Appendices to "Combining an additive and tree-based regression model simultaneously: STIMA"

published in the Journal of Computational and Graphical Statistics

Elise Dusseldorp, Claudio Conversano and Bart Jan Van Os

Appendix A

Algorithm 1. *Simultaneous Threshold Interaction Modeling Algorithm for Tree Growing.*

1. Initialization: $\ell = 0$; $M_0 = 1$.
   Input: L, $X_1, ..., X_J$, and Y. Estimate using least squares:

$$\hat{Y}_0 = \hat{\beta}_{00} + \sum_{j=1+d}^{J} \hat{\beta}_{j0} X_j,$$

   where $d = 0$ if X_1 is continuous, and $d = 1$ if X_1 is categorical; the first index of $\hat{\beta}_{00}$ and $\hat{\beta}_{j0}$ denotes the parameter type, and the second index denotes the model growing phase (i.e., the split number ℓ). Thus, $\hat{\beta}_{00}$ denotes the estimated intercept at split $\ell = 0$ and $\hat{\beta}_{j0}$ denotes the estimated slope of the j-th predictor at split $\ell = 0$.

2. $\ell = 1$.
   Estimate using least squares:

$$\hat{Y}_1 = \hat{\beta}_{01} + \sum_{j=1+d}^{J} \hat{\beta}_{j1} X_j + \underbrace{\sum_{m=1}^{M_1-1} \hat{\beta}_{J+m1} I(X_1 \in R_{m1})}_{};$$

   Characterization of $\sum_{m=1}^{M_1-1} \hat{\beta}_{J+m1} I(X_1 \in R_{m1})$:

   - X_1 categorical with K categories (each category value is denoted by c_m):
     $M_1 = K$; $\quad \sum_{m=1}^{M_1-1} \hat{\beta}_{J+m1} I(X_1 \in R_{m1}) = \sum_{m=1}^{K-1} \hat{\beta}_{J+m1} I(X_1 = c_m)$
   - X_1 continuous. Find s_1^* of X_1 as explained in step 3:
     $M_1 = 2$; $\quad \sum_{m=1}^{M_1-1} \hat{\beta}_{J+m1} I(X_1 \in R_{m1}) = \beta_{J+1} I(X_1 \leq s_1^*)$

3. For $\ell = \ell + 1$ to L: If splitting candidate X_j is categorical, (re)order its categories according to Algorithm 2 (see below). Select R_{m^*} such that

   - $s_j^* = \max_{s_j} f_{m\ell}^2 (X_j, s_j)$, $\forall j = 1, \ldots, J$ and $\forall s_j \in X_{1j}, \ldots, X_{Nj}$
   - $X_{j^*} = \max_j f_{m\ell}^2 \left(X_j, s_j^*\right)$, $\forall j = 1, \ldots, J$
   - $m^* = \max_m f_{m\ell}^2 (X_{j^*}, s_j^*)$, $\forall m = 1, \ldots, M_{\ell-1}$

   $f_{m\ell}^2 (X_j, s_j)$ is the increase in variance-accounted-for when moving from the trunk of size $\ell - 1$ to that of size ℓ. To compute the variance-accounted-for after a split on R_m, for each X_j and s_j combination the following model is estimated, using least

squares:

$$\begin{aligned}\hat{Y} &= \hat{\beta}_0 + \sum_{j=1+d}^{J} \hat{\beta}_j X_j + \sum_{m=1}^{M_{(\ell-1)}-1} \hat{\beta}_{J+m} I((X_1, \ldots, X_J) \in R_{m(\ell-1)}) + \\ & \beta_{J+M_{(\ell-1)}} I(X_j \leq s_j | i \in R_m).\end{aligned}$$

Replace the indicator variable for R_{m^*} by two new indicator variables for the child nodes of R_{m^*}. Renumber the nodes R_m after split ℓ, with $m = 1, \ldots, M_\ell$. Estimate using least squares:

$$\hat{Y}_\ell = \hat{\beta}_{0\ell} + \sum_{j=1+d}^{J} \hat{\beta}_{j\ell} X_j + \sum_{m=1}^{M_\ell - 1} \hat{\beta}_{J+m\ell} I((X_1, \ldots, X_J) \in R_{m\ell}).$$

4. Output:

$$\hat{Y}_L = \hat{\beta}_{0L} + \sum_{j=1+d}^{J} \hat{\beta}_{jL} X_j + \sum_{m=1}^{M_L - 1} \hat{\beta}_{J+mL} I((X_1, \ldots, X_J) \in R_{mL}).$$

Algorithm 2. *Preprocessing of a categorical splitting candidate X_j.*

1. Estimate the residual values of the regression trunk model grown so far:
$$\hat{Y}_{res} = Y - \hat{Y}_{(\ell-1)}.$$
2. Compute the average estimated residual value for each category c_m of X_j:
$$\bar{Y}_{res,c_m} = mean(\hat{Y}_{res} | X_j = c_m).$$
3. Consider X_j as an ordinal variable with order relation defined by $\bar{Y}_{res,c_m}$.

Appendix B

Values of the weights w to generate the two-way interaction population data, see Section 3. The continuous by continuous interaction data are generated from the model in (3.1) and the categorical by continuous interaction data from the model in (3.2). The values of f^2 are the effect sizes of the interaction term. The values of the weight of the interaction term are in boldface.

	Independent predictors situation					
	Continuous by continuous			Categorical by continuous		
w	$f^2 = 0$	$f^2 = 0.10$	$f^2 = 0.33$	$f^2 = 0$	$f^2 = 0.10$	$f^2 = 0.33$
w_1	0.41	0.31	0.21	0.44	0.24	0.20
w_2	0.41	0.31	0.16	0.44	0.45	0.30
w_3	0.41	0.32	0.16	0.00	0.00	0.00
w_4	**0.00**	**0.77**	**1.40**	**0.00**	**0.82**	**1.48**
	Correlated predictors situation					
	Continuous by continuous			Categorical by continuous		
w	$f^2 = 0$	$f^2 = 0.10$	$f^2 = 0.33$	$f^2 = 0$	$f^2 = 0.10$	$f^2 = 0.33$
w_1	0.34	0.25	0.15	0.50	0.28	0.16
w_2	0.34	0.25	0.15	0.50	0.52	0.25
w_3	0.34	0.25	0.14	0.00	0.00	0.00
w_4	**0.00**	**0.78**	**1.38**	**0.00**	**0.82**	**1.48**

Appendix C

Description of the ten benchmark datasets:

- **Abalone**. The dataset comes from the UCI Machine Learning Repository (Asuncion and Newman, 2007). The task is to predict the age of an abalone on the basis of some physical measurements. The age of the abalone is indicated by the variable "rings" plus 1.5.

- **Baseball**. The task is to predict the logarithm of the 1987 salary from twenty-two other variables observed in 263 professional baseball players.

- **Boston House Prices**. The data frame has 506 rows and 20 columns. The response is the median value of owner-occupied houses measured for each of 506 census tracts in the Boston area. Main interest is the effect of air pollution concentration (indicated by nitrogen oxide level).

- **Cars Origin**. This dataset is also known as "1985 Auto Import database". The aim is to verify whether the characteristics of a car, its insurance risk rating, and the relative average loss payment per insured vehicle can predict the car's price.

- **Employee**. This dataset is typically used in regression analysis to model the numeric variable "Salary" on the basis of a set of covariates related to 474 employees. It is included in the SPSS 14.0 statistical software package (called "Employee data.sav"; SPSS for Windows, 2005). The results of the analysis of these data are described in detail in section 5.4.

- **Eurostoxx**. It consists of a sample of 397 European Equities listed in the Eurostoxx 600 equity index. This index includes the major equities (in terms of trading volume) listed in the European markets, namely those that are part of EU countries plus some other major markets (such as UK, Switzerland, etc.). The Bloomberg data provider allows us to download updated information about companies listed in the Eurostoxx 600. The data at our disposal were collected during January 2004. For each equity (company) we obtained information concerning: 1-year return; Profit Margin; Price/Earnings ratio; Price/Book value; Sales growth; Long-term IBES EPS growth; 1-year IBES EPS growth; Current year IBES EPS growth; Equity Consensus; EV/Book value; ROA; ROE; Index weight; Equity beta; Weighted average cost of capital; Average 3-months volatility; 200-days volatility; Total Dividend yield; 1-year Dividend Yield, and Rating. The objective was to predict the 1-year return on the basis of the other financial information.

- **EUR_USD**. In this dataset, daily levels of the following variables were observed from 12/31/1998 to 12/24/1999: USD/EUR exchange rate; 3-month Libor rate fixing; 1-year German government bond yield; 5-year German government bond yield; 10-year German government bond yield; 1-month risk reversal index; 3-month risk reversal index; industrial confidence index (Euro zone), and overnight fixing rate. The level of the USD/EUR exchange rate has been estimated on the basis of the other variables.

- **Fev**. This dataset refers to the Kahn (2005) study about the relationship between respiratory function (measured by forced expiratory volume, Fev) and smoking. The aim is to predict Fev with respect to some personal characteristics, such as age, height, gender, and smoker/nonsmoker status.

- **Juice**. This dataset was originally used by Foster, Stine, and Waterman (1998), and refers to 1,070 purchases of juices of two brands (CH and MM) in some supermarkets. Some characteristics of these purchases have been discarded because of their minor importance. In our analysis, the aim is to predict the price of CH on the basis of the time of purchase (week), the price of MM, the discount applied on CH and MM, the type of store and the brand chosen (CH or MM).

- **Home Prices**. The data are a random sample of records of resale of houses from February 15 to April 30, 1993, from the files maintained by the Albuquerque Board of Realtors. This type of data is collected by multiple listing agencies in many cities, and is used by realtors as an information base. The aim is to estimate the price of each home taking into account its main characteristics, such as size, location and annual taxes.

Appendix D

Table 1: The performance of the compared models: each cell reports the 10-fold cross-validated error (1^{st} row), its standard error in brackets (2^{nd} row) and the number of parameters (3^{rd} row).

Model	Abalone	Base-ball	Boston	Cars	Emplo-yee	EUR/ USD	Euro-stoxx	Fev	Juice	Home prices	average rank	# times best	# times worst
LM	.564 (.028) 9	.457 (.035) 42	.278 (.034) 16	.198 (.032) 19	.163 (.025) 10	.070 (.007) 13	.799 (.027)* 21	.228 (.016) 5	.218 (.009) 10	.208 (.047) 8	6.5 (5.2) 4.2	0 (1) 2	3 (2) 1
Step LM	.530 (.021) 12	.457 (.036) 50	.324 (.058) 22	.152 (.020) 19	.153 (.024) 18	.070 (.007) 10	.879 (.029) 33	.223 (.015) 8	.192 (.008) 24	.287 (.067) 14	5.8 (4.3) 6.1	0 (0) 0	2 (1) 5
RPART	.522 (.020) 48	.300 (.030) 3	.243 (.036) 6	.194 (.037) 2	.299 (.045) 3	.135 (.014) 4	.947 (.139) 2	.274 (.020) 5	.089 (.007) 7	.418 (.073) 3	6.6 (6.8) 1.7	0 (0) 9	5 (6) 1
Step GAM	.473 (.018) 17	.226 (.027) 8	.197 (.035) 20	.168 (.031) 5	.165 (.026) 6	.107 (.009) 11	.438 (.094) 12	.213 (.016) 5	.166 (.008) 12	.205 (.037) 3	4.3 (4.8) 2.6	0 (0) 2	**0** (0) 0
Step GAM_int	.626 (.027) 11	.217 (.023) 10	.197 (.035) 20	.158 (.027) 6	.163 (.025) 6	.107 (.009) 11	.438 (.094) 12	.204 (.015) 11	.167 (.008) 12	.205 (.037) 3	4.5 (4.3) 3.6	0 (0) 1	1 (0) 1
MARS	.464 (.017)* 12**	.254 (.074) 13	.167 (.028) 21	.152 (.027) 17	.161 (.025) 8	.032 (.003)* 17**	.426 (.088) 17	.213 (.016) 8	.208 (.009) 10	.214 (.053) 5	3.9 (4.3) 4.5	0 (2) 0	**0** (2) 1
GUIDE	**.454** (.017)* 24	**.129** (.015)* 8**	**.089** (.009)* 31	.131 (.021) 10	.152 (.024)* 11	.048 (.006) 17	**.370** (.046) 10**	.206 (.015) 7	**.008** (.001)* 45	.198 (.043) 3	**1.7** (2.1) 4.8	**5** (4) 1	**0** (0) 3
STIMA	.478 (.018) 16	.182 (.018) 15	.150 (.019) 23**	**.087** (.011)* 18	**.147** (.023)* 10**	**.028** (.003)* 17**	.467 (.077) 12	**.200** (.013)* 7	.038 (.003) 18**	**.181** (.035)* 7**	2.0 (1.8) 5.4	**5** (5) 0	**0** (0) 2

Notes. The best performer is given in bold face; * refers to the model presenting the lowest standard error, and ** refers to the most parsimonious model when comparing the best and the second best model in terms of cross-validated error.

References

Asuncion, A. and Newman, D. J. (2007), "UCI machine learning repository," URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Foster, D. P., Stine, R. A., and Waterman, R. P. (1998), *Business Analysis Using Regression: A Casebook*, New York: Springer.

Kahn, M. (2005), "An exhalent problem for teaching statistics," *Journal of Statistics Education*, 13, Available on line at www.amstat.org/publications/jse/v13n2/datasets.kahn.html.

SPSS for Windows (2005), *Release 14.0.0.*, Chicago: SPSS inc.