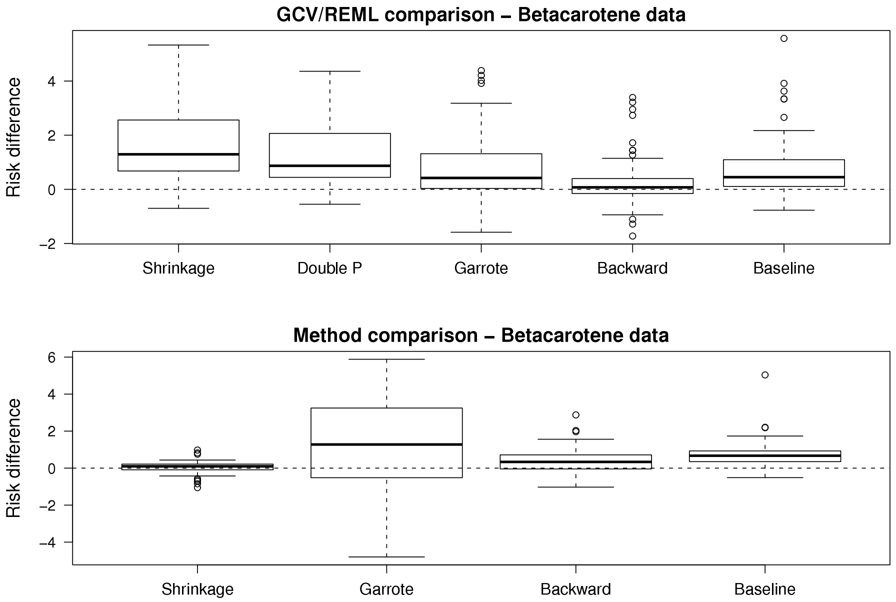

Fig. 11. The top boxplots report prediction risk comparisons (in units of 10^3) between GCV and REML for some of the methods discussed in Section 2 when using the beta-carotene dataset (see details in Section 4). The plots show the distributions of differences in prediction risk estimate between GCV and REML, which were obtained repeating 5-fold cross validation 100 times. In all cases a Wilcoxon signed rank test indicates the REML yields lower risk estimates as compared to GCV (p-value $< 10^{-19}$), except for Backward where this evidence is less strong (p-value < 0.022). The bottom boxplots report prediction risk comparisons between the four shrinkage methods used for the beta-carotene dataset and the double penalty approach, when REML estimation is employed. The plots show the distributions of differences in prediction risk estimate between each method and double penalty. In all cases a Wilcoxon signed rank test indicates that double penalty produces lower risk estimates as compared to the competing methods (p-value $< 10^{-18}$), except for Shrinkage where this evidence is less strong (p-value < 0.017).

To complete the analysis, following e.g. Bühlmann and Hothorn (2010), we looked at a synthetically enlarged problem. Specifically, we generated ten uniform variables with correlations approximately equal to 0.5 (see Section 3.1) and included them in the model containing the real predictors. This allowed us to check how many ineffective variables would be selected. The proposed approaches performed satisfactorily in that, overall, seven variables out of ten were eliminated. This was consistent with the simulation results.

5. Conclusions

In this article, we have proposed two effective shrinkage methods and extended the nonnegative garrote estimator to achieve component selection within GAMs, for situations in which there are moderate numbers of spurious covariates which it would be beneficial to eliminate. We have compared the empirical performance of the proposals to that of some available techniques via an extensive simulation study, and illustrated some of the procedures analysing data on plasma beta-carotene levels from a cross-sectional study conducted in the United States.

Our results show that, overall, the proposed shrinkage approaches perform significantly better than the competing methods in terms of predictive ability. As for the variable selection performance, the shrinkage and double penalty approaches are competitive as compared to the alternatives. The nonnegative garrote estimator is also competitive but not for the binomial case. As covariate correlation increases, the nonnegative garrote performance worsens. Matters improve when REML is employed for smoothing parameter estimation. Backward selection yields the best false positive rates. However, false negative rates indicate that this method eliminates influential covariates, especially for the low signal to noise ratio level-high covariate correlation scenario, worsening its predictive ability. GAM boosting performs comparatively poorly in the scenarios considered here, although its shrinkage performance could almost certainly be improved by changing the way variables are selected.

If the data have high information content then backward selection may be preferred over alternatives, otherwise our proposals yield the most reliable results. The main limitation of all the methods discussed here, except for GAM boosting and the Belitz & Lang approach, is that they cannot deal with situations in which $n < p$ (n is sample size and p is the number of predictors) and indeed require $n \geq kp$ where k is the average basis size used for the smoothers. For $p > n$, GAM boosting and the Belitz & Lang method appear to be the only methods available when an additive structure is considered appropriate. It is notable that both these methods rely on prediction error criteria such as AIC and GCV for smoothness selection, while for the other methods we found REML smoothness selection to generally yield superior results. It would be interesting to see whether REML based extensions to Belitz and Lang (2008) or GAM boosting could be produced which would improve the performance of these methods.

Acknowledgements

We wish to thank one anonymous referee for suggesting the comparison to Belitz and Lang (2008), and for suggestions on the penalty interpretation, and another anonymous referee for suggesting the simulations with larger numbers of nuisance covariates and some clarifications on GAM boosting.

References

Avalos, M., Grandvalet, Y., Ambroise, C., 2007. Parsimonious additive models. Computational Statistics and Data Analysis 51, 2851–2870.

Belitz, C., Lang, S., 2008. Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis 51, 6044–6059.

Binder, H., Tutz, G., 2008. A comparison of methods for the fitting of generalized additive models. Statistics and Computing 18, 87–99.

Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37, 373–384.

Bühlmann, P., Hothorn, T., 2010. Twin boosting: improved feature selection and prediction. Statistics and Computing 20, 119–138.

Bühlmann, P., Yu, B., 2003. Boosting with the L2 loss: regression and classification. Journal of the American Statistical Association 98, 324–339.

Cantoni, E., Flemming, J., Ronchetti, E., 2011. Variable selection in additive models by nonnegative garrote. Statistical Modelling 11, 165–180.

Cantoni, E., Hastie, T., 2002. Degrees of freedom tests for smoothing splines. Biometrika 89, 251–263.

Chib, S., Greenberg, E., 2007. Semiparametric modeling and estimation of instrumental variable models. Journal of Computational and Graphical Statistics 16, 86–114.

Daye, Z.J., Jeng, X.J., 2009. Shrinkage and model selection with correlated variables via weighted fusion. Computational Statistics and Data Analysis 53, 1284–1298.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. The Annals of Statistics 32, 407–451.

Gentle, J.E., 2003. Random Number Generation and Monte Carlo Methods. Springer-Verlag.

Greven, S., Kneib, T., 2009. On the behavior of marginal and conditional Akaike information criteria in linear mixed models. Johns Hopkins University, Department of Biostatistics Working Papers. Paper 179.

Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157, 89–100.

Härdle, W., Hall, P., Marron, J.S., 1988. How far are automatically chosen regression smoothing parameters from their optimum? Journal of the American Statistical Association 83, 86–95.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, London.

Hesterberg, T., Choi, N.H., Meier, L., Fraley, C., 2008. Least angle and l_1 penalized regression: a review. Statistics Surveys 2, 61–93.

Kauermann, G., Claeskens, G., Opsomer, J.D., 2009. Bootstrapping for penalized spline regression. Journal of Computational and Graphical Statistics 18, 126–146.

Kauermann, G., Tutz, G., 2001. Testing generalized linear and semiparametric models against smooth alternatives. Journal of the Royal Statistical Society Series B 63, 147–166.

Kimeldorf, G., Wahba, G., 1970. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. Annals of Mathematical Statistics 41, 495–502.

Lin, Y., Zhang, H.H., 2006. Component selection and smoothing in multivariate nonparametric regression. Annals of Statistics 34, 2272–2297.

Marra, G., Radice, R., 2010. Penalised regression splines: theory and application to medical research. Statistical Methods in Medical Research 19, 107–125.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall, London.

Nierenberg, D.W., Stukel, T.A., Baron, J.A., Dain, B.J., Greenberg, E.R., 1989. The skin cancer prevention study group. Determinants of plasma levels of beta-carotene and retinol. American Journal of Epidemiology 130, 511–521.

Nott, D.J., Leng, C., 2010. Bayesian projection approaches to variable selection in generalized linear models. Computational Statistics and Data Analysis 54, 3227–3241.

R Development Core Team, 2010. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0. URL: http://www.R-project.org.

Reiss, P.T., Ogden, R.T., 2009. Smoothing parameter selection for a class of semiparametric linear models. Journal of the Royal Statistical Society Series B 71, 505–524.

Scheipl, F., Greven, S., Kuchenhoff, H., 2008. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. Computational Statistics and Data Analysis 52, 3283–3299.

Shafik, N., Tutz, G., 2009. Boosting nonlinear additive autoregressive time series. Computational Statistics and Data Analysis 53, 2453–2464.

Similä, T., Tikka, J., 2007. Input selection and shrinkage in multiresponse linear regression. Computational Statistics and Data Analysis 52, 406–422.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B 58, 267–288.

Tutz, G., Binder, H., 2006. Generalized additive modeling with implicit variable selection by likelihood-based boosting. Biometrics 62, 961–971.

Tutz, G., Binder, H., 2007. Boosting ridge regression. Computational Statistics and Data Analysis 51, 6044–6059.

Wager, C., Vaida, F., Kauermann, G., 2007. Model selection for penalized spline smoothing using akaike information criteria. Australian and New Zealand Journal of Statistics 49, 173–190.

Wahba, G., 1985. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. The Annals of Statistics 13, 1378–1402.

Wood, S.N., 2003. Thin plate regression splines. Journal of the Royal Statistical Society Series B 65, 95–114.

Wood, S.N., 2006. Generalized Additive Models: An Introduction with R. Chapman & Hall, London.

Wood, S.N., 2008. Fast stable direct fitting and smoothness selection for generalized additive models. Journal of the Royal Statistical Society Series B 70, 495–518.

Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society Series B 73, 3–36.

Xue, L., 2009. Consistent variable selection in additive models. Statistica Sinica 19, 1281–1296.

Yuan, M., 2007. Nonnegative garrote component selection in functional anova models. In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics. Puerto Rico.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B 68, 49–67.

Zhang, H.H., Lin, Y., 2006. Component selection and smoothing for nonparametric regression in exponential families. Statistica Sinica 16, 1021–1041.

Zou, H., 2006. The adaptive LASSO and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.