Variable Selection in Additive Models by Nonnegative Garrote

Eva Cantoni°*, Joanna Mills Flemming†, Elvezio Ronchetti°

° Department of Econometrics

University of Geneva

1211 Geneva 4, Switzerland

† Department of Mathematics and Statistics

Dalhousie University

Halifax, Nova Scotia, Canada B3H 4J1

Eva.Cantoni@unige.ch, flemming@mathstat.dal.ca,

Elvezio.Ronchetti@unige.ch

March 2009

*Corresponding author

Abstract

We adapt Breiman's (1995) nonnegative garrote method to perform variable selection in nonparametric additive models. The technique avoids methods of testing for which no general reliable distributional theory is available. In addition it removes the need for a full search of all possible models, something which is computationally intensive, especially when the number of variables is moderate to high. The method has the advantages of being conceptually simple and computationally fast. It provides accurate predictions and is effective at identifying the variables generating the model. To illustrate our procedure, we analyze logbook data on blue sharks (Prionace glauca) from the United States pelagic longline fishery. In addition we compare our proposal to a series of available alternatives by simulation. The results show that in all cases our methods perform better or as these alternatives.

Keywords: Blue shark logbook data; cross-validation; nonnegative garrote; nonparametric regression; shrinkage methods; variable selection.

1 Introduction

Variable selection is an important issue in any statistical analysis, whether parametric or nonparametric in nature. Practically speaking, one is interested in determining the strongest effects that explain the response variable. Statistically speaking, variable selection is a way of reducing the complexity of the model, in some cases by admitting a small amount of bias to improve accuracy.

As a motivating example we consider data obtained by Julia Baum (see Baum, 2007 for further details) from the U.S. National Marine Fisheries Service Pelagic Observer Program (`http://www.sefsc.noaa.gov/pop.jsp`). Recently Myers et al. (2007) have been utilizing this data to investigate ecological impacts of eliminating top predators like sharks from oceanic food webs. Here we look specifically at catches of the most commonly caught shark, the blue shark (Prionace glauca), in the main areas where they are caught in the Northwest Atlantic, that is Northeast Coastal and Distant Atlantic (Area 6 and 7 as defined in Figure 1 in Baum et al., 2003). This avoids the presence of excess of zeros and puts us in position to propose a nonparametric additive model for the blue shark counts. Such a model is more flexible than its parametric counterpart in being able to accommodate covariates which are potentially nonlinearly related to some function of the response (i.e. the counts). The statistical goal is to simultaneously fit a nonparametric model and perform variable selection.

A nonparametric framework is more challenging than a parametric approach because of the lack of underlying assumptions that makes it difficult to define a general test approach for variable selection. Some notable exceptions exist, but only with strong restrictions: in special situations or for particular smoothers (see, e.g. Bock and Bowman, 1999 for local polynomials; Cantoni

and Hastie, 2002 for smoothing splines).

Subset selection is a well-known approach to variable selection: it selects a model containing a subset of available variables, according to a given optimality criterion and requires that one visits all possible models. This approach quickly becomes infeasible when the covariate dimension is too large even when efficient algorithms exist (e.g. *leaps and bounds* in the case of linear regression, see Furnival and Wilson, 1974). Stepwise procedures are a working compromise as they reduce the number of models for comparison. However, they suffer from dependence on the path chosen through the variable space and may be inconsistent. In addition, both subset selection and stepwise selection are discrete processes that either retain or discard one variable while shrinkage methods (e.g. ridge regression in the case of linear models) are continuous in this regard, which leads to lower variability.

Shrinkage methods have emerged and gained popularity (especially in the parametric context) in recent years. In addition, methods that simultaneously address estimation and variable selection now exist (e.g. LASSO, see Tibshirani, 1996, and LARS, see Efron, Hastie, Johnstone, and Tibshirani, 2004). In the nonparametric setting, a modified LASSO for additive models (method called PAM) has been proposed by Avalos, Grandvalet, and Ambroise (2007) and an adaptive LASSO suggested by Zou (2006). Within the boosting framework, two approaches in particular would be suitable in our context: the L_2 boost for additive models by Buhlmann and Yu (2003) and the GAMBoost of Tutz and Binder (2006) for generalized linear models. In addition, the method of COSSO has been proposed by Lin and Zhang (2006). Efficient algorithms for model selection with shrinkage methods have been provided by Yuan and Lin (2006).

Here, we propose a simple approach to variable selection for nonpara-

metric additive models based on the nonnegative garrote idea of Breiman (1995) which has simultaneously the properties of subset selection, shrinkage and stability as mentioned above. It has the advantage of being conceptually simple (like its original parametric counterpart) and computationally reasonable, and it can be used with any smoother. These desirable characteristics are not shared simultaneously by alternative methods with which we compare results. The idea was suggested in Cantoni, Flemming, and Ronchetti (2006) and independently in Yuan (2007) in the ANOVA framework. In this paper we provide in addition a detailed discussion on the choice of the smoothing parameters, a detailed comparison with several alternative approaches, and a full implementation of the model.

As we shall see in Sections 3 and 4, our proposal is a reliable variable selection procedure which is able to identify the true underlying model, with our procedure (C) (see Section 2.1) giving the best results in general.

The paper is organized as follows. We introduce the methodology in Section 2. Specifically, we discuss the automatic choice of the parameters involved (Sections 2.1 and 2.2) and provide guidelines for different options. In Section 3 we demonstrate our methodology using the blue shark data. Results from the simulation study follow in Section 4. Both demonstrations provide strong evidence that our proposal works well. A discussion (Section 5) closes the paper.

2 Methodology

A typical dataset of interest will consist of p explanatory variables $x_{1i}, \ldots, x_{pi}$ and a response variable Y_i for each of the $i = 1, \ldots, n$ independent individuals

for which we postulate an additive model of the form

$$Y_i = \alpha + \sum_{k=1}^{p} f_k(x_{ki}) + \epsilon_i, \tag{2.1}$$

for $i = 1, \ldots, n$.

Model (2.1) is often presented with only univariate functions for convenience, but it must be emphasized that this property is not necessary. In fact, component functions with two or three dimensions, as well as categorical variable terms (factors) and interactions between them can replace the univariate functions $f_k(x_k)$. Moreover, some of the functions in Model (2.1) may be defined parametrically, giving rise to a semiparametric model.

We suppose that the variables x_k have been centered by subtracting off their sample means. This is not a theoretical restriction, but rather a requirement to use Breiman's code, see Section 2.3 for further details.

Given an initial estimate $\hat{g}_k^{h_k}(x_k)$ of each function $f_k(x_k)$, the nonnegative garrote approach solves

$$\min_{c_k} \sum_{i=1}^{n} \left(y_i - \alpha - \sum_{k=1}^{p} c_k \hat{g}_k^{h_k}(x_{ki}) \right)^2 \tag{2.2}$$

under the constraints $c_k \geq 0$ and $\sum_{k=1}^{p} c_k \leq s$. The final estimate of $f_k(x_{ki})$ is $\hat{f}_k(x_{ki}) = c_k \hat{g}_k^{h_k}(x_{ki})$.

The parameters $h_1, \ldots, h_p$ are referred to as the smoothing parameters of the initial functions estimates $\hat{g}_1^{h_1}, \ldots, \hat{g}_p^{h_p}$. Alternatively one can consider the degrees of freedom (see Hastie and Tibshirani, 1990, p. 128). Most smoothing techniques (e.g. splines, loess, local polynomials), allow one parameter for each function [the AMlet technique (Sardy and Tseng, 2004) is an exception here in that it requires only a single parameter]. Note also that c_k depends on s, and s is regarded as an additional parameter. We will discuss the choice of these parameters in Sections 2.1 and 2.2 below.

Our proposal (2.2) generalizes the original proposal of Breiman (1995) which is recovered with $\hat{g}_k^{h_k}(x_k) = \hat{\beta}_k x_{ki}$, where $\hat{\beta}_k$ are the ordinary least squares estimates in the linear model $y_i = \alpha + \sum_{k=1}^{p} \beta_k x_{ki} + \epsilon_i$. In this parametric situation no choice of $h_1, \ldots, h_p$ is required.

The theoretical basis for our proposal can be traced back to the parametric case, where Zou (2006) has shown that the nonnegative garrote is essentially equivalent to the adaptive LASSO. This is a LASSO procedure with a weighted penalty function, where the weights are proportional to the inverse of the least squares estimators of the coefficients and are used to penalize different coefficients in the L_1 penalty. Under the conditions given in Zou (2006) in the parametric case, the adaptive LASSO is consistent and this implies the same property for the nonnegative garrote; see Zou (2006), Corollary 2, Section 3.4 or Yuan and Lin (2007). Notice however that, as pointed out by a referee, the proposed algorithm only scales the initial fit, and for typical smoothers this implies that the initial fit is itself consistent.

Given an initial estimate of all the additive functions in Model (2.1) and a value for s, the nonnegative garrote will automatically give in a single step a set of coefficients $c_1, \ldots, c_p$ that will provide information on the importance of each variable in the model. For instance, if $c_k = 0$, the variable x_k is considered uninformative and can be removed from the model. Alternatively the variable contribution to the model will be shrunk by some proportion c_k or left unchanged (if $c_k = 1$). Decreasing s has the effect of increasing the shrinkage of the nonzeroed functions and making more of the c_k become zero. The nonnegative garrote can be viewed as a method for comparing all possible models, but unlike subset selection, it avoids fitting each model separately, therefore making its use possible at low computational cost even for large values of p. Note that as in LASSO and in the parametric version

of the nonnegative garrote, the c_k as a function of s are not restricted to be strictly monotonic and they can even be larger than 1 for some values of s.

2.1 Choice of $h_1, \ldots, h_p$

In order for the method to perform well, it is important that the smoothing parameters $h_1, \ldots, h_p$ of the initial fits $\hat{g}_k^{h_k}$ be selected in a reasonable manner. They can either be set by the user (perhaps on the basis of asymptotic results, see Opsomer and Ruppert, 1998) or selected automatically with a data driven approach (e.g. cross-validation, see Härdle, 1990, Chapter 5). Here we take the second approach, specifically that proposed by Wood (2004). His procedure allows one to automatically select the smoothing parameters by addressing the problem in the more general framework of parameter estimation with multiple quadratic penalties.

We consider the following non exhaustive list of options with which to obtain an initial fit of the data:

(A) Estimate $h_1, \ldots, h_p$ automatically (by cross-validation, for example) on the basis of the p univariate nonparametric regressions $y_i = g_k(x_{ki}) + \epsilon_i$ for $k = 1, \ldots, p$, to produce $\hat{g}_k^{h_k}$.

(B) Given starting values $h_1^0, \ldots, h_p^0$ provided by the user, estimate $h_1, \ldots, h_p$ automatically (by cross-validation, for example) at each step of the backfitting algorithm (Hastie and Tibshirani, 1990, p. 91). This modified backfitting algorithm reads as follows:

1. Initialize: $\hat{\alpha} = \bar{y}$, $h_k = h_k^0$ for $k = 1, \ldots, p$, and $\hat{g}_k^{h_k} = \hat{g}_k^{h_k^0}$ for $k = 1, \ldots, p$.
2. Cycle: $j = 1, \ldots, p, 1, \ldots, p, \ldots$
   Produce estimates $\hat{g}_j^{h_j}$ by smoothing the partial residuals $\big(Y_i -$

$\hat{\alpha} - \sum_{k \neq j} \hat{g}_k^{h_k}(x_{ki})\big)$ on x_j, with h_j chosen automatically.

3. Continue Step 2 until the individual functions do not change.

(C) Estimate $h_1, \ldots, h_p$ automatically by minimizing a given criterion in the p dimensional space.

Procedure (C) is certainly the most desirable, but is not yet widely implemented in software packages. Procedure (A) is the simplest approach but neglects the correlation between covariates. Procedure (B) is a working compromise but is again effective only when there is little correlation between covariates. Note that the re-estimation of the smoothing parameter at each step of the backfitting algorithm might, in principle, affect the convergence of the backfitting algorithm. However, we never experienced this situation in our examples and simulations. We can expect procedure (C) to perform better than (B), which in turn will perform better than (A), but it is not clear a priori how large the differences will be.

2.2 Choice of s

The accuracy of the model can be measured through the (average) prediction error defined as

$$PE_s(\hat{\alpha}, \hat{f}_1^{h_1}(x_{1i}), \ldots, \hat{f}_p^{h_p}(x_{pi})) = \frac{1}{n} \sum_{i=1}^{n} E\Big(Y_i^{new} - \hat{\alpha} - \sum_{k=1}^{p} \hat{f}_k^{h_k}(x_{ki})\Big)^2, \quad (2.3)$$

where $s = \sum_{k=1}^{p} c_k$, $\hat{f}_k^{h_k}(x_{ki}) = c_k \hat{g}_k^{h_k}(x_{ki})$ and the expectation on the right hand side of Equation (2.3) is taken over Y_i^{new}. The best value of s is then defined as the minimizer of $PE_s(\hat{\alpha}, \hat{f}_1^{h_1}(x_{1i}), \ldots, \hat{f}_p^{h_p}(x_{pi}))$.

Of course, in practice $PE_s(\hat{\alpha}, \hat{f}_1^{h_1}(x_{1i}), \ldots, \hat{f}_p^{h_p}(x_{pi}))$ is not observable and needs to be estimated. V-fold cross-validation is an approach used to mimic

the behaviour of new observations coming into play, when only a single sample is available. It splits the data into V subsets. Denote by $\mathcal{I}_1, \ldots, \mathcal{I}_V$ the sets of the corresponding observation indices. For each value of s, the cross-validation estimator of (2.3) is then

$$\begin{aligned}\widehat{PE}_s(\hat{\alpha}, \hat{f}_1^{h_1}(x_{1i}), \ldots, \hat{f}_p^{h_p}(x_{pi})) = \\ = \frac{1}{V}\sum_{v=1}^{V}\frac{1}{|\mathcal{I}_v|}\sum_{i\in\mathcal{I}_v}\left(Y_i - \hat{\alpha}^{(-v)} - \sum_{k=1}^{p} c_k^{(-v)}\hat{g}_k^{h_k,(-v)}(x_{ki})\right)^2, \qquad (2.4)\end{aligned}$$

where $\hat{\alpha}^{(-v)}$, $\hat{f}_k^{h_k,(-v)}$ and $c_k^{(-v)}$ are obtained from the sample containing all the observations but those in $\mathcal{I}_v$. Values of V between 5 and 10 produce satisfactory results and are known to be a good balance between bias and variance in the estimation of PE_s, that is between the high variance if V is large (e.g. $V = n$ for leave-one out cross-validation) and the bias if V is smaller (because of the smaller size of the training set); see Breiman (1995) and Hastie, Tibshirani, and Friedman (2001, p. 214-7).

2.3 Implementation

Presently, considering all the procedures described in Section 2.1 requires the use of several different software packages. There are essentially two parts to our approach: the initial fit followed by the nonnegative garrote for variable selection. The user has the following options:

Initial fit:

- Procedure (A): `smooth.spline` function of Splus. Note that the function `smooth.spline` of R would produce the same results.
- Procedure (B): `addreg` function for Splus available from Statlib at `http://lib.stat.cmu.edu/S/` (funfits module, version 5.1, formerly

available from D. Nychka website).

- Procedure (C) `gam` from the R package `mgcv 1.3-29`, see Wood (2006).

We have used Splus (Version 7.0.0 for Linux 2.4.21 : 2005) and R version 2.6.2 (2008-02-08).

Nonnegative garrote:
We adapted the Fortran code L. Breiman had publicly available on his website. The algorithm makes use of a modification of the nonnegative least squares algorithm by Lawson and Hanson (1974). The predictors must be centered at zero by subtracting off their sample means. Note that for a given set of initial estimates $\hat{g}_k^{h_k}(x_k)$ for $k = 1, \ldots, p$, the nonnegative garrote Equation (2.2) is as simple as its parametric counterpart. We linked the Fortran code (note that redefinition of some of the input quantities was required) both within Splus and R and intend to distribute our routines as an R package. Based on the equivalence between the adaptive LASSO and the nonnegative garrote an alternative implementation may be to use the `lars` package in R. Also note that (2.2) is a quadratic optimization problem with constraints and therefore any program that can address this kind of problem could be used, e.g. function `pcls` in the R package `gam` or R package `quadprog`.

3 Example

[Figure 1 about here.]

In this section, we analyze the blue sharks dataset using our proposal. The model, with all the covariates can be written as

$$\log(\texttt{bluesharks}+1) = \alpha + f_1(\texttt{DOFY}) + f_2(\texttt{NLIGHTST}) + f_3(\texttt{SOAKTIME}) + \\ + \quad f_4(\texttt{AVGHKDEP}) + f_5(\texttt{OCEAND}) + f_6(\texttt{TEMP}) + \log(\texttt{TOTHOOKS}) + \epsilon, \quad (3.1)$$

where the covariates considered are day of the year (DOFY), number of light stick used (NLIGHTST), soak duration (amount of time from the midpoint of the gear setting to the midpoint of the gear hauling, SOAKTIME), hook depth as measured by the average of the minimum and the maximum of the hook depth (AVGHKDEP), ocean depth (OCEAND), surface water temperature (TEMP) and the total number of hooks (TOTHOOKS). Note that the total number of hooks measures the effort and is introduced as an offset to standardize the catch data as it is usual in fisheries science. Other covariates were available but were not used (for different reasons, including missingness issue and collinearity). The sample size is 91.

With smoothing parameters $h_1, \ldots, h_p$ automatically chosen according to Procedure (C) (see Section 2.1), we obtain the results as depicted in Figure 1. This plot identifies the strongest effects (the components that enter first in the model as s increases) which in this case are (in the order of appearance) TEMP, OCEAND and DOFY. The bold vertical line shows the value of s automatically chosen by 5-fold cross-validation (see Section 2.2). Those c_k which are zero for this value of s (=2.3) identify the variables that can be removed from the final model: SOAKTIME and NLIGHTST. The importance of AVGHKDEP is borderline. The other values of c_k are 0.86, 0.62 and 0.82 respectively, for TEMP, OCEAND and DOFY, indicating a shrinkage with respect to the initial fit. This shrinkage is more severe for OCEAND. The nonparametric model considered in our analysis is certainly a welcome alternative to a fully linear analysis as indicated by the nonlinear effects present in the final model, see

Figure 2. In particular, the day of the year has a complicated functional form, the ocean depth is likely a linear effect and the surface water temperature may well be approximated by a cubic term, but, of course, we would loose some nuances by doing so.

[Figure 2 about here.]

4 Simulation Study

In this section we compare the different procedures available within our proposal to a series of alternatives described in detail below. We will evaluate the prediction accuracy and the ability of each approach to extract the true underlying model.

Our nonnegative garrote proposal makes available 4 different options. Procedures (A) and (B) as described in Section 2.1, and two versions of Procedure (C), hereafter referred to as Procedures (C1) and (C2). Procedure (C1) uses the smoothing parameters obtained from the initial fit with the entire dataset on the cross-validated samples (80% of the data if $V = 5$) and Procedure (C2) re-estimates the smoothing parameter automatically on each of the cross-validated samples. This same distinction is not necessary for Procedures (A) and (B) because the software allows the specification of the degrees of freedom (instead of the smoothing parameters) which don't need to vary with the sample size.

Alternative approaches:
To contrast the results of our approach, we have considered the following alternatives:

- `gam` in `mgcv 1.3-29` with the default option for the spline basis, a thin plate regression spline (`bs="tp"`). This is the initial fit of our procedure

(C1) and (C2) and is considered a benchmark to evaluate the gain in term of ISE with the nonnegative garrote additional step. No variable selection is possible in this case.

- `gam` in `mgcv 1.3-29` with a thin plate regression spline with shrinkage (`bs="ts"`), which automatically allows for variable selection.

- `GAMBoost` from `GAMBoost 1.0` which implements the proposal by Tutz and Binder (2006).

- The COSSO proposal by Lin and Zhang (2006) via the Matlab code available on the authors' website at `http://www4.stat.ncsu.edu/~hzhang/pub.html`. There is also an R version, but we have been unable to get it running properly.

- The PAM approach presented in Avalos, Grandvalet, and Ambroise (2007) with the Matlab code available at `http://www.isped.u-bordeaux2.fr/ANNUAIRE/FR-M_AVALOS.htm`. The ISE measure is not available for this approach, given that the current code does not allow for prediction on a validation sample.

- A backward stepwise approach based upon a generalized cross-validation criterion, see the detailed description in Section 4.1 in Brumback, Ruppert, and Wand (1999).

We consider the generating process of Example 1 in Section 7 of Lin and Zhang (2006). It is a simple additive model in $\mathcal{R}^{10}$, where the underlying generating model for $i = 1, \ldots, 100$ is

$$Y_i = f(\mathbf{x}_i) + \epsilon_i = f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{4i}) + \epsilon_i, \qquad (4.1)$$

where $\mathbf{x}_i = (x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}, x_{10i})$ and

$$f_1(s) = 5s, \quad f_2(s) = 3(2s-1)^2, \quad f_3(s) = \frac{\sin(2\pi s)}{2 - \sin(2\pi s)},$$

$$f_4(s) = 6\big(0.1\sin(2\pi s) + 0.2\cos(2\pi s) + 0.3\sin^2(2\pi s) + 0.4\cos^3(2\pi s) + 0.5\sin^3(2\pi s)\big).$$

As a consequence there are 6 uninformative dimensions. The variables $X_1, \ldots, X_{10}$ are built according to the following "compound symmetry" design: $X_j = (W_j + tU)/(1+t)$, where $W_1, \ldots, W_{10}$ and U are i.i.d. from Uniform(0,1) which results in $Corr(X_j, X_k) = t^2/(1+t^2)$ for $j \neq k$. The uniform design corresponds to the case where $t = 0$. The values $t = 1$ and 3 produce covariates with correlations of 0.5 and 0.9, respectively. The error term ϵ_i is generated according to a centered normal distribution with variance equal to 1.74 (signal-to-noise ratio of 3 in the uniform case) in a first scenario and with variance equal to 3.9 (signal-to-noise ratio of 2 in the uniform case) in a second scenario. (Note that $Var(f_1(x_1)) = 2.08$, $Var(f_2(x_2)) = 0.80$, $Var(f_3(x_3)) = 3.30$ and $Var(f_4(x_4)) = 9.45$, see Lin and Zhang, 2006, p. 2284.)

[Table 1 about here.]

We measure the accuracy of the method being used to obtain $\hat{f}(\mathbf{x}) = \sum_{k=1}^{10} \hat{f}_k(x_k)$ via the integrated squared error (ISE), where $\text{ISE} = E_X\big((\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2\big)$, estimated by Monte Carlo using 10,000 test points generated from the same distribution as the training points. Note that some of the terms in $\hat{f}(\mathbf{x})$ could be zero as determined by the method being used, while $f(\mathbf{x})$ is the true generating model as defined by (4.1).

We begin by examining the predictive ability of each method. Table 1 presents the average ISE over the 100 simulations. We first comment on the set of results for a signal-to-noise ratio equal to 3. As expected the results

from the nonnegative garrote based on Procedure (A) yield the worst results and should not be recommended. Procedure (B) improves the results and does as well as the `gam` fit with no shrinkage. The (C1) and (C2) versions of our proposal improve over the initial `gam` fit and are the best performers. Note that Procedure (C2) performs better than Procedure (C1), showing that it is worthwhile to adjust the smoothing parameter to the sample size in the cross-validation approach. The `gam` fit with shrinkage is not as good as the Procedures (C1) and (C2) of the nonnegative garrote. GAMBoost is worst than all the nonnegative garrote options (except Procedure (A)). The COSSO performance is similar to the nonnegative garrote Procedure (C1) except for $t = 3$, and the stepwise approach behaves like a simple `gam` fit with no shrinkage, that is very slightly worst than COSSO. For the signal-to-noise ratio equal 2 scenario, the ISE is larger as expected. All the comments for the larger signal-to-noise ratio can be repeated here, except those regarding COSSO and the stepwise approach, which seem to suffer much more in presence of larger noise.

[Table 2 about here.]

Table 2 and 3 display the number of times (out of the 100 simulations) that each variable has been selected to appear in the final model for the a signal-to-noise ratio of 3 and 2 respectively.

As a general comment we can say that methods that are able to pick all the informative variables tend to retain more unnecessary variables. On the other hand, approaches that discard more unnecessary variables, miss the signal more often.

PAM is both less effective in identifying the generating signal and more prone to retain many irrelevant variables. Other techniques that do not discard the irrelevant variables are GAMBoost and gam with shrinkage. Within

the nonnegative garrote options, Procedures (A) and (B) are quite similar in terms of variable selection, despite their difference in ISE. This means that both methods pick up the relevant variables equally well, but that the functions are better estimated under (B). Procedures (C1) and (C2) are very good at identifying the signal, with Procedure (C2) performing also very well in discarding the irrelevant variables. COSSO can perform very well only in particular situations: high signal-to-noise ratio or low correlation between the covariates. In other situations, it tends to either miss the signal or to retain too many variables. In keeping with Shao (1993), who considers *good models* as those which contain the true generating model, our nonnegative garrote Procedure (C2) should be preferred. It performs very well over all the settings considered here. Note also that the presence of some extra variables in the final model does not seem to impact the predictive ability of our approaches (see Table 1).

[Table 3 about here.]

One has to be careful when reading the results in Table 2 for $t = 1$ and $t = 3$ since the X's are correlated in these cases, and consequently substitution can arise. We decide nevertheless to report the results in this manner, given that all of the methods under investigation are affected in the same way.

We also ran the nonnegative garrote procedures with $V = 10$ folds. The results (not reported here) were very similar.

5 Discussion

We have proposed a model selection approach based on nonnegative garrote for variable selection in nonparametric regression. We have compared (via

simulations) the performance of its four versions to available alternatives. In terms of predictive ability, Procedures (C1) and (C2) of our approach perform very well. Alternative Procedure (A) and (B) are not as good with respect to predictive ability, but are quite effective in identifying the underlying model, although additional spurious variables are included at times. In contrast, the alternatives considered do not perform as well in terms of ISE and/or in terms of retaining the correct variables. More precisely, the shrinkage approach within `gam` tends to include too many variables in the model, GAMBoost is not effective in terms of ISE and COSSO tends to select smaller models, sometimes missing important variables, and is sensitive to the signal-to-noise ratio. The stepwise approach shows a tendency to select very large models, including several irrelevant variables.

Wood and Augustin (2002) suggested an ad-hoc procedure to try to obtain a variable selection procedure from the automatic smoothing parameter selection. Their approach is based essentially on 3 criteria (see their Section 3.3). This involves some manual tuning and is very difficult to implement on a large scale.

Further work includes the extension of this approach to the entire GAM (non Gaussian) class of models and the consideration of resistance-robustness aspects building on work by Cantoni and Ronchetti (2001) and Cantoni, Mills Flemming, and Ronchetti (2005).

For practical applications like the blue sharks example discussed herein, our approach is particularly desirable. Our code is readily available and user-friendly, results are easily interpreted and most importantly nonlinear effects are quite apparent when present.

6 Acknowledgement

This work has been supported by grant 1214-66989 of the Swiss National Science Foundation. The authors would also like to thank Julia Baum and Sylvain Sardy for helpful discussions, David Conne for providing initial simulation results, the Editor, Associate Editor and a referee for useful remarks and references which improved an earlier version of the paper.

References

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

Baum, J. (2007). *Population- and community-level consequences of the exploitation of large predatory marine fishes*. Ph. D. thesis, Biology Department, Dalhousie University, Halifax, Canada.

Baum, J. K., Myers, R. A., Kehler, D. G., Worm, B., Harley, S. J., and Doherty, P. A. (2003). Collapse and conservation of shark populations in the northwest atlantic. *Science*, **299**, 389–392.

Bock, M. and Bowman, A. W. (1999). Comparing bivariate nonparametric regression models. Technical Report 99-1, Department of Statistics, University of Glasgow, Scotland.

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

Brumback, B. A., Ruppert, D., and Wand, M. P. (1999). Comment on "variable selection and function estimation in additive nonparametric

regression using a data-based prior". *Journal of the American Statistical Association*, **94**, 794–797.

Buhlmann, P. and Yu, B. (2003). Boosting With the L2 Loss: Regression and Classification. *Journal of the American Statistical Association*, *98*,(462), 324–339.

Cantoni, E., Flemming, J., and Ronchetti, E. (2006). Variable selection in additive models by nonnegative garrote. Technical Report 2006.05, Department of Econometrics, University of Geneva.

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

Cantoni, E., Mills Flemming, J., and Ronchetti, E. (2005). Variable selection for marginal longitudinal generalized linear models. *Biometrics*, **61**, 507–514.

Cantoni, E. and Ronchetti, E. (2001). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing*, **11**, 141–146.

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

Furnival, G. M. and Wilson, Robert W., J. (1974). Regression by leaps and bounds. *Technometrics*, **16**, 499–511.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning*. New York: Springer.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*.

London: Chapman & Hall.

Lawson, C. and Hanson, R. (1974). *Solving Least Squares Problems*. Englewood Cliffs, NJ: Prentice-Hall.

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

Myers, R. A., Baum, J. K., Shepherd, T. D., Powers, S. P., and Peterson, C. H. (2007). Cascading effects of the loss of apex predatory sharks from a coastal ocean. *Science*, **315**, 1846–1850.

Opsomer, J. D. and Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association*, **93**, 605–619.

Sardy, S. and Tseng, P. (2004). AMlet, RAMlet, GAMlet: Automatic nonlinear fitting of additive models, robust and generalized, with wavelets. *Journal of Computational and Graphical Statistics*, **13**, 283–309.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**, 486–494.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B, Methodological*, **58**, 267–288.

Tutz, G. and Binder, H. (2006). Generalized Additive Modeling with Implicit Variable Selection by Likelihood-Based Boosting. *Biometrics*, *62*,(4), 961–971.

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. (2006). *Generalized Additive Models: An Introduction with R.* Chapman Hall/CRC.

Wood, S. N. and Augustin, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, **157**, 157–177.

Yuan, M. (2007). Nonnegative garrote component selection in functional ANOVA models. *Proceedings of AI and Statistics, AISTATS*,.

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

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*,(2), 143–161.

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

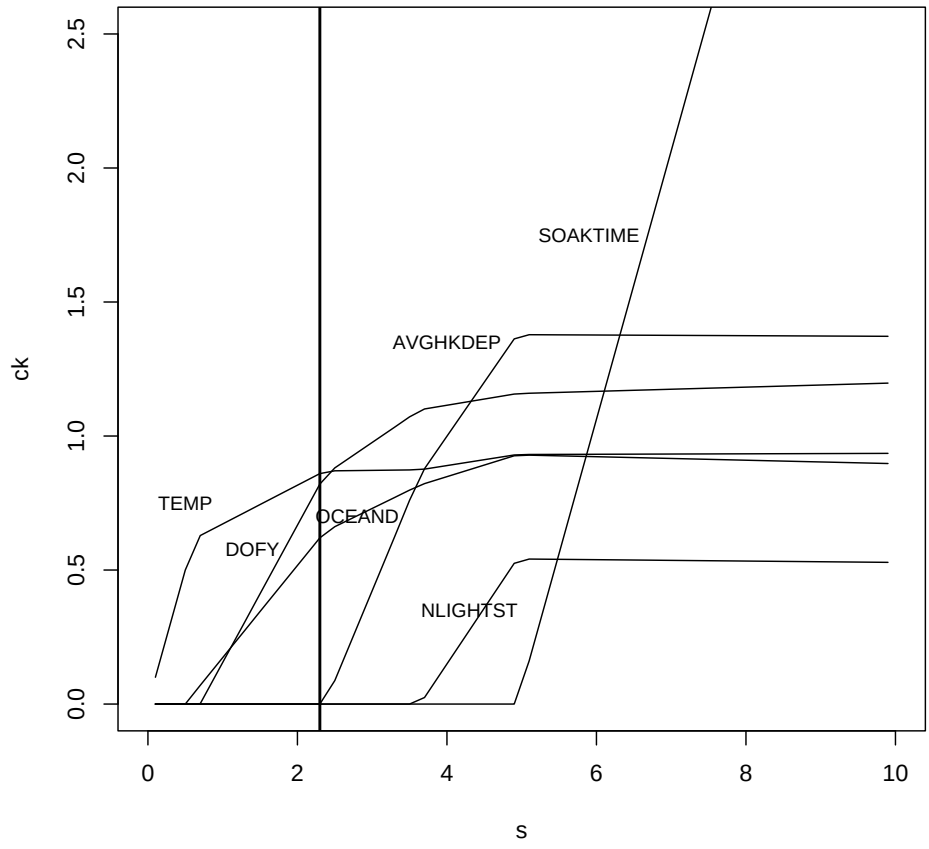

Figure 1: Shrinkage values c_k as a function of s for the blue sharks dataset. The bold vertical line indicates the value of s chosen by 5-fold cross-validation.

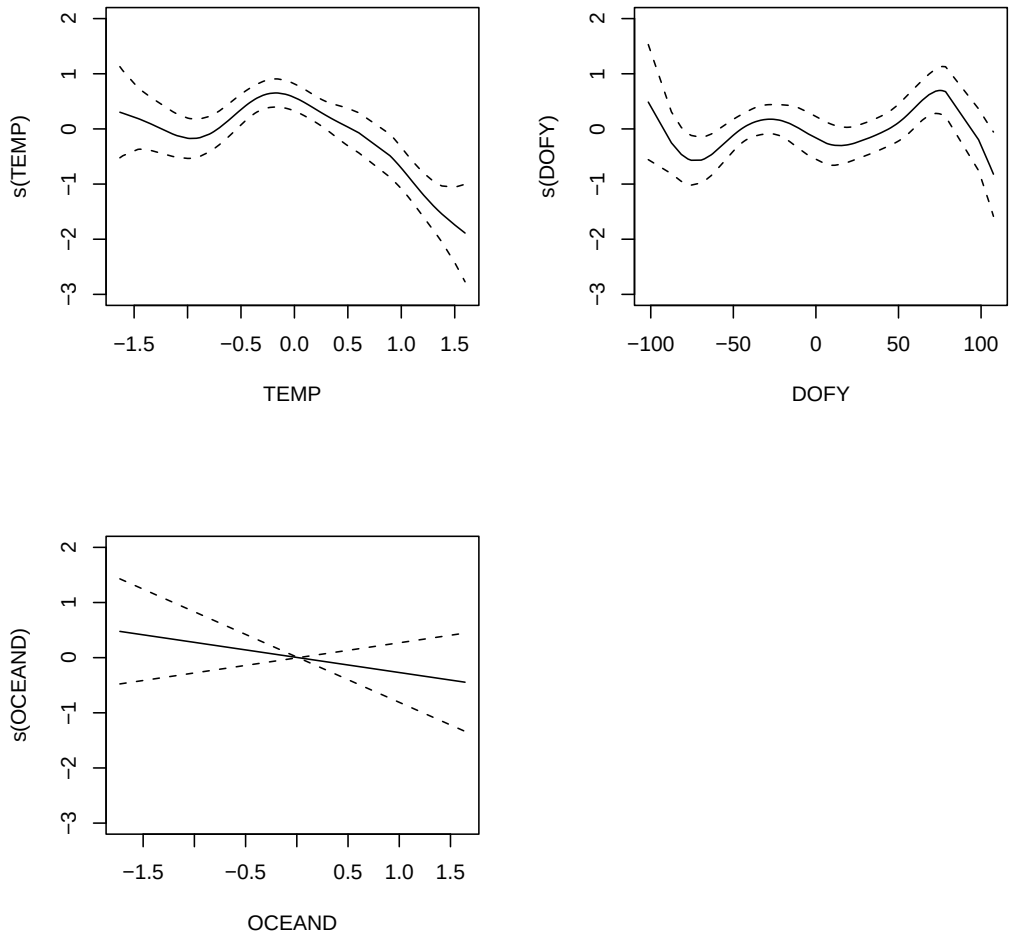

Figure 2: Fitted functions for the final model after variable selection by our nonparametric nonnegative garrote.

	t=0	t=1	t=3
Signal-to-noise ratio = 3			
NNG - Proc. (A)	1.71 (0.10)	1.24 (0.06)	1.12 (0.05)
NNG - Proc. (B)	0.81 (0.03)	0.85 (0.03)	0.95 (0.04)
gam (no shrinkage)	0.84 (0.03)	0.89 (0.03)	0.95 (0.06)
NNG - Proc. (C1)	0.72 (0.03)	0.75 (0.04)	0.71 (0.03)
NNG - Proc. (C2)	0.65 (0.03)	0.64 (0.04)	0.64 (0.03)
gam (shrinkage)	0.76 (0.03)	0.81 (0.03)	0.84 (0.05)
GAMBoost	1.10 (0.04)	1.31 (0.04)	1.09 (0.03)
COSSO	0.73 (0.03)	0.79 (0.03)	0.91 (0.04)
Stepwise GCV	0.82 (0.03)	0.87 (0.03)	0.93 (0.06)
Signal-to-noise ratio = 2			
NNG - Proc. (A)	2.33 (0.12)	1.93 (0.09)	1.94 (0.08)
NNG - Proc. (B)	1.71 (0.06)	1.83 (0.07)	1.67 (0.06)
gam (no shrinkage)	1.71 (0.06)	1.84 (0.07)	2.04 (0.12)
NNG - Proc. (C1)	1.49 (0.07)	1.64 (0.09)	1.46 (0.06)
NNG - Proc. (C2)	1.34 (0.06)	1.36 (0.07)	1.37 (0.05)
gam (shrinkage)	1.51 (0.06)	1.70 (0.07)	1.90 (0.12)
GAMBoost	1.87 (0.06)	1.96 (0.05)	1.66 (0.05)
COSSO	1.60 (0.06)	1.79 (0.08)	1.88 (0.08)
Stepwise GCV	1.63 (0.07)	1.79 (0.07)	2.02 (0.12)

Table 1: Average ISE (estimated by Monte Carlo over 10,000 points) over 100 simulations and its standard error within parentheses. $V = 5$ fold cross-validation is used. Empirical standard errors are given within parentheses.

Design	Technique	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
t=0	NNG - Proc. (A)	100	100	100	100	23	21	21	15	23	23
	NNG - Proc. (B)	100	100	100	100	23	20	27	22	33	15
	NNG - Proc. (C1)	100	100	100	100	28	27	35	35	22	30
	NNG - Proc. (C2)	100	100	100	100	19	16	19	20	13	19
	gam (shrinkage)	100	100	100	100	57	69	61	60	59	61
	GAMBoost	100	100	100	100	77	76	86	78	74	78
	COSSO	100	98	100	100	2	1	0	1	0	2
	PAM	100	100	100	100	93	95	92	92	97	94
	Stepwise GCV	100	100	100	100	29	43	40	30	24	37
t=1	NNG - Proc. (A)	100	100	100	100	13	22	24	28	20	20
	NNG - Proc. (B)	99	100	100	100	34	29	32	32	29	28
	NNG - Proc. (C1)	100	100	100	100	45	44	37	35	37	32
	NNG - Proc. (C2)	99	100	100	100	24	24	22	15	18	18
	gam (shrinkage)	100	100	100	100	65	65	67	59	66	61
	GAMBoost	100	100	100	100	68	76	72	72	75	74
	COSSO	95	74	100	100	3	12	4	4	10	3
	PAM	99	100	100	100	95	97	90	95	95	92
	Stepwise GCV	100	100	100	100	46	36	44	34	43	36
t=3	NNG - Proc. (A)	80	100	100	100	33	29	34	35	40	36
	NNG - Proc. (B)	87	100	100	100	36	43	34	44	37	46
	NNG - Proc. (C1)	90	100	100	100	46	38	39	40	36	41
	NNG - Proc. (C2)	79	100	100	100	24	22	23	26	19	22
	gam (shrinkage)	100	100	100	100	62	58	58	50	57	70
	GAMBoost	95	100	100	100	55	58	68	59	64	71
	COSSO	55	78	94	100	19	23	18	19	20	20
	PAM	87	99	100	100	82	76	82	84	82	72
	Stepwise GCV	94	100	100	100	40	36	40	39	44	45

Table 2: Frequency of appearance of the variables in 100 simulations for a signal-to-noise ratio of 3.

Design	Technique	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
t=0	NNG - Proc. (A)	100	100	100	100	26	22	22	17	28	23
	NNG - Proc. (B)	99	100	100	100	22	18	22	16	27	13
	NNG - Proc. (C1)	100	100	100	100	31	31	33	34	27	33
	NNG - Proc. (C2)	100	100	100	100	18	17	20	21	12	20
	gam (shrinkage)	100	100	100	100	57	69	58	59	55	61
	GAMBoost	100	100	100	100	72	76	80	73	71	79
	COSSO	100	78	100	100	6	6	3	7	1	5
	PAM	100	100	100	100	92	98	94	96	95	92
	Stepwise GCV	100	100	100	100	31	41	39	29	24	35
t=1	NNG - Proc. (A)	94	100	100	100	17	24	21	28	25	24
	NNG - Proc. (B)	93	100	99	100	31	30	22	23	23	22
	NNG - Proc. (C1)	96	100	100	100	50	46	45	39	38	36
	NNG - Proc. (C2)	94	100	100	100	24	22	23	16	19	21
	gam (shrinkage)	99	100	100	100	67	63	63	57	66	63
	GAMBoost	100	100	100	100	68	67	70	67	67	70
	COSSO	76	53	99	100	9	11	5	6	13	9
	PAM	88	98	94	100	89	73	84	89	83	76
	Stepwise GCV	98	100	100	100	44	32	47	37	40	34
t=3	NNG - Proc. (A)	57	94	93	100	33	32	25	33	37	42
	NNG - Proc. (B)	61	96	97	100	32	33	33	37	35	38
	NNG - Proc. (C1)	71	100	99	100	44	38	37	36	37	33
	NNG - Proc. (C2)	53	98	97	100	22	24	20	20	17	25
	gam (shrinkage)	92	100	100	100	60	57	60	51	57	68
	GAMBoost	80	100	98	100	57	55	64	59	63	68
	COSSO	42	60	82	100	31	30	24	35	27	30
	PAM	88	98	94	100	89	73	84	89	83	76
	Stepwise GCV	77	100	98	100	39	38	43	39	47	49

Table 3: Frequency of appearance of the variables in 100 simulations for a signal-to-noise ratio of 2.