Universidade Vigo

FWDselect: Variable selection algorithm in regression models

Marta Sestelo, Nora M. Villanueva and Javier Roca-Pardiñas

Report 13/02

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

Facultade de Ciencias Económicas e Empresariales
Lagoas-Marcosende, s/n · 36310 Vigo
Tfno.: +34 986 812440 - Fax: +34 986 812401
http://webs.uvigo.es/depc05/
E-mail: depc05@uvigo.es

Universidade Vigo

FWDselect: Variable selection algorithm in regression models

Marta Sestelo, Nora M. Villanueva and Javier Roca-Pardiñas

Report 13/02

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita: Universidade Vigo

Facultade de CC. Económicas e Empresariales
Departamento de Estatística e Investigación Operativa
As Lagoas Marcosende, s/n 36310 Vigo
Tfno.: +34 986 812440

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

`FWDselect`: Variable selection algorithm in regression models

Marta Sestelo[a,*], Nora M. Villanueva[a], Javier Roca-Pardiñas[a]

[a]*Department of Statistics and Operations Research, University of Vigo, C/ Torrecedeira 86, E-36280 Vigo, Spain.*

January 2013

Abstract

In multiple regression models, when there is a large number p of explanatory variables which may or may not be relevant for making predictions about the response, it is useful to be able to reduce the model. To this end, it is necessary to determine the best subset or subsets of q ($q \leq p$) predictors which will establish the model or models with the best prediction capacity. Here, we present a new approach to this problem, the `FWDselect` package which introduces a simple method to select the best model using different types of data (continuous, binary or poisson) and applying it in different contexts (parametric or nonparametric). The developed methodology includes two topics: i) to select the best combinations of q variables by using a new forward stepwise-based selection procedure and perhaps, most importantly, ii) to determine the number of covariates to be included in the model based on bootstrap resampling techniques. The software is illustrated using pollution data.

Keywords: variable selection, forward, bootstrap

*Corresponding author. Tel.: +34 986813948; fax: +34 986813644

Email address: `sestelo@uvigo.es` (Marta Sestelo)

1. Introduction

In a multivariate regression framework, the target response Y can depend on a set of p initial covariates $X_1, X_2, \ldots, X_p$ but in practical situations, one has to decide which covariates are "relevant" to describe this response. A question that tends to arise in regression models, and that has not been totally satisfactorily solved yet, is determining the best subset or subsets of q ($q \leq p$) predictors which will establish the model or models with the best predictive capability. This problem is particularly important when p is high and/or when there are redundant predictors. As a general rule, an increase in the number of variables to be included in a model provides an "apparently" better fit of the observed data. However, these estimates are not always satisfactory for different reasons. On the one hand, inclusion of irrelevant variables would increase the variance of the estimates, resulting to a partial loss of the predictive capability of the model; and on the other hand, inclusion of many variables would mean that the model would be difficult to interpret.

Model selection (and variable selection in regression, in particular) is a trade-off between bias and variance. This is the statistical principle of parsimony. Inference based on models with few variables can be biased, however, models that take into account too many variables may result in a lack of precision or false effects. These considerations call for a balance between under- and over-fitted models, the so-called model-selection problem (Forster, 2000).

To solve this problem, there are several procedures in the literature, e.g. shrinkage regression methods, such as ridge regression or the Lasso (least absolute shrinkage and selection operator) (Tibshirani, 1996; Hastie et al., 2003), the Bayesian approach, (Green, 1995; Kuo and Mallick, 1998; Park and Casella, 2008) or iterative procedures, such as stepwise, based on the use of some information criteria to compare the model obtained in the course of the simplification or complexification scheme. Several criteria have been used for this purpose, including Mallow's Cp or the Akaike Information Criteria (AIC) (Venables and Ripley, 1997; Miller, 2002).

Another option is to use a full information criteria-based approach, which compares all possible models and ranks them (Calcagno and de Mazancourt, 2010). On the one hand, this procedure enables us to find the "best" model —according to the criterion— and on the other hand, and more importantly, this method allows for better assessment of model-selection uncer-

tainty and better performance of multi-model inference than a single model would (Burnham and Anderson, 2002; Johnson and Omland, 2004; Calcagno and de Mazancourt, 2010). An example of this procedure is Roca-Pardiñas et al. (2009), where selection of variables is based on searching through all the possible subsets. Nevertheless, there is a problem associated with its use. If there is a large number of variables, this selection procedure may require an excessively high computational cost (e.g., if $p = 20$, the number of estimated models will be 1 048 575), and the problem thus becomes intractable.

In view of the above, we now propose and implement an adaptation of the previous method, a new forward stepwise-based selection procedure that greatly reduces computational costs. The methodology developed includes the following two topics: i) selecting the best combination of q variables by using a step-by-step procedure; and perhaps more importantly, ii) determining the number of covariates to be included in the model, based on bootstrap resampling techniques.

Several software or R packages (R Core Team, 2012) have been developed to carry out automated variable selection or model selection. For instance, the `meifly` package (Wickham, 2012) can be used to search through all the different models. In other case, this search is based on some algorithm as in `leaps` (Lumley and Miller, 2009) which uses a branch-and-bound algorithm or `subselect` (Orestes Cerdeira et al., 2011) that implements a simulated-anneling algorithm. When it comes to model selection with Generalized Additive Models, an option could be to use the `glmulti` package (Calcagno, 2012) or `bestglm` (Mcleod and Xu, 2011). Additionally, another procedure used by the R community seems to be the model selection oriented function `step`, builtin in package `stat` (Hastie and Pregibon, 1992). Here we introduce an alternative, we add over existing approach, a simple method for the R users to select the best model which can be applied to different types of data (such as binary, continuous or poisson) and in different contexts (parametric or nonparametric).

The developed methodology in this paper and implemented in `FWDselect` is tested on the prediction of atmospheric SO_2 pollution incidents. One of the problems that arises is to decide which temporal instants are relevant for prediction purpose, since inclusion of all the times may well degrade the overall performance of the prediction model.

The aim of this paper is to present a new software package for the statistical computing environment R. The functions contained in the package address the important aspect of selecting variables in the regression context.

We structure this paper as follows. Section 2 describes the forward algorithm used to select the best subset of size q, both the bootstrap techniques that were used to determine the number of variables to be included in the model and the step-by-step procedure used to select them. To assess the validity of these procedures, two simulation studies are provided in Section 3. In Section 4 we describe the implementation of package `FWDselect`. Finally, Section 5 illustrates the packages' capabilities using to predict a real pollution incident, and Section 6 concludes with some remarks.

2. Variable selection algorithm

This Section introduces the developed methodology and gives a description of the variable selection algorithm. The implemented procedure can be used with different types of models (parametric or nonparametric). Here, we explain it using a nonparametric regression model with continuous response.

Let $\mathbf{X} = X_1, X_2, \ldots, X_p$ be the set of p initial variables and Y the response, an additive regression model can be expressed as

$$Y = m(\mathbf{X}) + \varepsilon, \tag{1}$$

where

$$m(\mathbf{X}) = \alpha + m_1(X_1) + m_2(X_2) + \cdots + m_p(X_p)$$

where $m_j (j = 1, \ldots, p)$ are smooth and unknown functions and ε is the zero-mean error. Additionally, to guarantee the identification of the above model, a constant α is introduced in the model and it is required that the partial function satisfy

$$E[m_j(X_j)] = 0, \quad j = 1, \ldots, p. \tag{2}$$

This implies that $E[Y] = \alpha$.

To date, several approaches to estimating the model in (1) have been suggested in the statistical literature, e.g., Buja et al. (1989); Härdle and Hall (1993); Mammen et al. (1999). We estimate the latter using the backfitting algorithm (Opsomer, 2000). This algorithm cycles through the covariates X_j $(j = 1, \ldots, p)$ and estimates each m_j by applying local polynomial kernel smoothers to the partial residuals. These residuals are obtained by removing the estimated effects of the others covariates.

Given a sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ the steps of the estimation algorithm are as follows:

Initialize. Compute the initial estimates $\hat{\alpha} = \sum_{i=1}^n Y_i/n$ and $\hat{m}_j^0(X_{ij})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Step 1. For $j = 1, \ldots, p$ calculate the residuals by removing the estimated effects of all the others covariates:

$$r_z^j = Y_z - \hat{\alpha} - \sum_{s=1}^{j-1} \hat{m}_s(X_{zs}) - \sum_{s=j+1}^{p} \hat{m}_s^0(X_{zs}), \quad z = 1, \ldots, n$$

and compute for $i = 1, \ldots, n$ the local polynomial kernel estimator

$$\hat{m}_j(X_{ij}) = \hat{\alpha}_0(X_{ij})$$

with $\hat{\alpha}_0(X_{ij})$ the first of the vector $(\hat{\alpha}_0(X_{ij}), \hat{\alpha}_1(X_{ij}), \ldots, \hat{\alpha}_q(X_{ij}))$ which is the minimiser of

$$\sum\nolimits_{z=1}^{n} \left\{ r_z^j - \sum\nolimits_{s=0}^{q} \alpha_s \left(X_{ij}\right) \left(X_{zj} - X_{ij}\right)^s \right\}^2 \cdot K\left(\frac{X_{zj} - X_{ij}}{h_j}\right),$$

where K denotes a kernel function (a symmetric density), h_j the bandwidth associated with the estimation of $\hat{m}_j$ and q the order of the polynomial.

Finally, in order to meet the identifiability condition (2) the resulting estimate $\hat{m}_j(\cdot)$ is replaced by its centered version

$$\hat{m}_j(\cdot) - \frac{\sum_{i=1}^n \hat{m}_j(X_{ij})}{n}.$$

Step 2. Repeat **Step 1** with $\hat{m}_j^0$ replaced by $\widehat{m}_j$ until the convergence criterion

$$\frac{\sum_{i=1}^n \left[\hat{m}_j(X_{ij}) - \hat{m}_j^0(X_{ij})\right]^2}{\sum_{i=1}^n \hat{m}_j^0(X_{ij})^2} \leq \epsilon$$

for all the $j = 1, \ldots, p$ where ϵ is a small threshold.

In some circumstances, the generalized additive models extends the additive models by allowing for different distributions of the response. In these models the relationship between $E[Y|\mathbf{X}]$ and the covariates is defined as follows

$$E[Y|\mathbf{X}] = g(\alpha + m_1(X_1) + m_2(X_2) + \cdots + m_p(X_p)),$$

where g is an unknown function (the inverse of the link function). The selection procedure that we propose in this paper can also be used in this type of models.

It is important to highlight that, in situations involving a large number of variables, correct estimation of the response will be obtained on the basis of selecting the appropriate predictors. In the case that we have information a priori about which of the initial set of variables are relevant, it would be possible to apply a likelihood ratio test (Neyman and Pearson, 1928) or a F-test type (Seber and Wild, 1989; Seber, 1997) in a parametric framework, or a generalized likelihood ratio test (Fan et al., 2001; Fan and Jiang, 2005, 2007) in a nonparametric one. However, in situations where we do not have information in advance, it will be necessary to select the model according to a selection algorithm.

As we mentioned, there are described in the literature traditional or classical parametric procedures to select the appropriate model. These procedures try to simplify the maximum model —containing all possible explanatory variables— to a reduced model that only contains the variables which provide important information about the response. These methods involve two topics, the choice of the selection criterion —a criterion which will order all possible models from "best" to "worst"— and the choice of the selection procedure —a procedure to locate this "best" model.

In relation with the first issue, many different criteria have been suggested through time. The most common criteria could be: i) the coefficient of determination or R^2 which refers to the proportion of the total amount of variation in the data which can be explained by the fitted model, ii) the F-test criterion, which tests if a reduced model provides as good fit to the data as the maximum model and iii) the Mallows's Cp criterion (Mallows, 1973) which compares the unbiased estimate of the error variance between the reduced and the maximum model.

According to the selection strategy, the traditional procedures deal with: i) the all possible models procedure, where all possible models are fitted and compared using some criteria to choose the best one, ii) the forward selection and backward elimination procedures, which concentrate on deciding if each of the explanatory variables should, or should not, be included in the final model, and iii) the stepwise regression procedure, developed from the pre-

vious, to improve the possibility of achieving the best model. For example, in the forward selection procedure, we start with an "empty" model without explanatory variables and we add the variable with the lowest p-value of the F-test for significance of a single variable. The procedure ends when no more variables can be added in the model at a critical significance level (e.g., 10%). The difference with the stepwise regression procedure is that, in the latter, each time a new variable is added to the model, the significance of each of the variables already in the model is re-examined. The backward elimination is a reversed version of the forward selection. Instead of starting with a model without variables, we start with the maximum model and remove the variable with the highest p-value one by one.

These last procedures have some limitations, such as the statistical significance is lost after applying successive tests to choose the added or removed variable in each step, it is not possible to test the number of significance variables in the model (obtaining a p-value) and finally, with these methods is not possible, given a number q, to obtain the "best" q variables. According to this, we propose a procedure that includes two topics: i) selecting the best combination of q variables; and ii) determining the minimum number of covariates to be included in the model. Both topics are explained as follows.

2.1. Selecting the best variables

The first topic of our procedure is, given a number q ($q \leq p$), to select the best combination of q variables. For this purpose, one option is to use the method described in Roca-Pardiñas et al. (2009), which requires all possible models to be considered. When confronted with a large number of variables, however, the computational cost of the procedure can be very high or even prohibitive. In view of this, we use a new method that speeds up the process and is described step-by-step below.

Let $X_{j_1}, \ldots, X_{j_k}$ be a subset of variables of size k ($k \leq q$). We define $IC_{j_1,\ldots,j_k}$ as one possible information criterion (such as AIC, deviance, residual variance, etc.) of the nonparametric model

$$Y = \alpha + m_{j_1}(X_{j_1}) + m_{j_2}(X_{j_2}) + \ldots + m_{j_k}(X_{j_k}) + \varepsilon', \tag{3}$$

where ε' is the zero-mean error.

Here, we use the residual variance obtained by cross-validation. Given a sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ we define $\hat{\sigma}^2_{j_1,\ldots,j_k}$ as the sample variance obtained by cross-validation according to the following expression

$$\hat{\sigma}^2_{j_1,\ldots,j_k} = n^{-1} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i^{(-i)} \right)^2,$$

where $\hat{Y}_i^{(-i)}$ indicates the estimate of Y_i leaving out the i-th element of the sample obtained by fitting the model in (3)[1]. Based on this information criterion, IC, the proposed procedure is an automatic forward stepwise method that, given a number q, selects the q variables $X_{l_1}, \ldots, X_{l_q}$ which minimises the following expression

$$(l_1, l_2, \ldots, l_q) = \underset{\substack{(j_1,\ldots,j_q) \\ 1 \leq j_1 \leq \cdots \leq j_q \leq p}}{\arg\min} \; IC_{j_1,\ldots,j_q}. \tag{4}$$

Step 1: The elements of the vector of indices $(l_1, l_2, \ldots, l_q)$ are selected consecutively in the following manner:

- Firstly, determine the variable of the first position X_{l_1} where

$$l_1 = \underset{\substack{j_1 \\ 1 \leq j_1 \leq p}}{\arg\min} \; IC_{j_1}.$$

  Note that all possible models of one variable must be estimated.

- Fix the first variable obtained previously, X_{l_1} , and obtain the second one, X_{l_2}, with

$$l_2 = \underset{\substack{j_2 \\ 1 \leq j_2 \leq p \\ j_2 \neq l_1}}{\arg\min} \; IC_{l_1,j_2}.$$

[1] In the case of using a generalized additive model, it is useful to introduce the deviance term which behaves like the residual sum of squares of a linear model. Its expression is

$$D = -2 \sum_{i=1}^{n} \{ l(\hat{\mu}_i) - l(y_i) \},$$

where $l(\hat{\mu}_i)$ and $l(y_i)$ are the individual log-likelihood of the proposed model and satured model (including all data).

- Fix X_{l_1} and X_{l_2}, and obtain the third one, X_{l_3}, where

$$l_3 = \underset{\substack{j_3 \\ 1 \leq j_3 \leq p \\ j_3 \notin \{l_1, l_2\}}}{\arg\min} \; IC_{l_1, l_2, j_3}.$$

- Fix $X_{l_1}, X_{l_2}, \ldots, X_{l_{q-1}}$, and repeat the procedure analogously until the q-th variable, X_{l_q}, with

$$l_q = \underset{\substack{j_q \\ 1 \leq j_q \leq p \\ j_q \notin \{l_1, \ldots, l_{q-1}\}}}{\arg\min} \; IC_{l_1, \ldots, j_q}$$

Step 2: Once variables $X_{l_1}, X_{l_2}, \ldots, X_{l_q}$ have been selected, run through positions $j = 1, \ldots, q$ and replace each l_j element as follows, only if the obtained IC is less than the minimum criterion obtained with the previous l_j,

$$l_j = \underset{j_j \notin \{l_1, \ldots, l_{j-1}, l_{j+1}, \ldots, l_q\}}{\arg\min} \; IC_{l_1, \ldots, l_{j-1}, j_j, l_{j+1}, \ldots, l_q}.$$

Step 3: Step 2 is repeated until there is no change in the selected covariates, i.e., the algorithm stops when it has gone through a complete cycle without changing any of the q positions.

As we mentioned, given a number q, the algorithm selects the best model of q variables attending to an information criterion. Any criterion can be used without correcting it taking into account the number of variables. This is possible because the models which are compared have always the same number of variables. Additionally, it should be highlighted that the solution that we obtain from (4) is an approximation of the optimal one. This solution could to be achieved based on searching through all the possible subsets however this procedure supposes a very high computational cost. Therefore, we provide a method that, although it does not reach the optimal solution, could be close to it.

2.2. Testing the number of significant variables

Previously, the best subset of q variables is selected according to an information criterion. However, the question that arises in this procedure is to know the optimal number q. Thus, the second topic in our methodology is to decide the number of covariates that should be included in the model, i.e, determining the number of significant variables.

Accordingly, we propose a procedure to test the null hypothesis of q significant variables in the model versus the alternative in which the model contains more than q variables. Based on the general model

$$Y = m(\mathbf{X}) + \varepsilon \quad \text{where} \quad m(\mathbf{X}) = \alpha + m_1(X_1) + m_2(X_2) + \ldots + m_p(X_p),$$

the following strategy is considered: for a subset of size q, considerations will be given to a test for the null hypothesis

$$H_0\,(q) : \sum_{j=1}^{p} I_{\{m_j \neq 0\}} \leq q$$

versus the general hypothesis

$$H_1 : \sum_{j=1}^{p} I_{\{m_j \neq 0\}} > q.$$

Given a i.i.d. sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^{n}$, with $\mathbf{X} = (X_1, \ldots, X_p)$, to test the above null hypothesis we propose the following strategy:

Step 1. Obtain the best subset of q variables using for this purpose the selection algorithm exposed in Subsection 2.1. For simplicity of notation, the $\mathbf{X}$ vector will be considered as $\mathbf{X} = (X_1, \ldots, X_q, X_{q+1}, \ldots, X_p)$ and the variables selected by the algorithm will be the first q. Note that this is not a constraint, we are just reordering the $\mathbf{X}$. Therefore, the regression function under the null model is

$$m_0(\mathbf{X}) = \alpha + m_1(X_1) + \ldots + m_q(X_q). \tag{5}$$

Step 2. Obtain the nonparametric estimates of the null model, $\hat{m}_0(\mathbf{X}_i)$, compute the residuals as $r_i = Y_i - \hat{m}_0(\mathbf{X}_i)$ and obtain the nonparametric estimates of $g(\mathbf{X}_i)$ according to the model[2]

$$r_i = g(\mathbf{X}_i) + \varepsilon \quad \text{where} \quad g(\mathbf{X}) = \alpha + g_{q+1}(X_{q+1}) + \ldots + g_p(X_p). \tag{6}$$

Finally, we propose the following four test statistics, based on the estimations of g (T_1 and T_2) and on the differences of the residual sum of squares (T_3 and T_4) —closely related to the test introduced by Dette (1999) and by Fan and Jiang (2005)— respectively:

$$T_1 = \sum_{i=1}^{n} |\hat{g}(\mathbf{X}_i)| \qquad \text{and} \qquad T_2 = \sum_{i=1}^{n} \hat{g}(\mathbf{X}_i)^2,$$

$$T_3 = RSS_0 - RSS_1 \qquad \text{and} \qquad T_4 = \frac{RSS_0 - RSS_1}{RSS_1},$$

being $RSS_0 = \sum_{i=1}^{n} \left(Y_i - \hat{m}_0(\mathbf{X}_i)\right)^2$ and $RSS_1 = \sum_{i=1}^{n} \left(Y_i - \hat{m}_0(\mathbf{X}_i) - \hat{g}(\mathbf{X}_i)\right)^2$.

It is important to stress that, if the null hypothesis holds, T —which represents T_1, T_2, T_3 and T_4— should be close to zero. Thus, the test rule for checking $H_0(q)$ with a significance level of α is that the null hypothesis is rejected if T is larger than its $(1-\alpha)$-percentile. To obtain the critical values of T, we apply the wild bootstrap method. The testing procedure consists on the following steps:

Step 1: Obtain T from the sample data as explained above.

Step 2: Obtain the estimates, for $i = 1, \ldots, n$, of $\hat{m}_0(\mathbf{X}_i)$ based on the null model in (5).

Step 3: For $b = 1, \ldots, B$, simulate the bootstrap sample $\left\{\mathbf{X}_i, Y_i^{\bullet b}\right\}_{i=1}^{n}$ with $Y_i^{\bullet b} = \hat{m}_0(\mathbf{X}_i) + \varepsilon_i^{\bullet b}$, with $\varepsilon_i^{\bullet b}$ being

$$\hat{\varepsilon}_i^{\bullet b} = \begin{cases} \hat{\varepsilon}_i \cdot \frac{(1-\sqrt{5})}{2} & \text{with probability } p = \frac{5+\sqrt{5}}{10} \\ \hat{\varepsilon}_i \cdot \frac{(1+\sqrt{5})}{2} & \text{with probability } p = \frac{5-\sqrt{5}}{10} \end{cases}$$

[2] In situations where the number of initial covariates is very high we propose a minor modification of the procedure. To obtain the estimates of g, now we only include one covariate in the model in (6). This unique covariate will be chosen from $X_{q+1}, \ldots, X_p$ applying the selection algorithm exposed in Subsection 2.1.

where $\hat{\varepsilon}_i = Y_i - \hat{m}_0(\mathbf{X}_i)$ are the residuals of the null model, and compute the bootstrap estimates of $T^{\bullet b}$.

The test rule based on T is given by rejecting the null hypothesis if $T > T^{1-\alpha}$, where $T^{1-\alpha}$ is the empirical $(1-\alpha)$-percentile of values $T^{\bullet b}$ $(b = 1, \ldots, B)$.

Applying this test to $q = 1, \ldots, p-1$ could be an important issue in a covariate selection procedure. If $H_0(q)$ is not rejected, only the subset of the covariates $X_{j_1}, \ldots, X_{j_q}$ will be retained, and the remaining variables will be eliminated from the model. In all other cases, the test is repeated with $q+1$ variables until the null hypothesis is not rejected. For example, if $H_0(1)$ is not rejected just one variable should be included into the model. If this hypothesis is rejected it will be required to test $H_0(2)$. If this new hypothesis is again rejected, $H_0(3)$ should be tested and so on until a certain $H_0(q)$ is accepted.

3. Simulation studies

This Section reports the results of two simulation studies conducted both to assess the validity of our method and to compare it against other existing methodologies created to perform automated variable selection or model selection. Accordingly, the validation of the approach relying on the bootstrap-based test is followed by the comparison with the `regsubsets` function from the *R* package `leaps` (Lumley and Miller, 2009), the `step` function (Hastie and Pregibon, 1992; Venables and Ripley, 1997) built into the `stats` package and the Lasso method (Tibshirani, 1996) implemented, for example, in the `glmnet` package (Friedman et al., 2013).

3.1. Simulation 1. Bootstrap-based test

Here, we report the results of a simulation study designed to assess the validity of the bootstrap-based test conducted to determine the number of variables to be included in the model. We focus our attention on situations where there is correlation between covariates.

We consider a vector of 5 covariates, $\mathbf{X} = (X_1, \ldots, X_5)$, and a continuous response, Y, generated in accordance with

$$Y = m(\mathbf{X}) + \varepsilon \qquad \text{being} \qquad m(\mathbf{X}) = \sum_{j=1}^{5} m_j(X_j) \tag{7}$$

with

$$m_j(X_j) = \begin{cases} 2\sin(2\pi X_j) & \text{if} \quad j \in \{1,2\} \\ 2\,a\sin(2\pi X_j) & \text{if} \quad j \in \{3,4,5\} \end{cases}$$

and ε being the error distributed in accordance with a $N(0, \sigma(\mathbf{X}))$ with $\sigma(\mathbf{X}) = 0.5 + 0.05\,|m(\mathbf{X})|$. The explanatory covariates were generated with the following expression: $X_j = (U_j + tU)/(1+t)$, where $U_1, \ldots, U_5, U$ are i.i.d. random variables from uniform distribution $[0, 1]$. To check the performance of the test for different levels of correlation between covariates, a constant t is included. The used values (corresponding correlation shown in brackets) are $t = 0$ (0.0), $t = 1$ (0.5) and $t = 2$ (0.8).

One thousand independent samples $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ were generated from the model (7) for the purpose of testing the following null hypothesis

$$H_0(2) : \sum_{j=1}^{5} I_{\{m_j \neq 0\}} \leq 2 \qquad \text{versus} \qquad H_1 : \sum_{j=1}^{5} I_{\{m_j \neq 0\}} > 2$$

Note that the constant a governs the number of non-informative covariates. While the value $a = 0$ corresponds to the null hypothesis $H_0(2)$, with three predictors (X_3, X_4, X_5) being uninformative, as the value of a rises, so do the effects of X_3, X_4 and X_5. To test the above hypothesis, we use the bootstrap procedure described in Section 2, specifically using $B = 1000$ bootstrap samples to calculate type I error and $B = 500$ bootstrap samples to calculate the power under the alternative. Both type I error and power are calculated on the basis of 1000 simulations runs.

Estimated type I errors registered by the tests at different significance levels, t values and sample sizes are displayed in Table 1. All the test statistics perform reasonably well, with the level coming relatively close to the nominal size, specially with large sample sizes. I should be note, however, that they seems to reject the null hypothesis more often than they should when H_0 is true. The test based on T_4 seems to produce better approximations of the nominal level than the others.

We also study power performance for the alternatives as a function of a. Power results for the test with different sample sizes and taking different t values into account are shown in Figure 1. As expected, the probability of rejection rises with the increase in the constant a and sample size. Additionally, in Figure 2 it can be observed that the power of the test depends on

the correlation between predictors, so that power decreases as correlation increases (high values of t). Additionally, as can be seen in Table 2 and Figure 1, the T_1 and T_2 yield slightly upper power than T_3 and T_4, situation more remarkable when the correlation increases.

Table 1: Estimated type I error (in %) for $t = 0$, 1, 2, for different sample sizes and nominal levels (1, 5, 10, 15 and 20%).

		t=0					t=1					t=2				
n	Test	1%	5%	10%	15%	20%	1%	5%	10%	15%	20%	1%	5%	10%	15%	20%
200	T_1	1.1	7.0	13.1	20.2	25.7	1.3	6.3	10.6	17.4	23.3	1.3	7.0	13.1	19.8	26.7
	T_2	0.8	6.9	13.1	19.2	26.0	1.1	5.8	11.4	17.3	24.0	1.0	6.6	12.9	19.9	25.3
	T_3	0.8	7.6	12.6	20.3	25.3	1.0	7.7	12.9	17.7	24.0	0.9	6.3	12.8	18.8	24.2
	T_4	0.7	5.2	10.4	16.7	21.7	0.8	6.6	11.2	14.9	20.0	0.7	5.0	9.7	14.9	19.9
500	T_1	0.7	6.3	11.5	17.0	22.6	0.6	5.0	9.2	13.8	19.5	1.1	5.8	11.1	17.9	22.3
	T_2	0.6	5.7	11.0	16.9	22.0	0.3	4.8	9.1	15.1	19.9	1.0	6.2	11.4	17.3	22.8
	T_3	0.9	6.3	11.4	18.1	23.2	0.3	4.0	8.3	14.7	19.4	0.9	6.9	11.7	17.8	23.1
	T_4	1.1	5.9	10.6	16.4	21.5	0.3	3.7	8.1	13.6	18.4	1.0	5.9	10.4	16.5	21.5
1000	T_1	0.9	6.4	11.8	16.9	22.8	0.9	6.2	10.7	16.6	21.6	1.1	5.4	10.8	16.9	23.7
	T_2	0.9	6.9	12.0	17.1	22.7	1.2	5.6	10.7	16.6	21.6	1.3	6.1	10.9	17.2	23.6
	T_3	0.8	6.1	11.0	15.8	20.9	1.1	5.1	9.2	15.5	20.3	1.5	5.7	10.9	19.3	23.7
	T_4	0.7	5.8	10.5	15.4	20.0	1.1	5.0	8.8	15.3	19.5	1.4	5.7	10.5	18.3	22.8

In view of the results shown above, our procedure could be said to determine the number of variables correctly. At this point, it is also important to evaluate if the selection of variables performs reasonably well. We therefore apply the step-by-step procedure proposed in Section 2.1 to select the best subset of variables of size $q = 2$. The data were generated in accordance with the above scenario, with the a value being kept at 0. The results of this selection, based on 1000 simulation runs with sample sizes of $n = 200$, 500 and 1000, and t values of t 0, 1 and 2, are successful, with the right variables (X_1 and X_2) being selected 100% of the times in all cases.

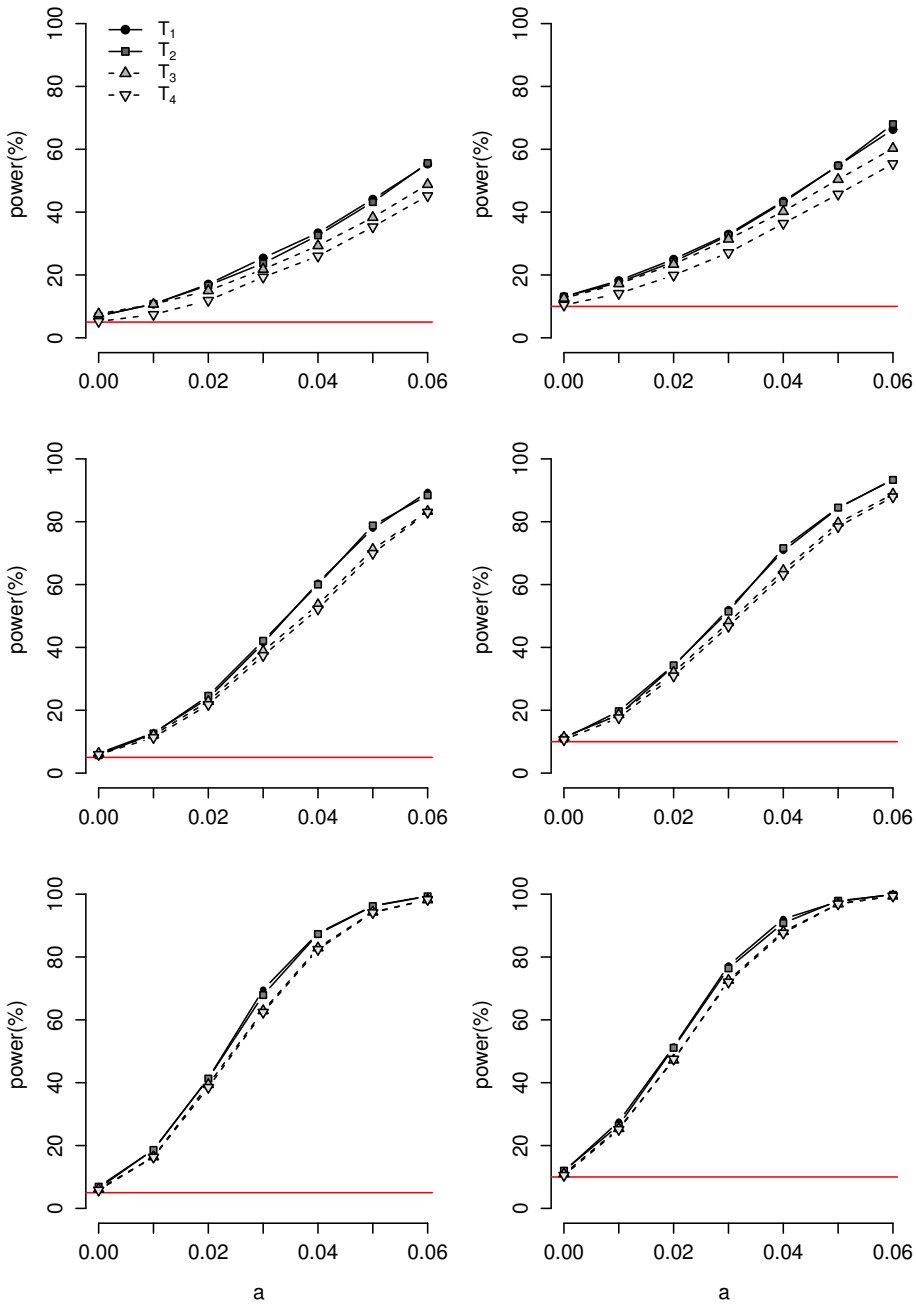

Figure 1: Percentage of rejections for T_1, T_2, T_3 and T_4 on a increasing for nominal levels of 5% and 10% (left and right plot, respectively), keeping $t = 0$. Upper panel: rejections for sample size $n = 200$. Middle panel: rejections for sample size $n = 500$. Lower panel: rejections for sample size $n = 1000$.

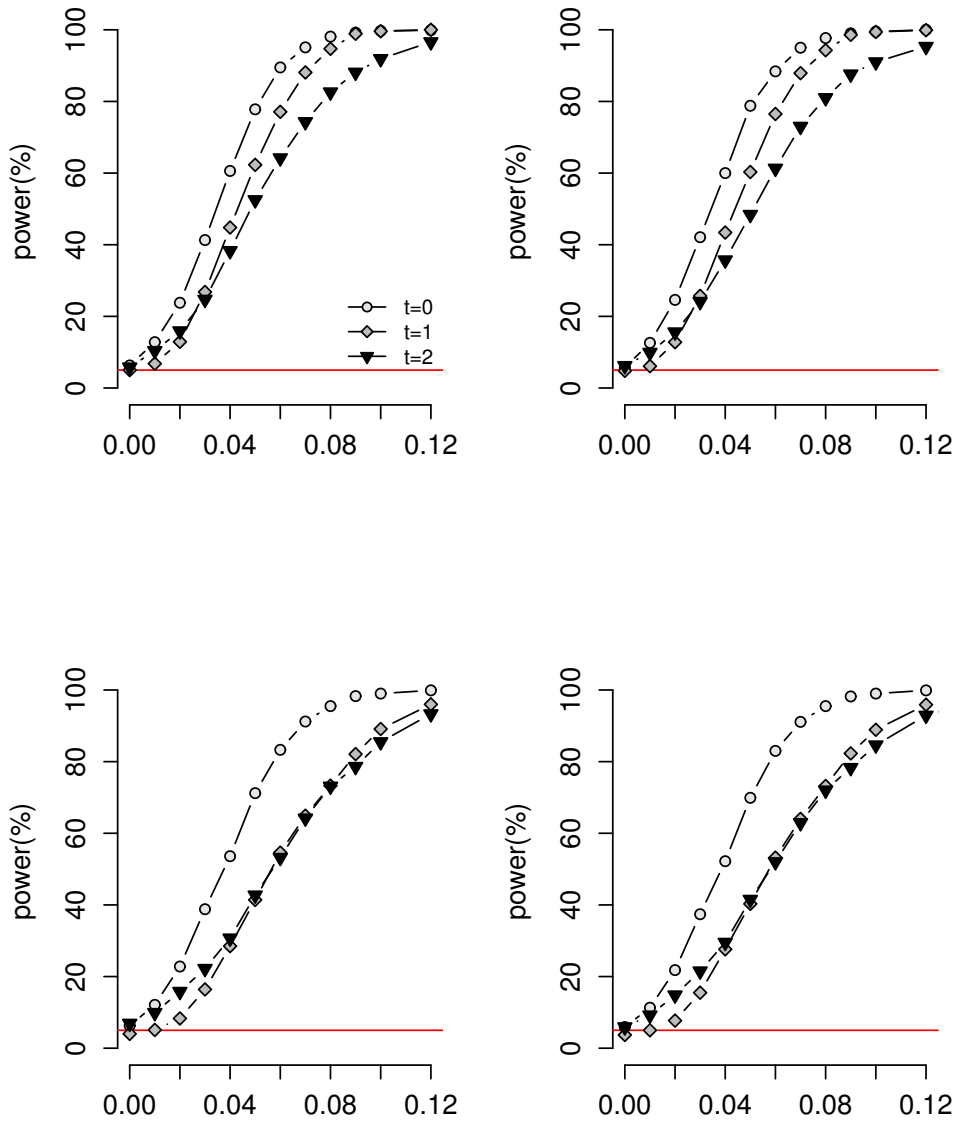

Figure 2: Percentage of rejections on a increasing for the tests based on T_1 (left upper panel), T_2 (right upper panel), T_3 (left lower panel) and T_4 (right lower panel) for different correlation values ($t = 0$, $t = 1$ and $t = 2$), keeping the nominal level at 5% and for a sample size of $n = 500$.

Table 2: Percentage of rejections (in %) for all the test statistics with $t = 0$, 1, 2, for different a values, sample sizes and nominal levels (1, 5, 10, and 20%).

			t=0				t=1				t=2			
a	n	Test	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%
0.02	200	T_1	3.7	17.1	25.0	37.6	2.2	9.8	17.5	30.5	2.4	12.6	21.7	36.1
		T_2	3.5	16.7	24.2	37.8	2.1	9.5	17.2	30.7	2.2	12.5	21.0	35.5
		T_3	2.8	15.0	23.4	38.4	2.2	9.6	14.4	29.6	2.0	10.1	20.0	34.8
		T_4	2.9	11.9	19.9	32.8	1.4	7.7	12.3	24.9	1.8	8.2	15.5	28.8
	500	T_1	9.8	23.8	34.0	47.9	3.6	12.9	21.5	35.2	4.2	15.9	25.5	43.4
		T_2	8.3	24.6	34.3	48.3	3.8	12.8	20.6	35.7	3.5	15.6	23.1	41.1
		T_3	7.7	22.8	32.1	46.9	1.6	8.3	15.6	29.9	2.9	15.8	23.7	40.0
		T_4	7.3	21.8	30.9	45.2	1.3	7.7	14.4	28.7	3.1	14.8	22.1	37.2
	1000	T_1	19.8	41.4	51.5	66.5	10.5	25.6	35.3	49.4	8.1	27.9	37.4	55.5
		T_2	18.3	41.3	51.1	65.2	10.2	25.2	34.4	48.8	7.6	26.7	36.3	54.5
		T_3	15.7	39.4	47.2	63.6	4.7	19.5	26.4	41.4	7.0	24.5	35.1	51.3
		T_4	15.4	38.5	47.5	62.9	5.2	19.8	26.3	41.3	6.9	23.7	34.8	50.7
0.04	200	T_1	14.2	33.4	43.5	59.1	7.8	23.9	33.1	48.2	5.7	23.5	33.7	50.2
		T_4	12.9	32.6	43.1	60.1	7.0	23.3	32.3	47.3	4.9	21.6	32.3	50.1
		T_3	10.2	29.3	40.2	55.8	5.1	17.7	26.7	40.5	4.4	19.3	30.2	46.3
		T_4	9.5	26.0	36.4	51.1	4.1	14.2	22.5	35.4	4.1	16.6	25.3	41.6
	500	T_1	35.1	60.6	70.7	81.8	19.9	44.8	55.3	68.7	13.5	38.3	50.8	68.7
		T_2	35.2	60.0	71.6	82.9	19.4	43.4	53.6	67.4	11.8	35.7	48.1	66.4
		T_3	28.7	53.6	64.5	78.1	9.3	28.5	39.0	52.4	9.9	30.7	44.5	61.6
		T_4	29.6	52.2	63.2	76.7	9.2	27.6	37.3	50.7	10.5	29.5	41.2	60.3
	1000	T_1	69.6	87.6	92.1	96.2	50.9	75.0	82.3	90.2	33.8	62.9	73.7	86.0
		T_2	68.5	87.3	90.8	96.0	49.6	73.6	82.1	89.9	30.6	62.2	72.2	82.9
		T_3	61.6	82.9	88.1	94.8	28.0	50.9	60.9	71.9	26.5	54.7	64.8	79.5
		T_4	62.1	82.4	87.6	94.3	28.6	51.1	60.2	71.6	26.8	53.7	64.3	79.2
0.06	200	T_1	29.7	55.3	66.3	79.3	20.3	43.4	54.3	67.8	12.3	37.1	48.8	64.7
		T_2	27.5	55.6	68.0	80.3	18.4	42.1	53.4	67.2	10.5	34.5	46.4	62.8
		T_3	22.9	48.8	60.3	75.7	10.3	28.6	38.5	52.7	8.0	30.5	42.5	60.0
		T_4	22.1	45.2	55.4	70.8	9.5	25.8	34.8	48.4	7.9	25.9	37.0	54.0
	500	T_1	71.2	89.5	93.5	96.9	55.0	77.1	84.9	91.0	32.7	64.2	73.3	85.2
		T_2	70.2	88.4	93.3	96.8	52.3	76.5	83.5	90.3	29.9	61.3	72.6	84.7
		T_3	61.0	83.3	88.8	95.0	29.0	54.6	63.3	73.6	24.5	53.2	65.3	79.9
		T_4	61.0	83.0	87.9	94.3	28.3	53.2	62.9	73.1	24.4	52.0	63.7	78.7
	1000	T_1	96.2	99.4	100	100	90.5	97.4	98.4	99.6	67.1	88.7	93.5	96.7
		T_2	95.8	99.3	99.8	100	90.0	96.8	98.5	99.7	65.8	86.5	91.5	96.4
		T_3	93.1	98.2	99.4	99.8	61.5	80.5	85.7	91.1	56.4	81.9	88.3	93.7
		T_4	93.1	98.3	99.4	99.8	61.5	80.7	85.5	91.0	56.2	80.6	87.9	93.7

3.2. Simulation 2. Comparing methodologies

To date, several procedures that carry out automatic variable selection have been reported in the literature. We therefore want to compare the proposed methodology with some of these existing methods. We choose the *R* `regsubsets` function of the `leaps` package, which selects the best variables for each subset of size q without determining the number of variables that users have to include in the model; the `step` function which selects a formula-based model using the AIC; and the package `glmnet` in which the Lasso method is implemented.

We start by showing the results of the simulation study which compares our procedure (denoted here as `selection` function) to the above-mentioned `regsubsets` function. This function is based on all subsets or, in other words, exhaustive variable selection using the AIC. The method identifies the best subsets of linear predictors using a branch-and-bound algorithm (Miller, 2002).

As mentioned previously, our procedure is able to select the best predictor in different regression contexts (parametric or nonparametric). However, as the `regsubsets` function is only suitable for linear frameworks, in order to evaluate the behaviour of both methods different scenarios to one used in the previous Subsection are used. Two new scenarios are thus considered, namely: (a) a linear scenario in which the explanatory variable depends on two covariates; and, (b) another linear scenario, in which three informative variables appear. These scenarios were generated according with the model in (7) with

$$\text{(a)} \quad m(\mathbf{X}) = X_1 + 2\,a\,X_2 + 3\,a\,X_3 + 0.5\,X_4 + 2\,a\,X_5,$$

$$\text{(b)} \quad m(\mathbf{X}) = X_1 + 2\,a\,X_2 + 3\,a\,X_3 + 0.5\,X_4 + 2\,X_5.$$

In both cases, ε is the error distributed in accordance with a $N(0, \sigma(\mathbf{X}))$ with $\sigma(\mathbf{X}) = 0.75 + 0.05\,|m(\mathbf{X})|$. The vector of covariates $\mathbf{X}$ was generated as in the previous Subsection, while the a value was kept at zero.

The results of the selection for a given subset of size q, $q = 2$ for scenario (a) and $q = 3$ for scenario (b), based on 1000 simulated samples with sample sizes of $n = 200$, 500 and 1000, and t values of 0, 1 and 2, are the same for both methods (Table 3, only shown $n = 200$). Performance is good, in that the proportion of mistakes decreases as sample size increases. This proportion rises as correlation (value of t) and the number of selected variables increases.

Table 3: Percentage of mistakes selecting the best subset of size $q = 2$ —scenario (a)— and $q = 3$ —scenario (b)— for `regsubset` and `selection` functions based on 1000 simulation runs for $n = 200$ and for different t values.

		Scenario (a)		Scenario (b)	
n	t	`regsubset`	`selection`	`regsubset`	`selection`
200	0	0.0	0.0	0.0	0.0
	1	0.1	0.1	0.1	0.1
	2	2.4	2.4	3.5	3.5

For instance, with $n = 200$, $t = 1$ and selecting 3 variables, the methods are only wrong 0.1% of the times. With $t = 0$ none of the methods make mistakes in selection, and that for $n = 500$ or more, they are equally successful, even with different values of t.

The reviewed literature features other methodologies for jointly determining the number and choice of variables. One example of this is the model-selection oriented function `step` (Hastie and Pregibon, 1992; Venables and Ripley, 1997) which uses a stepwise selection procedure based on AIC. We try to assess its performance by means of a simulation study. Initially, to compare the methodologies in the same framework, we used the scenario proposed in Subsection 3.1, applying the function to a model of class `gam`. However, due to its poor performance (in each case we obtained a new model with four variables), we decided to replace it with a linear scenario —specifically scenario (a) described above— in which the `step` function could be applied to a recommended model of class `lm`.

Table 4 shows the models selected by this function, based on 1000 simulation runs for different sample sizes and $t = 0$. The method performs correctly, selecting the right variables (X_1 and X_4), 59% of the times ($n = 200$), 58.9% ($n = 500$) and 59.4% ($n = 1000$). Note that, if we would want to compare it with our procedure, testing the null hypothesis of a model with two variables, type I error would be quite high (41% with $n = 200$, 41.1% with $n = 500$ and 40.6% with n =1000). Additionally, it should also be pointed out that 90% of the mistakes are due to selection of one more variable.

Table 4: Selected models with their percentages, by the `step` function based on 1000 simulation runs for different sample sizes and $t = 0$.

Model	n: 200	500	1000
1, 4	59.0	58.9	59.4
1, 2, 4	12.6	11.8	9.6
1, 3, 4	10.9	11.7	12.7
1, 4, 5	10.7	11.5	12.2
1, 2, 3, 4	2.0	1.9	2.1
1, 2, 4, 5	2.3	2.0	1.6
1, 3, 4, 5	1.9	1.5	1.9
1, 2, 3, 4, 5	0.6	0.7	0.5

Finally, we compare our procedure with the Lasso method, applying for this purpose the functions implemented in the `glmnet` package (Friedman et al., 2013). The Lasso is a shrinkage method that minimises the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant (λ). Because the nature of this constraint it tends to produce some coefficients that are exactly zero and hence gives interpretable models (Tibshirani, 1996). The algorithm computes an entire path of solutions (in λ) for any particular model, leaving the user to select a particular solution. However, the authors of this methodology propose to consider two posible λ. The first option is the minimum lambda (λmin), which minimises an estimate of prediction error based on tenfold cross-validation[3]. The second is the lambda obtained with the "one-standard error" rule (λ1se), which applies this rule also with cross-validation to obtain the largest value of λ such that error is within one standard error of the minimum.

In relation with this method, we focus our attention in assesing only the validity of the procedure selecting correctly the number of variable that have to be included in the model. To this end, one thousand independent samples

[3]The original sample is randomly partitioned into ten equal size subsamples. Of the ten subsamples, one subsample is retained as the validation data for testing the model, and the remaining nine subsamples are used as training data. The cross-validation process is then repeated ten times (the folds), with each of the ten subsamples used exactly once as the validation data.

Table 5: Proportion of mistakes selecting the correct number of variables ($q = 2$) based on the use of the minimum lambda (λmin) and on "one-standard error" rule (λ1se) for different sample sizes.

	$t = 0$		$t = 1$		$t = 2$	
n	λmin	λ1se	λmin	λ1se	λmin	λ1se
200	75.6	3.7	81.1	17.6	82.7	42.2
500	74.2	1.1	78.8	9.5	85.1	34.9
1000	76.4	0.2	78.4	5.0	83.3	24.3

were generated from the scenario (a). The proportion of mistakes selecting the correct number of variables ($q = 2$) based on the use of the minimum lambda (λmin) and on "one-standard error" rule (λ1se) are displayed in Table 5. The performance using the minimum lambda is unsatisfactory, with a proportion of mistakes close to 75% even with $t = 0$. Insofar as the use of the other lambda (λ1se), the percentage of mistakes decreases as the sample size grows and it increases as the correlation rises (value of t). If we would want to compare these last results to ours, testing $H_0(2)$ of a model with two variables, type I error would be quite high for $t = 1$ and $t = 2$.

4. `FWDselect` in practice

This Section introduces an overview of how the package is structured. `FWDselect` is a shortcut for "Forward selection" and this is its major functionality: to provide a forward stepwise-based selection procedure. This software helps the user select relevant variables and evaluate how many of these need to be included in a regression model. In addition, it enables both numerical and graphical outputs to be displayed.

Our package includes several functions that enable users to select the variables to be included in linear models, generalized linear models or generalized additive models. The functions within `FWDselect` are briefly described in Table 6.

Users can obtain the best combinations of q variables by means of the main function which is `selection`. Additionally, if one wants to obtain the results for more than one subset size, it is possible to apply the `qselection` function, which returns a summary table showing the different subsets, selected variables and information criterion values. The object obtained with

Table 6: Summary of functions in the `FWDselect` package.

Function	Description
`selection`	Main function for selecting a subset of q variables. Note that the selection procedure can be used with `lm`, `glm` or `gam` functions.
`print.selection`	Method of the generic print function for `selection` objects, which returns a short summary.
`qselection`	Function that enables users to obtain the selected variables for more than one size of subset. Returns a table showing the chosen covariates to be introduced into the models and their information criteria.
`plot.qselection`	Visualisation of `qselection` objects. It plots the cross-validation information criterion for several subsets with size q chosen by users.
`test`	Function that applies a bootstrap based test for covariate selection. It helps determine the number of variables to be included in the model.

this last function is the argument required for `plot`, which provides a graphical output. Finally, to determine the number of variables that should be introduced in the model, only the `test` function needs to be applied. Table 7 provides a summary of the arguments of the `selection`, `qselection` and `test` functions.

4.1. An example with pollution data

The usage of this package is tested on the prediction of atmospheric SO_2 pollution incidents. Combustion of fuel oil or coal releases sulphur dioxide into the atmosphere in different quantities. Current Spanish legislation governing environmetrical pollution controls the vicinity of potential point sources of pollution, such as coal-fired power stations. It places a limit on the mean of 24 successive determinations of SO_2 concentration taken at 5-minute intervals. An emission episode is said to occur when the series of bi-hourly means of SO_2 is greater than a specific level, r. In this framework, it is of interest for a plant, both economically and environmentally, to be able to predict, when the legal limit will be exceeded with sufficient time for effective countermeasures to be taken.

In previous studies (García-Jurado et al., 1995; Prada-Sánchez et al., 2000; Prada-Sánchez and Febrero-Bande, 1997; Roca-Pardiñas et al., 2004),

Table 7: Summary of `selection`, `qselection` and `test` functions.

	`selection()` arguments
`x`	A data frame containing all the covariates.
`y`	A vector with the response values.
`q`	An integer specifying the size of the subset of variables to be selected.
`criterion`	The cross-validation-based information criterion to be used. Default is the deviance. Other functions provided are the coefficient of determination ("`R2`") and residual variance ("`variance`").
`method`	A character string specifying which regression method is used, "`lm`" (linear model), "`glm`" (generalized linear model) or "`gam`" (generalized additive model).
`family`	This is a family object specifying the distribution and link to use in fitting.
`seconds`	A logical value. If TRUE then, rather than returning the single best model only, the function returns a few of the best models.
`nmodels`	Number of secondary models to be returned.
	`qselection()` arguments
`x`	A data frame containing all the covariates.
`y`	A vector with the response values.
`qvector`	A vector with more than one variable-subset size to be selected.
`criterion`	The cross-validation-based information criterion to be used. Default is the deviance. Other functions provided are the coefficient of determination ("`R2`") and residual variance ("`variance`").
`method`	A character string specifying which regression method is used, "`lm`" (linear model), "`glm`" (generalized linear model) or "`gam`" (generalized additive model).
`family`	This is a family object specifying the distribution and link to use in fitting.
	`test()` arguments
`x`	A data frame containing all the covariates.
`y`	A vector with the response values.
`method`	A character string specifying which regression method is used, "`lm`" (linear model), "`glm`" (generalized linear model) or "`gam`" (generalized additive model).
`family`	This is a family object specifying the distribution and link to use in fitting.
`nboot`	Number of bootstrap repeats.
`speedup`	A logical value. If `TRUE` (default), the testing procedure is accelerated by a minor change in the statistic.
`unique`	A logical value. If `TRUE`, the test is performed only for one null hypothesis, given by the argument `num.h0`.
`num.h0`	If `unique` is `TRUE`, `num.h0` is the integer number q of $H_0(q)$ to be tested.

semiparametric, partially linear models and generalised additive models with unknown link functions were applied to the prediction of atmospheric SO_2 pollution incidents in the vicinity of a coal/oil-fired power station. Here, we present a new approach to this problem, whereby we try to predict a new emission episode, focusing our attention on the importance of ascertaining the best combinations of time instants for the purpose of obtaining the best prediction. Bearing this in mind, the selection of the optimal subset of variables could be a good approach to this issue.

Let t be the present time, and X_t the value obtained by the series of bi-hourly means for SO_2 at instant t (5-minute temporal instants). Setting $r = 150$ μg/m^{3}N as the maximum value permitted for the SO_2 concentration, and half-an-hour (6 instants) as the prediction horizon, it is of interest to predict $Y = X_{t+6}$, with the best vector of $X_l = (X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-17})$. Note that one of the problems that arises is to decide which temporal instants $(X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-17})$ are relevant for prediction purposes, since inclusion of all the times X_l may well degrade the overall performance of the prediction model. Based on this, we demonstrate the package capabilities using these data. An excerpt of the data frame included in the package is shown below:

```
R> library(FWDselect)
R> data(pollution)
R> head(pollution)
   In17  In16  In15  In14  In13  In12  In11  In10   In9   In8
1  3.02  3.01  3.01  3.01  3.01  3.03  3.03  3.03  3.03  3.03
2 16.49 16.55 16.42 16.35 16.56 16.75 16.74 16.72 16.63 16.53
3  4.78  4.56  4.48  4.46  4.38  4.29  4.34  4.85  5.75  7.17
4  5.30  5.29  5.28  5.23  5.14  4.92  4.73  4.27  3.96  3.67
5 68.83 63.76 59.14 51.63 42.21 34.04 30.07 26.70 24.28 22.90
6  9.78  9.62  9.46  9.43  9.37  9.27  9.07  9.22  9.21  9.11

    In7   In6   In5   In4   In3   In2   In1   In0   InY
1  3.03  3.03  3.03  3.03  3.03  3.03  3.03  3.03 10.78
2 16.32 16.08 15.77 15.47 14.81 14.30 13.70 13.35 10.65
3  8.39  9.56 10.36 10.47 10.43 10.42 10.44 10.21 10.23
4  3.47  3.23  3.09  3.04  3.01  3.00  3.00  3.00  3.00
5 22.08 20.64 17.28 13.30  9.58  6.92  5.38  4.77  4.52
6  9.00  8.92  9.06  9.01  8.89  8.67  8.47  8.42  7.92
```

The variables from `In17` to `In0` correspond to the registered values of SO_2 at a specific temporal instant. `In0` denotes the zero instant (X_t), `In1` corresponds to the 5-min temporal instant before (X_{t-1}), `In2` is the 10-min temporal instant before (X_{t-2}), and so on until the last variable. The last column of the data-frame (`InY`) refers to the response variable, $Y = X_{t+6}$, the temporal instant that we wish to predict. For this purpose, we propose the underlying generalised additive model

$$Y = m_0(X_t) + m_1(X_{t-1}) + \ldots + m_{17}(X_{t-17}) + \varepsilon \tag{8}$$

where m_j, with $j = 0, \ldots, t-17$, are smooth and unknown functions and ε is the error which is assumed to have mean zero. To estimate the model in (8), `FWDselect` allows penalised regression splines, implemented in the `mgcv` library (Wood, 2003, 2004, 2011).

It may often be of interest to determine the best subset of variables of size q needed to predict the response. The question that naturally arises in this application is, what is the best temporal instant for predicting an emission episode. This is easy to ascertain with the function `selection`

```
R> x=pollution[,-19]
R> y=pollution[,19]
R> obj1=selection(x,y,q=1,method="gam",
+ criterion="deviance")
R> obj1

***************************************************
Best subset of size q = 1 : In0

Information Criterion Value - deviance : 421847.9
***************************************************
```

Additionally, if the selected variables for more than one subset size wish to be known, this package contains the `qselection` function, which returns a table for the different subsets with the chosen variables and their information criteria, `criterion=c(`"`R2`", "`deviance`", "`variance`").

```
R> obj2=qselection(x,y,qvector=c(1:7),method="gam",
+   criterion="deviance")
[1] "Selecting subset of size 1 ..."
[1] "Selecting subset of size 2 ..."
[1] "Selecting subset of size 3 ..."
[1] "Selecting subset of size 4 ..."
[1] "Selecting subset of size 5 ..."
[1] "Selecting subset of size 6 ..."
[1] "Selecting subset of size 7 ..."

R> obj2
  q  deviance                          selection
1 1 421847.87                                In0
2 2 192723.93                           In0, In2
3 3  212786.1                      In0, In2, In1
4 4 249584.21                 In0, In3, In1, In5
5 5 435614.94            In0, In3, In1, In7, In6
6 6 512708.69      In0, In3, In1, In5, In6, In7
7 7 1315273.3 In0, In2, In1, In5, In6, In8, In4
```

The above function output is a useful display that greatly helps determine the most relevant variables. A plot of this object can easily be obtained by using the following input command:

```
R> plot(obj2)
```

Figure 3 shows the deviance values (obtained by cross-validation) corresponding to the different subsets. In each subset, `q` represents the number of temporal instants included in the model. These models appear in Table 8. Note, however, that only the results until subset of size $q = 7$ are shown because, from this size onwards, the rest of the obtained models considerably worse results.

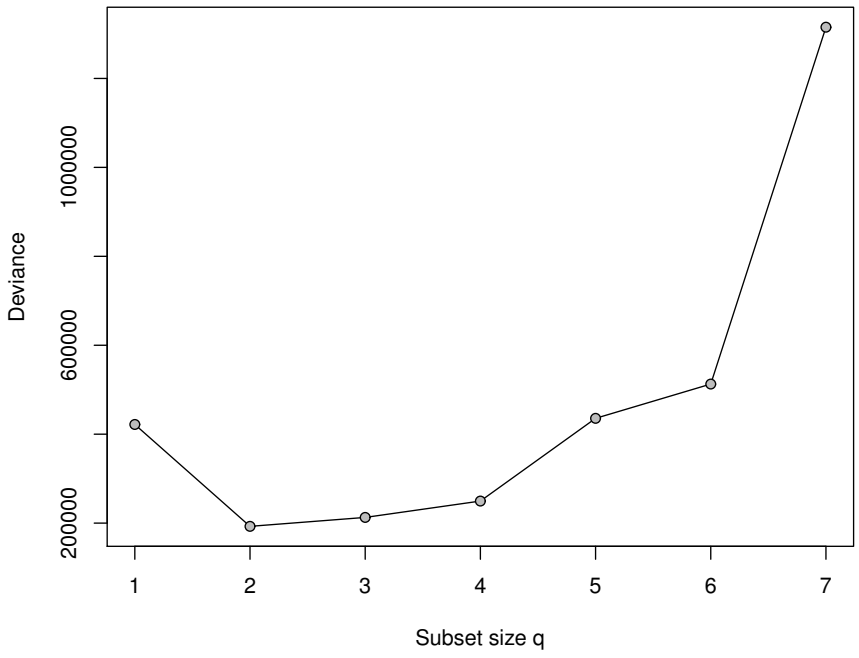

Figure 3: For each subset of size q, cross-validation deviance obtained by the best model.

Table 8: Deviance obtained with each selected model of size q, with $t, 1, \ldots, 17$ being temporal instants $(X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-17})$.

q	Deviance	t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	421847.9	x																	
2	192723.9	x		x															
3	212786.1	x	x	x															
4	249584.2	x	x		x		x												
5	435614.9	x	x		x			x	x										
6	512708.7	x	x		x		x	x	x										
7	1315273.0	x	x	x		x	x	x		x									

The performance of the proposed predictors was then evaluated in a real pollution incident. The corresponding data are found in the `episodeS02` data set, also included in this package. The corresponding data frame is illustrated as follows:

```
R> data(episode)
R> head(episode)
  In17 In16 In15 In14 In13 In12 In11 In10  In9  In8  In7  In6
1 3.02 3.02 3.03 3.10 3.10 3.10 3.10 3.22 3.27 3.33 3.36 3.38
2 3.02 3.03 3.10 3.10 3.10 3.10 3.22 3.27 3.33 3.36 3.38 3.47
3 3.03 3.10 3.10 3.10 3.10 3.22 3.27 3.33 3.36 3.38 3.47 3.50

   In5  In4  In3  In2  In1  In0  InY  time
1 3.47  3.50 3.56 3.61 4.28 4.60 5.45 00:00
2 3.50  3.56 3.61 4.28 4.60 4.68 6.20 00:05
3 3.56  3.61 4.28 4.60 4.68 4.78 6.85 00:10
```

The course of the incident is depicted in Figure 4. Temporal instants are plotted on the horizontal axis and the real 2-hour mean SO_2 concentration that we seek to predict ($Y = X_{t+6}$) is represented by a grey line. Corresponding the predictions to the episode obtained by applying the different models achieved with the `qselection` function are shown in the same figure. These predictions are obtained using the `predict.gam` function of the `mgcv` package. The prediction obtained with the inclusion of just one variable in the model, X_t, is far from the optimum. However, the addition of one more variable, X_{t-2}, resulted in a remarkable increase in the model predictive capability. It makes possible for predictions close to real values to be obtained. Lastly, it can be seen that the incorporation of one more variable or temporal instant (X_{t-1}) in the model does not produce any improvement in pollution-incident prediction. Numerically speaking, the same results can be observed by taking into account the Mean Square Error for each model (Table 9).

The question that now arises is what is the minimum number of variables that must be used in order to obtain the best prediction. It is possible to deduce that there is an optimal intermediate point between the number of variables that enters the model (preferably low) and the deviance value (preferably also low). To find this point, the described test for the null hypothesis $H_0(q)$ is applied for each size, q. To this end, the following call is:

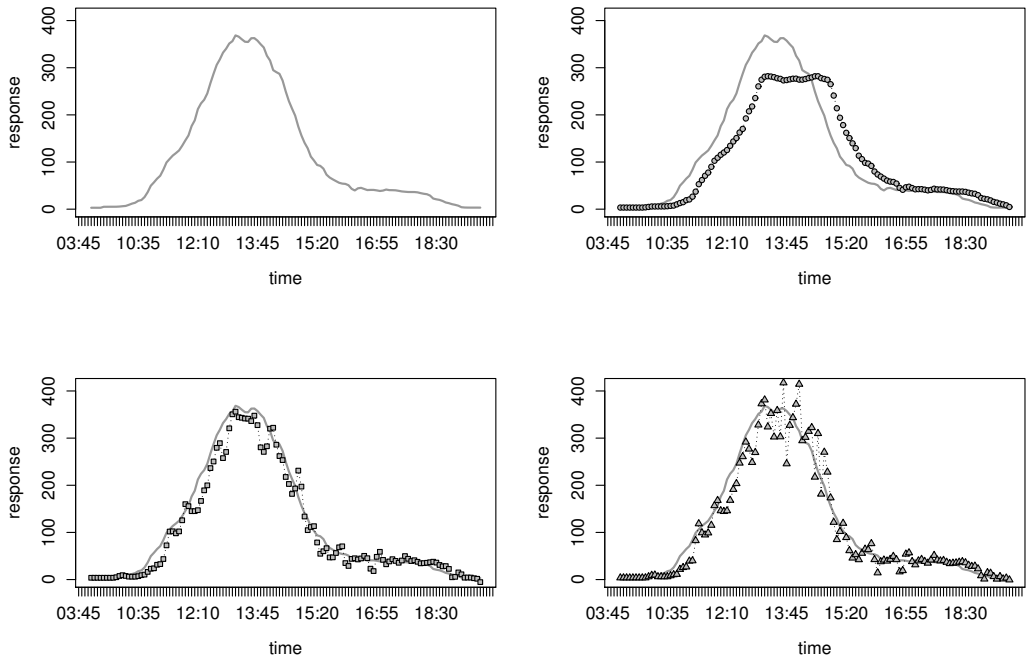

Figure 4: Example of an SO_2 pollution incident that occurred on 4 July 2003. Temporal instants are shown on the horizontal axis. The grey line represents the known response of SO_2 levels in $\mu g/m^3 N$. Estimation of SO_2 levels with one, two and three covariates are represented by circles, squares and triangles respectively.

Table 9: Mean Square Error of the selected models.

Model	MSE
$Y = X_t$	1 682.14
$Y = X_t + X_{t-2}$	366.44
$Y = X_t + X_{t-2} + X_{t-1}$	556.49

```
> test(x,y,method="gam",speedup=FALSE)
[1] "Processing IC bootstrap for H_0 ( 1 )..."
[1] "Processing IC bootstrap for H_0 ( 2 )..."

***********************************
  Hypothesis Statistic pvalue Decision
1    H_0 (1)  11145.41      0 Rejected
2    H_0 (2)   4516.66    0.7 Accepted
```

The deduction to be drawn from these results obtained is that, for a 5% significance level, the null hypothesis is rejected with $q = 1$ and accepted thereafter. From these results, it can be concluded that the best temporal instants for prediction of an emission episode would be X_t and X_{t-2}.

5. Conclusions

Throughout the paper, we have displayed the implementation in R of a new algorithm for the problem of variable selection in a regression framework. The `FWDselect` package provides R users with a simple method for ascertaining the relevant variables for prediction purposes and how many of these should be included in the model. The proposed method is a new forward stepwise-based selection procedure that selects a model containing a subset of variables according to an information criterion, and also takes into account the computational cost. Bootstrap techniques have been used to determine the minimum number of variables needed to obtain an appropriate prediction.

In some situations, several statistically equivalent optimal models of size q may exist. In such cases, `FWDselect` allows the user to visualise those models and select the one that most interesting one. In addition, the software provides the user with a way of easily obtaining the best subset of variables using different types of data in different contexts, by applying the `lm`, `glm` and `gam` functions already implemented in R. The use of these classical R functions nevertheless entails a high computational cost. Hence, a further interesting extension would be the implementation of this package using Fortran (Gehrke, 1995) as the programming language. R users could profit from this advantage in a future version of this package.

The goal of this package is to afford the research community with a new tool in the selection framework. Nevertheless, our intention is not to replace other currently available approaches but rather to provide a practical solution to this challenge.

Acknowledgements

The authors gratefully acknowledge the financial support from the projects MTM2008-03129 and MTM2011-23204 of the Spanish Ministry of Science and Innovation (FEDER support included) and from 10 PXIB 300 068 PR of the Xunta de Galicia.

References

Buja, A., Hastie, T., Tibshirani, R., 1989. Linear smoothers and additive models. The Annals of Statistics 17, 453–510.

Burnham, K., Anderson, D., 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer.

Calcagno, V., 2012. `glmulti`: Model selection and multimodel inference made easy. *R* package version 1.0.4.

Calcagno, V., de Mazancourt, C., 2010. `glmulti`: An *R* Package for easy automated model selection with (Generalized) Linear Models. Journal of Statistical Software 34, 1–29.

Dette, H., 1999. A consistent test for the functional form of a regression based on a difference of variance estimators. The Annals of Statistics 27, 1012–1040.

Fan, J., Jiang, J., 2005. Nonparametric inferences for additive models. Journal of the American Statistical Association 100, 890–907.

Fan, J., Jiang, J., 2007. Nonparametric inference with generalized likelihood ratio tests. TEST 16, 409–444.

Fan, J., Zhang, C., Zhang, J., 2001. Generalized Likelihood Ratio Statistics and Wilks Phenomenon. The Annals of Statistics 29, 153–193.

Forster, M.R., 2000. Key concepts in model selection: Performance and generalizability. Journal of Mathematical Psychology 44, 205–231.

Friedman, J., Hastie, T., Tibshirani, R., 2013. `glmnet`: Lasso and elastic-net regularized generalized linear models. *R* package version 1.8-5.

García-Jurado, I., González-Manteiga, W., Prada-Sánchez, J.M., Febrero-Bande, M., Cao, R., 1995. Predicting using box-jenkins, nonparametric, and bootstrap techniques. Technometrics 37, 303–310.

Gehrke, W., 1995. Fortran 95 Language Guide. Springer, London.

Green, P.J., 1995. Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika 82, 711–732.

Härdle, W., Hall, P., 1993. On the backfitting algorithm for additive regression models. Statistica Neerlandica 47, 43–57.

Hastie, T., Tibshirani, R., Friedman, J.H., 2003. The Elements of Statistical Learning. Springer.

Hastie, T.J., Pregibon, D., 1992. Generalized linear models, in: Chambers, J.M., Hastie, T.J. (Eds.), Statistical Models in S. Wadsworth & Brooks/Cole, p. 335.

Johnson, J.B., Omland, K.S., 2004. Model selection in ecology and evolution. Trends in Ecology & Evolution 19, 101–108.

Kuo, L., Mallick, B., 1998. Variable selection for regression models. The Indian Journal of Statistics (Special Issue on Bayesian Analysis) 60, 65–81.

Lumley, T., Miller, A., 2009. `leaps`: Regression Subsets Selection. *R* package version 2.9.

Mallows, C.L., 1973. Some Comments on CP. Technometrics 15.

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

Mcleod, A.I., Xu, C., 2011. `bestglm`: Best subset glm. *R* package version 0.33.

Miller, A., 2002. Subset Selection in Regression. Chapman and Hall/CRC, Boca Raton, FL.

Neyman, J., Pearson, E.S., 1928. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part II. Biometrika 20A, 263–294.

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

Orestes Cerdeira, J., Duarte Silva, A.P., Cadima, J., Minhoto, M., 2011. `subselect`: Selecting variable subsets. *R* package version 0.11-3.

Park, T., Casella, G., 2008. The Bayesian Lasso. Journal of the American Statistical Association 103, 681–686.

Prada-Sánchez, J.M., Febrero-Bande, M., 1997. Parametric, non-parametric and mixed approaches to prediction of sparsely distributed pollution incidents: a case study. Journal of Chemometrics 11, 13–32.

Prada-Sánchez, J.M., Febrero-Bande, M., Cotos-Yáñez, T., González-Manteiga, W., Bermúdez-Cela, J.L., Lucas-Domínguez, T., 2000. Prediction of SO2 pollution incidents near a power station using partially linear models and an historical matrix of predictor-response vectors. Environmetrics 11, 209–225.

R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Roca-Pardiñas, J., Cadarso-Suárez, C., Tahoces, P.G., Lado, M.J., 2009. Selecting variables in non-parametric regression models for binary response. An application to the computerized detection of breast cancer. Statistics in Medicine 28, 240–259.

Roca-Pardiñas, J., González-Manteiga, W., Febrero-Bande, M., Prada-Sánchez, J.M., Cadarso-Suárez, C., 2004. Predicting binary time series of SO2 using generalized additive models with unknown link function. Environmetrics 15, 729–742.

Seber, G.A.F., 1997. Linera Regression Analysis. Wiley.

Seber, G.A.F., Wild, C., 1989. Nonlinear Regression. Wiley.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. Roy. Statist. Soc. Ser. B 58, 267–288.

Venables, W.N., Ripley, B.D., 1997. Modern applied statistics with S-Plus. Springer. second edition.

Wickham, H., 2012. `meifly`: Interactive model exploration using `GGobi`. *R* package version 0.2.

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

Wood, S.N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99, 673–686.

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.