Generalized Additive Multi-Mixture Models: Some Applications

Claudio Conversano

November 1, 2001

Outline

- ▷ Generalized Additive Multi-Model (GAM-M)
- ▷ Generalized Additive Multi-Mixture Model (GAM-MM)
- ▷ A Combined Strategy for Data Mining
- ▷ Examples
- ▷ Discussion

Generalized Additive Multi-Model (GAM-M)

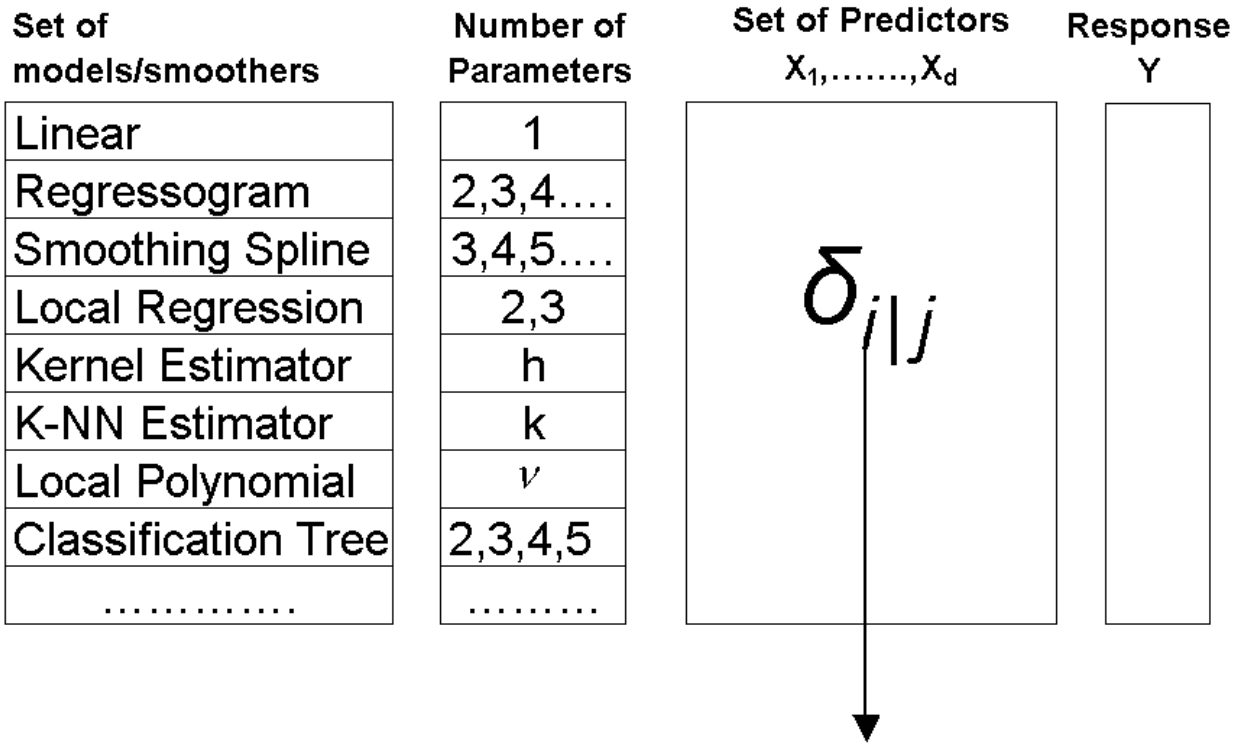
Definition. *Let assume that the response Y has an exponential family density $\rho_Y(y;\theta;\phi)$; the mean $\mu = E(Y|X_1,\ldots,X_d)$ is linked to the predictors via:*

$$g(\mu) = \alpha + \sum_{j=1}^{d}\sum_{i=1}^{K} \delta_{ij} f_{ij}(X_j, \theta_{ij})$$

where:

- ▷ f_{ij} *is a model/smoother;*
- ▷ θ_{ij} *is a vector of parameters;*
- ▷ δ_{ij} *is a dummy parameter;*
- ▷ α *is the intercept term.*

The GAM-M idea



$$g(\mu) = \alpha + \sum_{j=1}^{d} \sum_{i=1}^{K} \delta_{ij} f_{ij}\left(X_j, \theta_{ij}\right)$$

with $\sum_i \delta_{ij} = 1$.

Possible Specifications

$$g(\mu) = \alpha + \sum_{j=1}^{d} \sum_{i=1}^{K} \delta_{ij} f_{ij}(X_j, \theta_{ij})$$

Linear Regression model:

$$g(\mu) = \alpha + \sum_{j=1}^{d} X_j \beta_j$$

Generalized Additive Model:

$$g(\mu) = \alpha + \sum_{j=1}^{d} f_j(X_j)$$

Generalized Linear Model:

$$g(\mu) = \eta$$

The GAM-M Algorithm

1. *Select Model:*

   fit $\widehat{f}_{ij}(X_j, \theta_{ij})$ $\forall(i,j)$;

   choose $f_{i^*j}(X_j, \theta_{i^*j})$ w.r.t. a g.o.f. measure;

   $\delta_{i^*j} \leftarrow 1$ and $\delta_{ij} \leftarrow 0$ $\forall i \neq i^*$.

2. *Initialize:*

   $\widehat{f}_{i^*j}(X_j, \widehat{\theta}_{i^*j}) \leftarrow 0$ $\forall j$;

   $\widehat{\alpha} \leftarrow g(\sum_i^n y_i/n)$; $r \leftarrow 0$.

3. *Update:*

   $\widehat{g}^{(r)}(X_j) \leftarrow \widehat{\alpha} + \sum_{j=1}^{d} \widehat{f}_{i^*j}^{(r)}(X_j, \widehat{\theta}_{i^*j})$;

   $\widehat{p} \leftarrow \frac{e^{\widehat{g}^{(r)}(X_j)}}{1+e^{\widehat{g}^{(r)}(X_j)}}$;

   $z \leftarrow \widehat{g}^{(r)}(X_j) + \frac{(y-\widehat{p})}{[\widehat{p}(1-\widehat{p})]}$;

   $w \leftarrow \widehat{p}(1-\widehat{p})$;

   $r \leftarrow r+1$;

   update $\widehat{f}_{i^*j}(X_j, \widehat{\theta}_{i^*j})$ (*inner loop*).

4. *Verify:*
   Check convergence of the log-likelihood criterion:
   $D(\mathbf{y}, \hat{\mathbf{p}}) = -2\sum[y_n \ln \widehat{p}_n + (1-y_n)\ln(1-\widehat{p}_n)]$

The GAM-M Algorithm *(inner loop)*

$r \leftarrow r + 1$;

$\widehat{f}_{i^*j}^{(r)}(X_j, \widehat{\theta}_{i^*j}) \leftarrow \widehat{f}_{i^*j}^{(r-1)}(X_j, \widehat{\theta}_{i^*j})\ \forall j$;

$\forall j$ fit the model $f_{i^*j}^{(r)}(X_j, \widehat{\theta}_{i^*j})$ to:

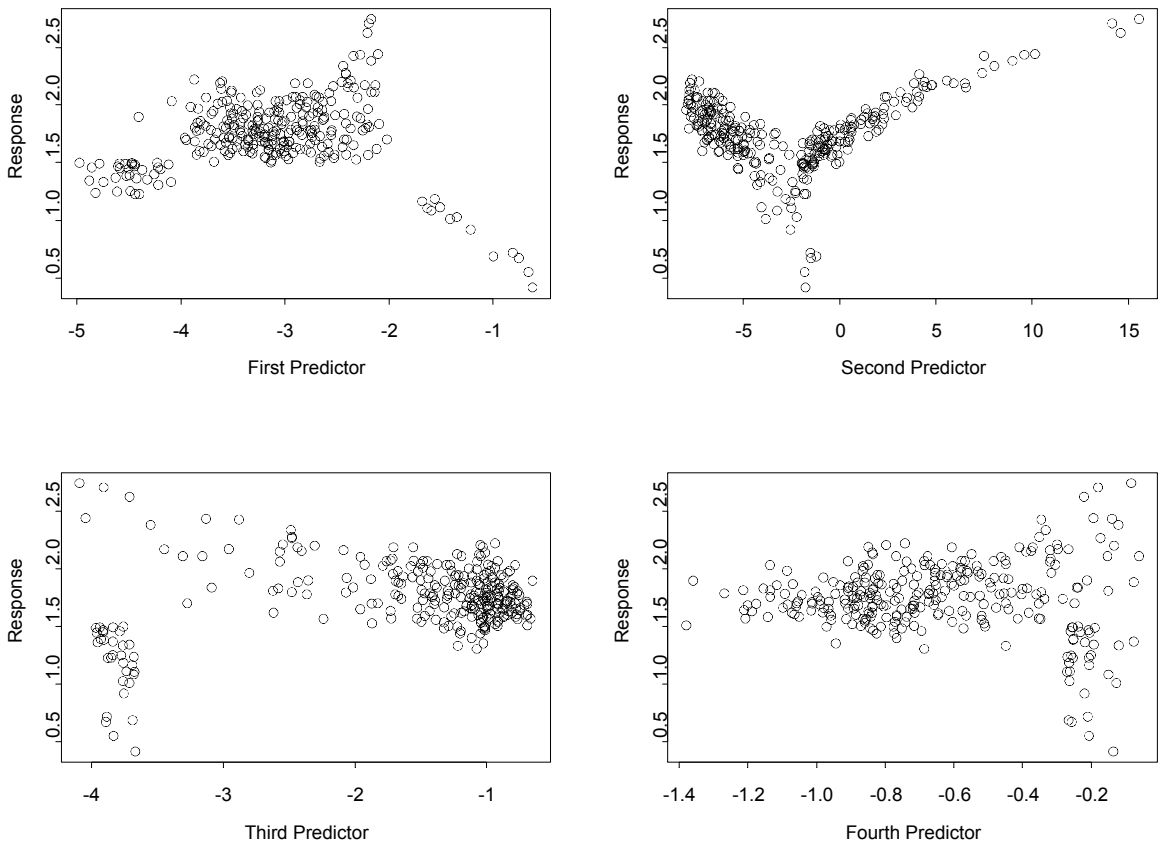
$$e_j{}^{(r)} = \{Y - \sum_{s \neq j} \sum_i \delta_{i^*s} \widehat{f}_{i^*s}^{(r)}(X_s, \widehat{\theta}_{i^*s})\}$$

yielding to $\widehat{f}_{i^*j}^{(r)}(X_j, \widehat{\theta}_{ij})$.

GAM-M (Some Examples)

Numeric Response

1) Simulation Study: *n=500; d=4.*

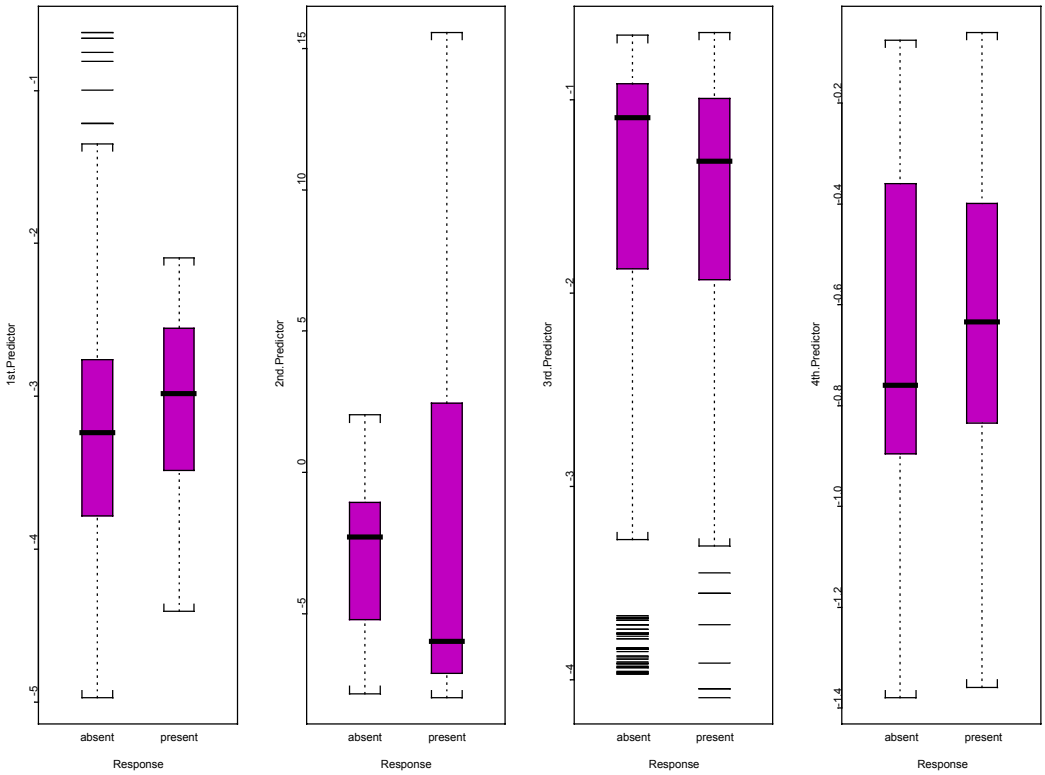


2) `Auto-Mpg` Data Set: *n=398; d=8.*

GAM-M (Some Examples)

Categorical Response

1) Simulation Study: *n=500; d=4.*



2) `Bupa` Data Set: *n=345; d=7.*

A Comparison with Standard GAM

1) Numeric Response Case:

Decrease in the Residual Deviance

	Simulated Data		Auto-Mpg	
Iteration	GAM	GAM-M	GAM	GAM-M
1	7,22	8,10	8,44	9,96
2	5,63	5,16	8,05	9,01
3	5,21	4,25	7,63	8,45
4	5,01	3,49	7,54	7,95
5	4,86	2,49	7,47	7,13
6	4,75	2,10	7,35	6,46
7	4,66	1,87	7,27	5,93
8	4,58	1,41	7,23	5,27
9	4,51	1,25	7,19	4,98
10	4,45	0,94	7,16	4,63

2) Categorical Response Case:

Log-likelihood Statistics and Miss-classification Error

Simulated Data			
Model	**Iteration**	**Log-likelihood**	**Miss-Classification Error**
GAM	1	385	0,198
GAM-M	1	242	0,036
GAM	2	332	0,150
GAM-M	2	154	0,016
Bupa Liver Disorder			
Model	**Iteration**	**Log-likelihood**	**Miss-Classification Error**
GAM	1	195	0,136
GAM-M	1	171	0,104
GAM	2	180	0,084
GAM-M	2	126	0,052

GAM-M Drawbacks

- Model Assignment

- Identifiability

- Overfitting

- Inaccuracy

Some Hints

- *Combining Mixtures*
  $\Rightarrow$ to prevent from incorrect model selection

- *Model Fit Scoring*
  $\Rightarrow$ to avoid overfitting and improve identifiability

- *Bootstrap Averaging*
  $\Rightarrow$ to improve accuracy

Generalized Additive Multi-Mixture Model (GAM-MM)

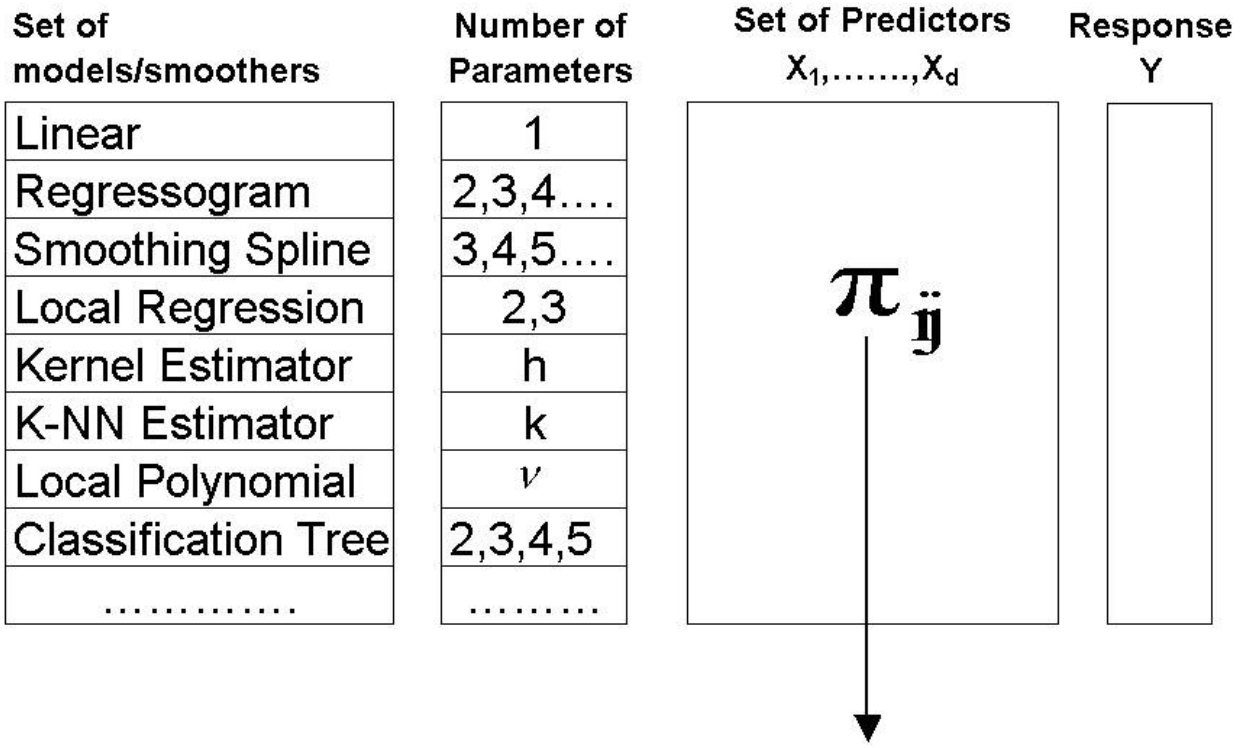
Definition. *Let assume that the response Y has an exponential family density $\rho_Y(y;\theta;\phi)$; the mean $\mu = E(Y|X_1,\ldots,X_d)$ is linked to the predictors via:*

$$g(\mu) = \alpha + \sum_{j=1}^{d}\sum_{i=1}^{K} \pi_{ij} f_{ij}(X_j, \theta_{ij})$$

where:

- ▷ f_{ij} *is a model/smoother;*
- ▷ θ_{ij} *is a vector of parameters;*
- ▷ π_{ij} *is a mixing parameter;*
- ▷ α *is the intercept term.*

The GAM-MM idea



$$g(\mu) = \alpha + \sum_{j=1}^{d} \sum_{i=1}^{K} \pi_{ij} f_{ij}\left(X_j, \theta_{ij}\right)$$

with $\sum_i \pi_{ij} = 1$.

The additive part of the model consists in the sum of $d \times K$ terms.

GAM-MM Estimation Procedure

Basic Steps:

1. *Mixing Parameters Estimation;*

2. *Variable Selection;*

3. *Model Fitting.*

Mixing Parameters Estimation(*via bootstrap*)

Definition. *For the j-th predictor X_j, ($j = 1, \ldots, d$), the mixing parameter estimation resulting from the b-th bootstrap sample is a synthesis of the score reached by the K models in the set $F = \{f_1, \ldots, f_K\}$, namely:*

$$\rho_{i,j,(b)} = \psi\{\phi[f_i(X_j, b)], \nu(f_i)\}$$

such that $\sum_i \rho_{i,j,(b)} = 1$, and:

- $\phi[f_i(X_j, b)]$ *is a non decreasing function, ranging in* $[0, 1]$, *expressing the adequacy of each candidate model to the current predictor;*

- $\nu(f_i)$ *is a decreasing function, ranging in* $[0, 1]$, *expressing the cost for the non complexity induced by the i-th model;*

- $\psi\{\phi[f_i(X_j, b)], \nu(f_i)\}$ *is a normalized scoring function depending upon a combination of $\phi(.)$ and $\nu(.)$.*

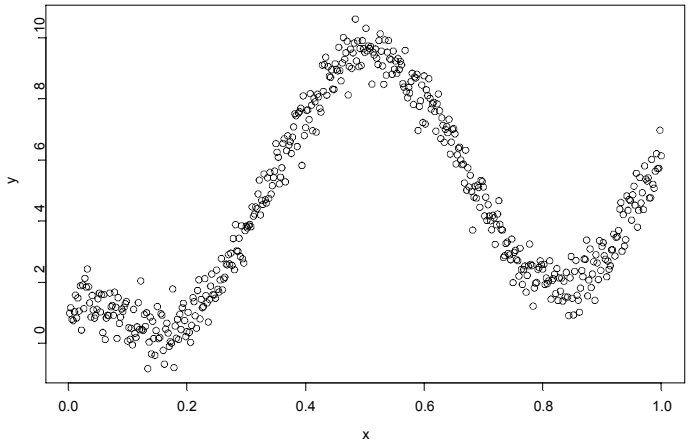
An Example of Scoring-Link Function

Definition. *Let indicate with* $RSS[f_i(X_j, b)]$ *and* $RSS[f_1(X_j, b)]$ *the g.o.f. measures provided by the* i*-th and the most parsimonious model respectively. Equivalently, the number of parameters of the* i*-th model and the less parsimonious are* $\tau(f_i)$ *and* $\tau(f_K)$ *respectively. The scoring-link function is:*

$$\rho_{i,j,(b)} = \left\{1 - \frac{RSS[f_i(X_j, b)]}{RSS[f_1(X_j, b)]}\right\}\left[1 - \frac{\tau(f_i)}{\tau(f_K)}\right]$$

As a result, the estimated mixing parameters defined provide a set of scores of the models fitted to each predictor.

Coherence of $\rho_{i,j,(b)}$

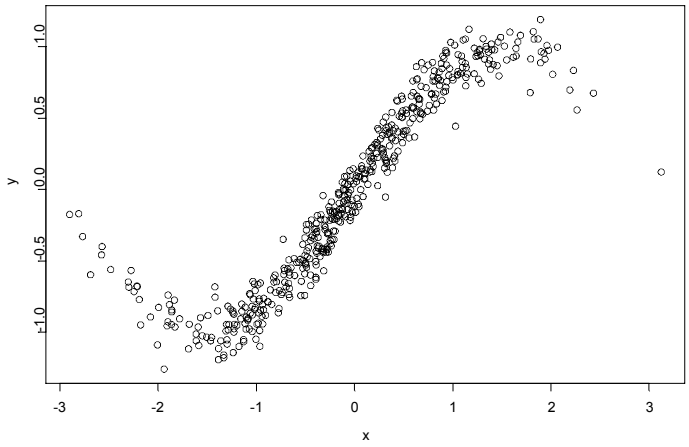


n=500
X~U[0,1]

$y_i = 5–4\ \sin(3\pi\ x_i)+\log(x_i\ \pi)+\varepsilon_i$

ε~N(0,0.5)

<u>selected model</u>: lowess (h=0.5)

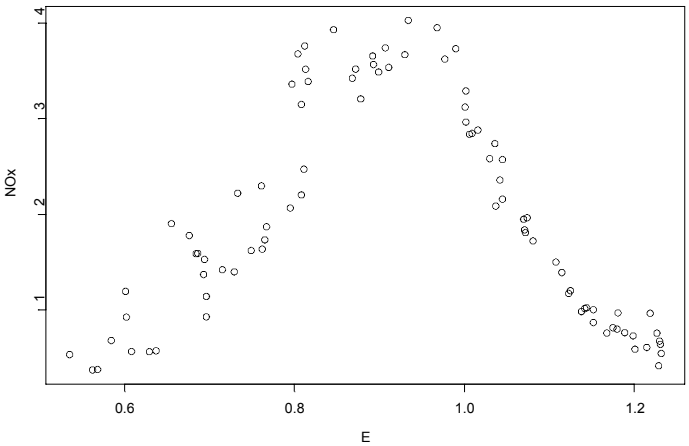


n=500
X~N[0,1]

$y_i = \sin(x_i)+ \varepsilon_i$

ε~N(0,1/9)

<u>selected model</u>: spline (df=4)



Ethanol Data

NOx=Concentration of Nitric Oxide
E=Equivalence Ratio

<u>selected model</u>: lowess (h=0.75)

Coherence of $\rho_{i,j,(b)}$ *(first example)*

Model	τ_i	RSS_i	$v(f_i)$	$\phi(.)$	ρ_i *(%)*
intercept	1	5263.7	0.93	0.00	0.00
linear	2	4257.5	0.86	0.19	2.04
tree(2)	2	2570.9	0.86	0.51	5.45
tree(3)	3	2507.7	0.78	0.52	5.12
lowess(.66,1)	3	1287.4	0.78	0.75	7.38
lowess(.75,1)	3	1494.0	0.78	0.72	7.00
tree(4)	4	902.0	0.71	0.83	7.36
spline(3)	4	1119.9	0.71	0.79	6.99
lowess(.5,1)	**4**	**704.4**	**0.71**	**0.87**	**7.69**
spline(4)	5	529.0	0.64	0.90	7.19
lowess(.66,2)	5	293.7	0.64	0.94	7.55
lowess(.75,2)	5	577.2	0.64	0.89	7.12
spline(5)	6	256.8	0.57	0.95	6.76
lowess(.33,1)	6	237.6	0.57	0.95	6.78
lowess(.5,2)	7	138.7	0.50	0.97	6.05
lowess(.25,1)	8	157.1	0.43	0.97	5.17
lowess(.33,2)	10	126.3	0.28	0.97	3.47
lowess(.25,2)	13	121.5	0.07	0.97	0.87

Bagging Estimation

Definition. *The most appropriate model for* X_j *in the* b*-th bootstrap sample is such that:*

$$i^* = \arg\max[\rho_{i,j,(b)}]$$

deriving from the indicator variable $\delta_{i,j,(b)} = 1$ *if* $i = i^*$ *and zero otherwise.*

Defining a weighting system w_{bj} *for the* B *bootstrap samples, the* **bagging estimation** *of the mixing parameter is:*

$$\pi_{i,j} = \sum_{b=1}^{B} \delta_{i*,j,(b)} w_{bj}$$

Estimation Procedures:
- Majority Vote (*GAM-M);*
- *Voting (GAM-MM*$_v$*);*
- *Weighted Average Scoring (GAM-MM*$_{was}$*).*

Ordering Predictors

Basic Steps

1. *predictors are ordered according to the values of the mixing parameters;*

2. *in each loop of the first iteration of the backfitting algorithm, evaluate the gain in the model fitting provided by the s-th predictor (with the associated mixture) on the partial residuals of the model with $s-1$ predictors, by:*

$$\arg\max_{s} \left\{ C[X_{(s)}] = \frac{[RSS_{(s-1)} - RSS_{(s)}]/RSS_{(s-1)}}{[g_{(s-1)} - g_{(s)}]/g_{(s-1)}} \right\}$$

3. *Repeat step 2 for $s = 1, \ldots, d$ until defining the additive combination of mixtures $\sum_s \sum_i \pi_{i,s} \widehat{f}_{i,s}(X_s, \widehat{\theta}_{i,s})$.*

A financial application

Data: Daily Returns of MIB30 Index and its components from 3/1/96 to 12/5/98 (594 obs.).

Aim: Index Tracking.

Selected Equities:

Banking: Mediobanca, Unicredit;

Industrial: Fiat, Pirelli;

Insurance: Generali, Ras;

Tmt: Telecom, Tim;

Utilities: Edison, Italgas.

Assumptions:

▷ 30 day average trading volume of portfolio components influences 30 day MIB30 historical volatility;

▷ Portfolio Weights are proportional to the decrease in the Total RSS.

Estimation of the Mixing Parameters

Model	Unicredit	Edison	Fiat	Generali	Italgas	M.banca	Pirelli	Ras	Telecom	Tim
intercept	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
linear	**0.80**	0.00	0.00	**0.90**	**0.62**	**0.14**	0.00	**0.36**	**0.89**	0.00
tree(2)	**0.13**	**0.36**	**0.06**	**0.10**	**0.34**	**0.86**	**0.38**	**0.54**	**0.07**	0.00
tree(3)	**0.07**	**0.01**	**0.02**	0.00	**0.04**	0.00	**0.13**	**0.10**	**0.04**	0.00
tree(4)	0.00	**0.65**	**0.02**	0.00	0.00	0.00	**0.49**	0.00	0.00	**1.00**
spline(3)	0.00	0.00	**0.71**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
spline(4)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
spline(5)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lowess(.25,1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lowess(.33,1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lowess(.5,1)	0.00	0.00	**0.04**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lowess(.66,1)	0.00	0.00	**0.15**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lowess(.75,1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lowess(.25,2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lowess(.33,2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lowess(.5,2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lowess(.66,2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lowess(.75,2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Selected Portfolio

d	*RSS(s)*	*g(s)*	*C(Xs)*	Weights		△
Intercept	52,28	1,00	-	GAM-MM Port.	Buy&Hold Port.	Weights
Pirelli	44,33	3,11	0,49	38,80%	3,53%	35,3%
Ras	40,04	2,10	0,19	20,94%	3,37%	17,6%
Italgas	38,36	2,04	0,13	8,20%	1,70%	6,5%
Mediobanca	37,07	2,00	0,14	6,30%	4,38%	1,9%
Tim	35,08	4,00	0,14	9,72%	22,46%	-12,7%
Edison	33,86	3,35	0,17	5,95%	3,21%	2,7%
Fiat	32,6	3,86	0,19	6,15%	9,72%	-3,6%
Unicredit	31,99	2,07	0,21	2,98%	9,21%	-6,2%
Telecom	31,86	2,04	0,05	0,63%	22,42%	-21,8%
Generali	31,79	2,00	0,03	0,34%	20,00%	-19,7%
				100%	100%	

A financial application (II)

Aim: estimate and predict levels of US$/EUR Exchange Rate

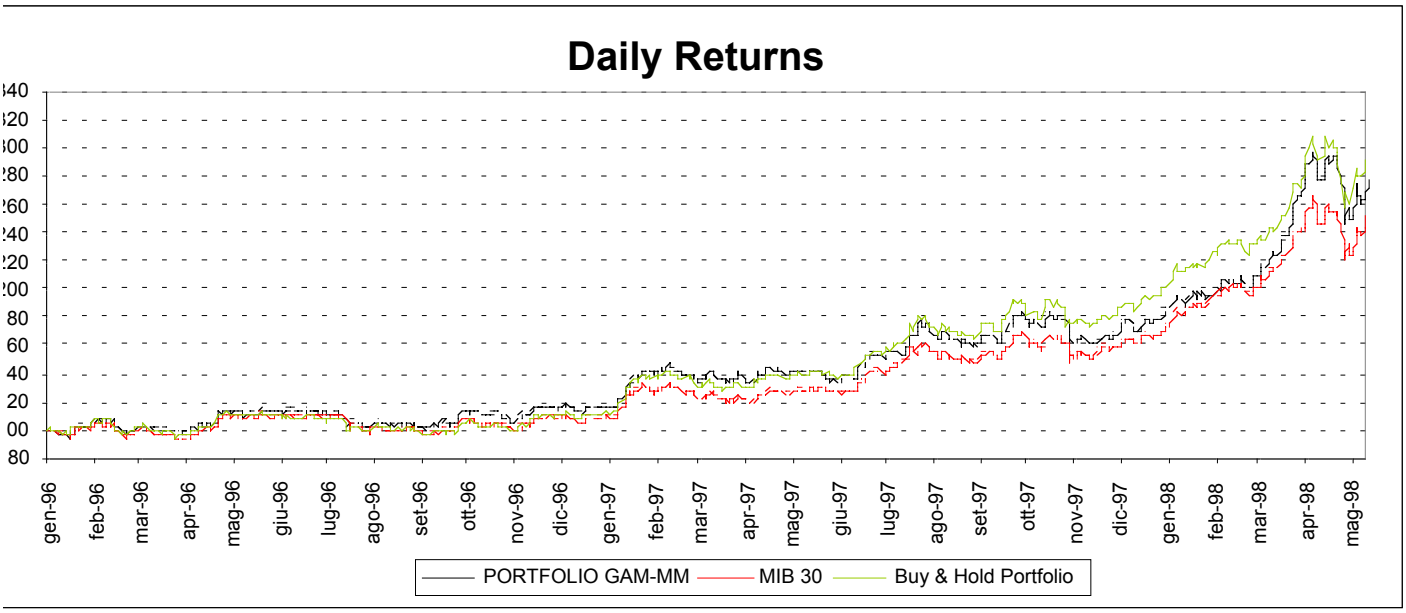
Data: Daily levels of the following variables:

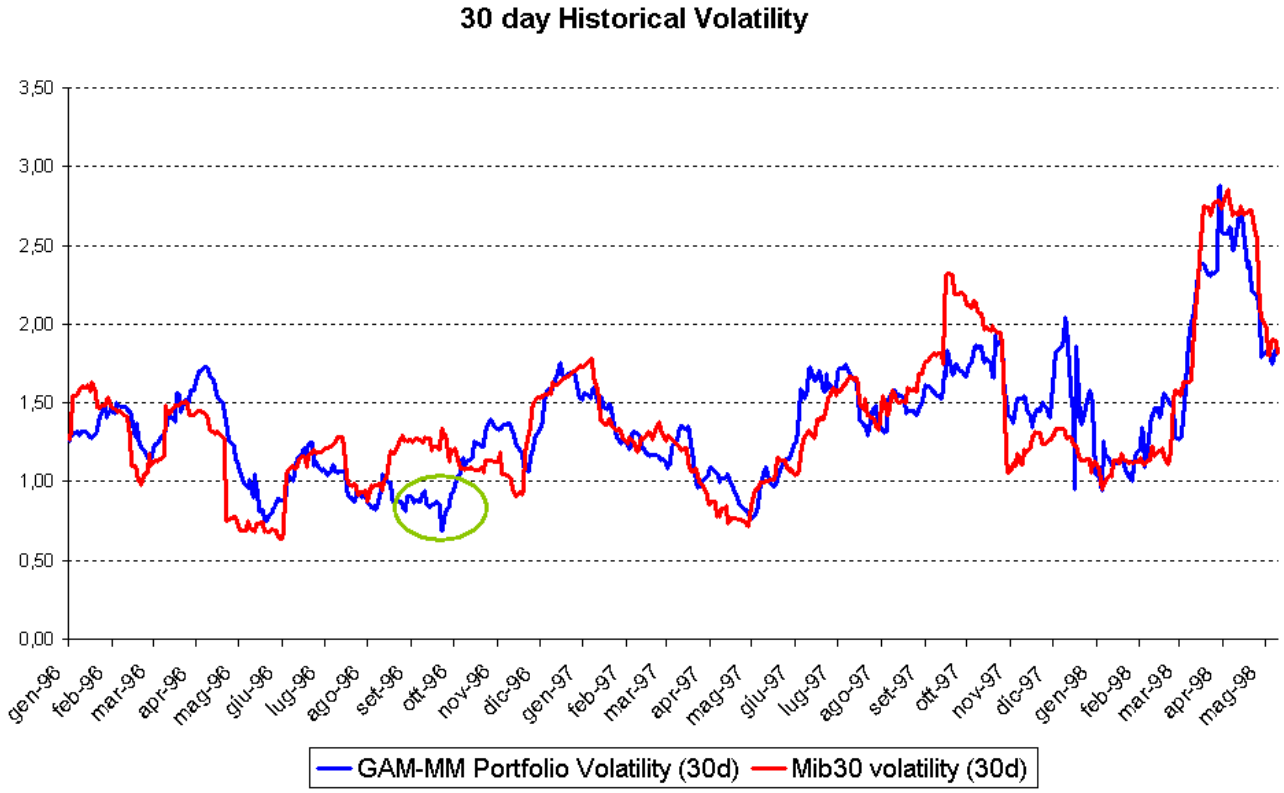
- `exr`: US$/EUR Exchange Rate;
- `3m`: DEM Libor Fixing (3 months);
- `y1`: German Government Bond (1 year);
- `y5`: German Government Bond (5 years);
- `y10`: German Government Bond (10 years);
- `rr1`: Risk Reversal (1 month);
- `rr3`: Risk Reversal (3 months);
- `ic`: Industrial Confidence (Euro Zone);
- `ov`: EUR Overnight Fixing;

observed from 31/12/98 to 24/12/99 (255 obs.):

Comparing Portfolios

	Returns	Volatility	Downside Risk	Max Gain	Max Loss
GAM-MM PORTFOLIO	178,6%	1,4257	89,6	5,8%	-5,2%
MIB 30	150,6%	1,5861	97,4	5,3%	-8,4%
BUY & HOLD PORTFOLIO	186,3%	1,4856	99,3	6,3%	-6,9%





US$/EUR Exchange Rate estimation and prediction

Selected Predictors and Residual Sum of Squares

GAM-MM			GAM		
d	$g_{(s)}$	*RSS*	*d*	$g_{(s)}$	*RSS*
ic	1,02	4,79	y1	1,00	7,52
y10	1,00	3,58	lo(y10)	3,50	0,06
rr1	1,13	3,06	rr3	1,00	4,36
y1	1,25	2,24	ic	1,00	4,28

Forecast Error for GAM and GAM-MM

	GAM-MM		GAM	
	Mean	*Standard Error*	*Mean*	*Standard Error*
t+1	0,174	0,363	0,234	0,475
t+2	0,263	0,443	0,459	0,612
t+3	0,581	0,648	0,648	0,864
t+4	0,669	0,874	0,756	1,005
t+5	0,779	1,077	0,917	1,441

A PMA Strategy for Data Mining

Basic Steps

1. **P***artitioning*
   Use decision trees to find internally homogeneous samples, defining a set of Assignment Rules such that:

$$d(\mathbf{X}) = Y$$

2. **M***odelling*
   Model each sample h with GAM-MM:

$$g(\mu) = \alpha_h + \sum_{j=1}^{d_h} \sum_{i=1}^{K} \pi_{ij} f_{ij}(X_j, \theta_{ij})$$

   for $h = 1, \ldots, H$

3. **A***malgamation*
   Construct a set of Complex Decision Rules:

$$g(\mu) = \sum_{h=1}^{H} \delta_h g(\mu_h)$$

   for $h = 1, \ldots, H$

A Data Mining Application

The Data:
"Adult" Data Set (US Census Bureau)
$www.ics.uci.edu\backslash \sim mlearn$

Response Variable:
Salary (greater or less than $50,000)

Attributes:
10 Socio-demographic indicators (age, education, race, capital loss, capital gain, etc.), 5 continuous and 5 nominal.

Number of Instances:
48842 (30725 without missing values).

Partitioning Summary

Exploratory Tree:
- Learning Sample: 20505 units
- Test Sample: 10840 units (33.33%)

Response Variable:
- Class 1: Salary $> 50,000$ (24.78%)
- Class 2: Salary $\leq 50,000$ (75.22%)

Number of final samples: 31

Number of amalgamated samples: 11

Results of the PMA Approach

Final Error Rate in the Amalgamated Samples

Node Number	*Number of Istances*	*Error Rate (Tree) %*	*Error Rate (GAM) %*	*Error Rate (GAM-MM) %*
3	5853	12.03	12.03	0.53
5	1166	0.94	0.94	0.94
9	744	14.24	14.24	5.64
17	4630	47.32	36.69	.3.04
64	508	18.50	18.50	2.56
67	4245	41.27	36.56	1.18
130	158	0.05	0,05	0,05
133	579	14.68	14.68	0.20
1051	225	41.78	39.11	5.49
1053	138	13.04	13.04	13.04
4216	1206	27.61	27.61	2.74

Estimated Mixing Parameters in the Sample n.9

Models or Smoothers	***Predictors***						
	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Intercept	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Linear	0,14	0,15	0,23	0,14	0,12	**0,22**	0,11
Tree(2)	0,03	0,01	0,00	0,08	0,00	0,01	0,07
Tree(3)	0,16	0,15	0,17	0,15	0,14	0,16	0,12
Spline(3)	0,08	0,09	0,03	0,09	0,08	0,06	0,10
Tree(4)	**0,17**	**0,17**	**0,26**	**0,17**	**0,19**	0,20	**0,15**
Spline(4)	0,08	0,08	0,03	0,07	0,09	0,06	0,10
Tree(5)	0,14	0,15	0,21	0,14	0,16	0,15	0,13
Spline(5)	0,08	0,06	0,02	0,06	0,08	0,06	0,09
Lowess	0,09	0,09	0,02	0,07	0,10	0,04	0,10
Lowess(2)	0,03	0,03	0,02	0,02	0,03	0,03	0,03

Recent papers

Conversano, C., Mola, F., Siciliano, R. (2001). *Partitioning Algorthm and Combined Model Integration for Data Mining,* Computational Statistics, **16**, 3, 323-339.

Conversano, C., (2001). "Generalized Additive Multi-Mixture Models: Scoring Models and Predictors", in Klein B. e Korsholm (Eds) *New Trends in Statistical Modelling, Proceedings of the 16th International Workshop on Statistical Modelling,* University of Southern Denmark, pag. 103-110.

Conversano, C., Siciliano, R., Mola, F. (2001). "Generalized Additive Multi-Mixture Models for Data Mining", *Computational Statistics and Data Analysis,* in press.

Conversano, C. (2000). *Combining Mixture of Models through Model Scoring Criteria and Bootstrap Averaging,* Technical Report.

Conversano, C., Mola, F., Siciliano, R., (2000). "Generalized Additive Multi-Model for Classification and Prediction", in Kiers, H.A.L., Rasson, J.P., Groenen, P.J.F., Schader, M. (eds.): *Data Analysis, Classification, and Related Methods,* Springer, Berlin, 205-210.

Conversano, C., Siciliano, R., Mola, F. (2000). "Supervised Classifier Combination Through Generalized Additive Multi-Model", in Kittler, J., Roli, F. (eds.): *Proceedings of the First International Workshop on Multiple Classifier Systems,* Lecture Notes in Computer Science, Springer Verlag, 167-176.