Variable Selection in Additive Models with an Application to Logbook Data on Blue Sharks

Joanna Mills Flemming

Assistant Professor

Department of Mathematics and Statistics
Dalhousie University
Halifax, NS, CANADA

TIES, June 2008

Collaborators & Funding

- Eva Cantoni
- Elvezio Ronchetti
- Julia Baum

- NSERC
- NICDS

Variable Selection

- Important issue in any statistical analysis
- Determine strongest effects that explain the response variable
- Reduces model complexity by admitting a small amount of bias

Motivating Example

- US National Marine Fisheries Service Pelagic Observer Program
- Catches of the blue shark, *Prionace glauca*
- Northeast Coastal and Distant Atlantic

Goals:

- *Statistical:*
  - propose an additive model
  - accommodate covariates which are potentially nonlinearly related to some function of the response (counts)
  - simultaneously fit a model and perform variable selection
- *Ecological:*
  - Are blue shark counts decreasing?

Approaches

- *Subset selection:* quickly becomes infeasible when the covariate dimension is too large
- *Stepwise procedures:* suffer from dependence on the path chosen through the variable space and may be inconsistent
- *Shrinkage methods:* have emerged and gained popularity in recent years
- **Methods that simultaneously address estimation and variable selection now exist**: modified LASSO, COSSO

Nonnegative Garrote

- Simple approach to variable selection for additive models
- Based on nonnegative garrote idea of Breiman (1995)
- Simultaneously has properties of subset selection, shrinkage and stability
- Computationally reasonable

Nonnegative Garrote

- Simple approach to variable selection for additive models
- Based on nonnegative garrote idea of Breiman (1995)
- Simultaneously has properties of subset selection, shrinkage and stability
- Computationally reasonable

Nonnegative Garrote

- Simple approach to variable selection for additive models
- Based on nonnegative garrote idea of Breiman (1995)
- Simultaneously has properties of subset selection, shrinkage and stability
- Computationally reasonable

Nonnegative Garrote

- Simple approach to variable selection for additive models
- Based on nonnegative garrote idea of Breiman (1995)
- Simultaneously has properties of subset selection, shrinkage and stability
- Computationally reasonable

Methodology

Additive Model

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

Solves

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

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 $\hat{f}_k(x_{ki}) = c_k \hat{g}_k^{h_k}(x_{ki})$.

Interpretation

- $h_1, \cdots, h_p$ are smoothing parameters of the initial function estimates $\hat{g}_1^{h1}, \cdots, \hat{g}_p^{h_p}$.
- c_k depends on s and s is regarded as an additional parameter.
- Decreasing s has the effect of increasing the shrinkage of the nonzeroed functions and making more of the c_k become zero.
- *Given an initial estimate of all the additive functions in the model and a value for s* our method will automatically give a set of coefficients $c_1, \cdots c_p$ that will provide information on the importance of each variable in the model.

Choice of $h_1, \cdots, h_p$

- Smoothing parameters of initial fits must be selected in a reasonable manner
- $\Rightarrow$ We select to use an automatically data driven approach

Choice of s

- Best value of s will be that which minimizes the PE
- $\Rightarrow$ Estimate the PE by V-fold cross-validation

Implementation

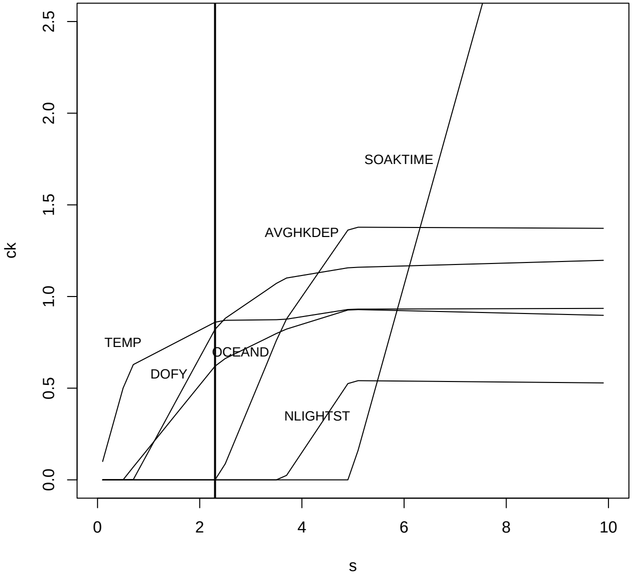
- Two parts:
- $\Rightarrow$ *gam* from the *mgcv* library in R
- $\Rightarrow$ Modified fortran code of Breiman and linked with R

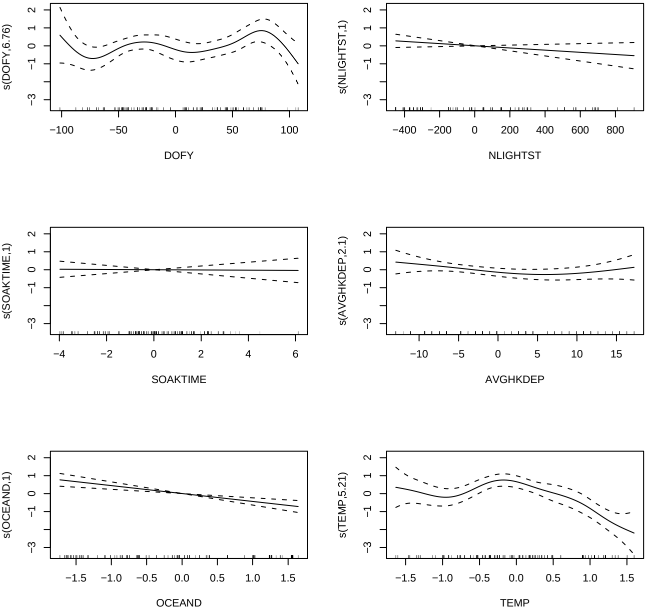
Blue Sharks Dataset

Model

$$log(bluesharks+1) = \alpha + f_1(DOFY) + f_2(NLIGHTST) + f_3(SOAKTIME) + f_4(AVGHKDEP) + f_5(OCEAND) + f_6(TEMP) + log(TOTHOOKS)$$

- Sample size is 91
- Strongest effects are TEMP, OCEAND and DOFY
- SOAKTIME and NLIGHTST can be removed
- AVGHKDEP borderline
- DOFY complicated functional form, TEMP approximately quadratic

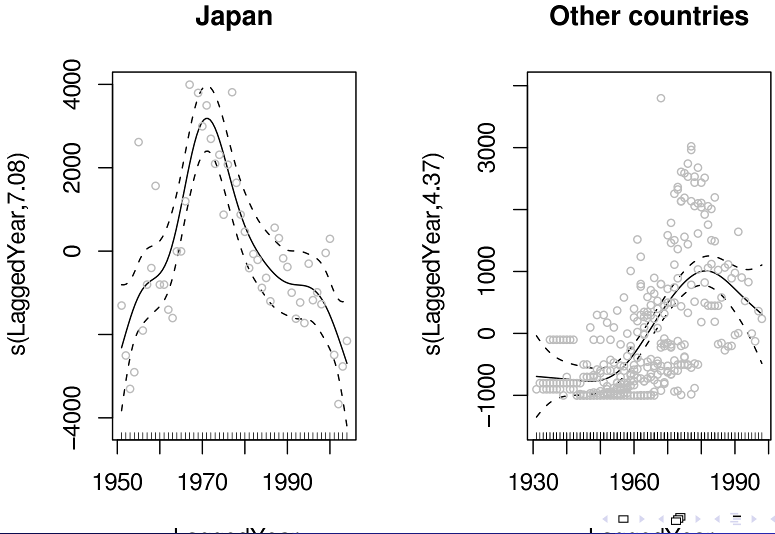

ck
s
SOAKTIME
AVGHKDEP
TEMP
OCEAND
DOFY
NLIGHTST
0.0
0.5
1.0
1.5
2.0
2.5
0
2
4
6
8
10


s(DOFY,6.76)
DOFY
s(NLIGHTST,1)
NLIGHTST
s(SOAKTIME,1)
SOAKTIME
s(AVGHKDEP,2.1)
AVGHKDEP
s(OCEAND,1)
OCEAND
s(TEMP,5.21)
TEMP

Discussion

- In terms of predictive ability, as well or better than competitors
- Code readily available and user-friendly

- **Patterns of expansion and depletion of invertebrate fisheries on a global scale**



The Future

- Extension to GAMs
- Robustness aspects