A case study on using generalized additive models to fit credit rating scores

Marlene Müller, marlene.mueller@itwm.fraunhofer.de

This version: July 8, 2009, 14:32

Fraunhofer ITWM

Contents

Application: Credit Rating

- Basel II: capital requirements of a bank are adapted to the individual credit portfolio
- core terms: determine rating score and subsequently default probabilities (PDs) as a function of some explanatory variables
- further terms: loss given default, portfolio dependence structure
- in practice: often classical logit/probit-type models to estimate linear predictors (scores) and probabilities (PDs)
- statistically: 2-group classification problem

risk management issues

- credit risk is ony one part of a bank's total risk:
  $\leadsto$ will be aggregated with other risks
- credit risk estimation from historical data:
  $\leadsto$ stress-tests to simulate future extreme situations
  $\leadsto$ need to easily adapt the rating system to possible future changes
  $\leadsto$ possible need to extrapolate to segments without observations

Application: Credit Rating

- Basel II: capital requirements of a bank are adapted to the individual credit portfolio
- core terms: determine rating score and subsequently default probabilities (PDs) as a function of some explanatory variables
- further terms: loss given default, portfolio dependence structure
- in practice: often classical logit/probit-type models to estimate linear predictors (scores) and probabilities (PDs)
- statistically: 2-group classification problem

risk management issues

- credit risk is ony one part of a bank's total risk:
  $\rightsquigarrow$ will be aggregated with other risks
- credit risk estimation from historical data:
  $\rightsquigarrow$ stress-tests to simulate future extreme situations
  $\rightsquigarrow$ need to easily adapt the rating system to possible future changes
  $\rightsquigarrow$ possible need to extrapolate to segments without observations

(Simplified) Development of Rating Score and Default Probability

- raw data:

  X_j measurements of several variables ("risk factors")

- (nonlinear) transformation:

$$X_j \to \widetilde{X}_j = m_j(X_j)$$

  $\leadsto$ handle outliers, allow for nonlinear dependence on raw risk factors

- rating score:

$$S = w_1 \widetilde{X}_1 + \ldots + w_d \widetilde{X}_d$$

- default probability:

$$PD = P(Y = 1 | \boldsymbol{X}) = G(w_1 \widetilde{X}_1 + \ldots + w_d \widetilde{X}_d)$$

  (where G is e.g. the logistic or gaussian cdf $\leadsto$ logit or probit)

(Simplified) Development of Rating Score and Default Probability

- raw data:

  X_j measurements of several variables ("risk factors")

- (nonlinear) transformation:

$$X_j \to \widetilde{X}_j = m_j(X_j)$$

  $\rightsquigarrow$ handle outliers, allow for nonlinear dependence on raw risk factors

- rating score:

$$S = w_1 \widetilde{X}_1 + \ldots + w_d \widetilde{X}_d$$

- default probability:

$$PD = P(Y = 1 | \boldsymbol{X}) = G(w_1 \widetilde{X}_1 + \ldots + w_d \widetilde{X}_d)$$

  (where G is e.g. the logistic or gaussian cdf $\rightsquigarrow$ logit or probit)

(Simplified) Development of Rating Score and Default Probability

- raw data:

  X_j measurements of several variables ("risk factors")

- (nonlinear) transformation:

$$X_j \to \widetilde{X}_j = m_j(X_j)$$

  $\leadsto$ handle outliers, allow for nonlinear dependence on raw risk factors

- rating score:

$$S = w_1 \widetilde{X}_1 + \ldots + w_d \widetilde{X}_d$$

- default probability:

$$PD = P(Y = 1|X) = G(w_1 \widetilde{X}_1 + \ldots + w_d \widetilde{X}_d)$$

  (where G is e.g. the logistic or gaussian cdf $\leadsto$ logit or probit)

(Simplified) Development of Rating Score and Default Probability

- raw data:

  X_j measurements of several variables ("risk factors")

- (nonlinear) transformation:

$$X_j \to \widetilde{X}_j = m_j(X_j)$$

  $\rightsquigarrow$ handle outliers, allow for nonlinear dependence on raw risk factors

- rating score:

$$S = w_1 \widetilde{X}_1 + \ldots + w_d \widetilde{X}_d$$

- default probability:

$$PD = P(Y = 1|X) = G(w_1 \widetilde{X}_1 + \ldots + w_d \widetilde{X}_d)$$

  (where G is e.g. the logistic or gaussian cdf $\rightsquigarrow$ logit or probit)

Fraunhofer ITWM

(Simplified) Development of Rating Score and Default Probability

- raw data:

  X_j measurements of several variables ("risk factors")

- (nonlinear) transformation:

$$X_j \to \widetilde{X}_j = m_j(X_j)$$

  $\rightsquigarrow$ handle outliers, allow for nonlinear dependence on raw risk factors

- rating score:

$$S = w_1 \widetilde{X}_1 + \ldots + w_d \widetilde{X}_d$$

- default probability:

$$PD = P(Y = 1|\boldsymbol{X}) = G(w_1 \widetilde{X}_1 + \ldots + w_d \widetilde{X}_d)$$

  (where G is e.g. the logistic or gaussian cdf $\rightsquigarrow$ logit or probit)

Aim of this Talk

case study on (cross-sectional) rating data

- compare different approaches to generalized additive models (GAM)
- consider models that allow for additional categorical variables
  $\rightsquigarrow$ partial linear terms (combination of GAM/GPLM)

► generalized additive models allow for a simultaneous fit of the transformations from the raw data, the linear rating score and the default probabilities

Outline of the Study

- credit data case study: 4 credit datasets

dataset	sample	defaults	regressors continuous	regressors discrete	regressors categorical
German Credit	1000	30.00%	3	–	17
Australian Credit	678	55.90%	3	1	8
French Credit	8178	5.86%	5	3	15
UC2005 Credit	5058	23.92%	12	3	21

  - differences between different approaches?
  - improvement of default predictions?

- simulation study: comparison of additive model (AM) and GAM fits

  - differences between different approaches?
  - what if regressors are concurve? (nonlinear version of multicollinear)
  - do sample size and default rate matter?

Generalized Additive Model

- logit/probit are special cases of the generalized linear model (GLM)

$$E(Y|\boldsymbol{X}) = G\left(\boldsymbol{X}^\top \beta\right)$$

- "classic" generalized additive model

$$E(Y|\boldsymbol{X}) = G\left\{c + \sum_{j=1}^{p} m_j(X_j)\right\} \quad m_j \text{ nonparametric}$$

- generalized additive partial linear model (semiparametric GAM)

$$E(Y|\boldsymbol{X}_1, \boldsymbol{X}_2) = G\left\{c + \boldsymbol{X}_1^\top \beta + \sum_{j=1}^{p} m_j(X_{2j})\right\} \quad m_j \text{ nonparametric}$$

linear part

- allows for known transformation functions
- allows to add / control for categorical regressors

Generalized Additive Model

- logit/probit are special cases of the generalized linear model (GLM)

$$E(Y|\boldsymbol{X}) = G\left(\boldsymbol{X}^\top \beta\right)$$

- "classic" generalized additive model

$$E(Y|\boldsymbol{X}) = G\left\{c + \sum_{j=1}^{p} m_j(X_j)\right\} \quad m_j \text{ nonparametric}$$

- generalized additive partial linear model (semiparametric GAM)

$$E(Y|\boldsymbol{X}_1, \boldsymbol{X}_2) = G\left\{c + \boldsymbol{X}_1^\top \beta + \sum_{j=1}^{p} m_j(X_{2j})\right\} \quad m_j \text{ nonparametric}$$

linear part

- allows for known transformation functions
- allows to add / control for categorical regressors

Generalized Additive Model

- logit/probit are special cases of the generalized linear model (GLM)

$$E(Y|\boldsymbol{X}) = G\left(\boldsymbol{X}^\top \beta\right)$$

- "classic" generalized additive model

$$E(Y|\boldsymbol{X}) = G\left\{c + \sum_{j=1}^{p} m_j(X_j)\right\} \quad m_j \text{ nonparametric}$$

- generalized additive partial linear model (semiparametric GAM)

$$E(Y|\boldsymbol{X}_1, \boldsymbol{X}_2) = G\left\{c + \boldsymbol{X}_1^\top \beta + \sum_{j=1}^{p} m_j(X_{2j})\right\} \quad m_j \text{ nonparametric}$$

linear part

- allows for known transformation functions
- allows to add / control for categorical regressors

R "Standard" Tools

two main approaches for GAM in R

- **gam::gam** $\rightsquigarrow$ backfitting with local scoring (Hastie and Tibshirani; 1990)
- **mgcv::gam** $\rightsquigarrow$ penalized regression splines (Wood; 2006)

$\rightsquigarrow$ compare these procedures under the default settings of gam::gam and mgcv::gam

competing estimators:

- **logit** binary GLM with $G(u) = 1/\{1 + \exp(-u)\}$ (logistic cdf as link)
- **logit2**, **logit3** binary GLM with 2nd / 3rd order polynomial terms for the continuous regressors
- **logitc** binary GLM with continuous regressors categorized (4–5 levels)
- **gam** binary GAM using gam::gam with $s()$ terms for continuous
- **mgcv** binary GAM using mgcv::gam

German Credit Data

- from http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html

dataset name	sample	defaults	regressors continuous	regressors discrete	regressors categorical
German	1000	30.00%	3	–	17

- 3 continuous regressors: age, amount, duration (time to maturity)
- use 10 CV subsamples for validation
- stratified data (true default rate $\approx$ 5%)

- important findings:
  - some observation(s) that seem to confuse mgcv::gam in one CV subsample ($\longrightarrow$ see following slides)
  - however, mgcv::gam seems to improve deviance and discriminatory power w.r.t. gam::gam
  - estimation times of mgcv::gam are between **4** to **7** times higher than for gam::gam (not more than around a second, though)
  - if we only use the continuous regressors: both GAM estimators are comparable to logit cubic additive functions

German Credit Data

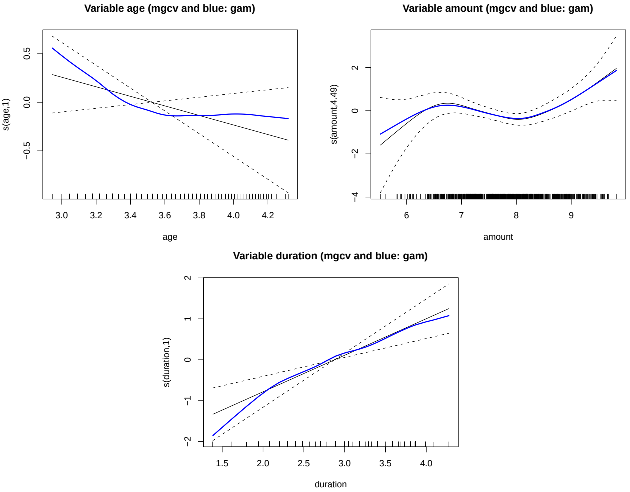
- from http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html

dataset name	sample	defaults	regressors continuous	regressors discrete	regressors categorical
German	1000	30.00%	3	–	17

- 3 continuous regressors: age, amount, duration (time to maturity)
- use 10 CV subsamples for validation
- stratified data (true default rate $\approx$ 5%)

- important findings:
  - some observation(s) that seem to confuse mgcv::gam in one CV subsample (→ see following slides)
  - however, mgcv::gam seems to improve deviance and discriminatory power w.r.t. gam::gam
  - estimation times of mgcv::gam are between **4** to **7** times higher than for gam::gam (not more than around a second, though)
  - if we only use the continuous regressors: both GAM estimators are comparable to logit cubic additive functions

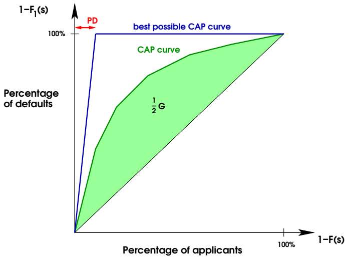
German Credit Data: Additive Functions



How to Compare Binary GLM Fits?

- preferably by out-of-sample validation $\rightsquigarrow$ block cross-validation approach: leave out subsamples of x% from the fitting procedure, estimate from the remaining (100-x)% and calculate validation criteria from the x% left-out
- two criteria for comparison: deviance ($\rightarrow$ goodness of fit) and accuracy ratios AR from CAP curves ($\rightarrow$ discriminatory power)
- CAP curve (Lorenz curve) and the accuracy ratio AR:
  - plot the empirical cdf of the fitted scores against the empirical cdf of the fitted default sample scores (precisely $1 - \widehat{F}$ vs. $1 - \widehat{F}(.|Y = 1)$)
  - AR is the area between CAP curve and diagonal in relation to the corresponding area for the best possible CAP curve (best possible $\cong$ perfect separation)
  - relation to ROC: compares $\widehat{F}(.|Y = 0)$ and $\widehat{F}(.|Y = 1)$ and it holds

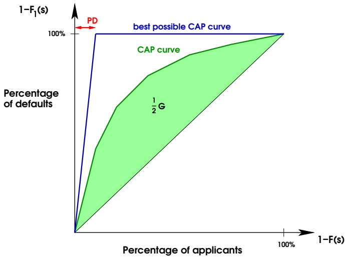
$$AR = 2\,AUC - 1$$



How to Compare Binary GLM Fits?

- preferably by out-of-sample validation $\rightsquigarrow$ block cross-validation approach: leave out subsamples of x% from the fitting procedure, estimate from the remaining (100-x)% and calculate validation criteria from the x% left-out
- two criteria for comparison: deviance ($\rightarrow$ goodness of fit) and accuracy ratios AR from CAP curves ($\rightarrow$ discriminatory power)
- CAP curve (Lorenz curve) and the accuracy ratio AR:
  - plot the empirical cdf of the fitted scores against the empirical cdf of the fitted default sample scores (precisely $1 - \widehat{F}$ vs. $1 - \widehat{F}(.|Y = 1)$)
  - AR is the area between CAP curve and diagonal in relation to the corresponding area for the best possible CAP curve (best possible $\cong$ perfect separation)
  - relation to ROC: compares $\widehat{F}(.|Y = 0)$ and $\widehat{F}(.|Y = 1)$ and it holds

$$\mathrm{AR} = 2\,\mathrm{AUC} - 1$$



Fraunhofer
ITWM

German Credit Data: Comparison

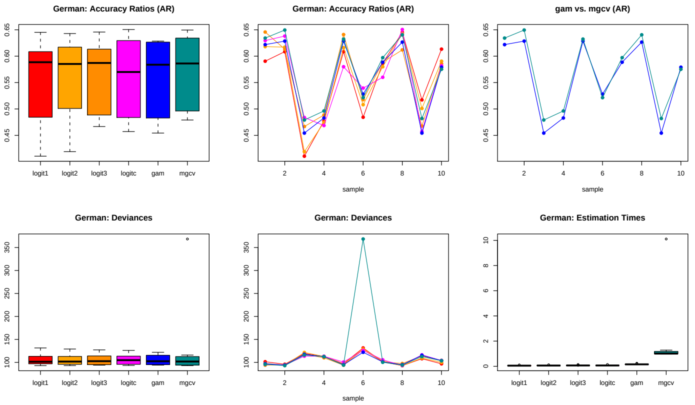


Figure: Out of sample comparison (blockwise CV with 10 blocks) for various estimators, accuracy ratios from CAP curves (upper panels), deviance values and estimation times (lower panels)

Fraunhofer ITWM

German Credit Data: Models with only Continuous Regressors

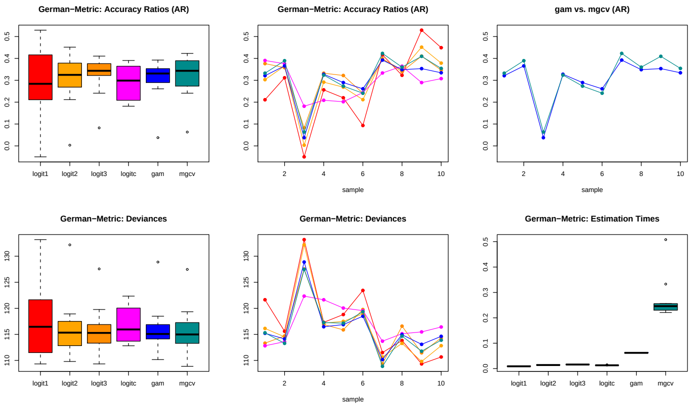


Figure: Out of sample comparison (blockwise CV with 10 blocks) for various estimators, accuracy ratios from CAP curves (upper panels), deviance values and estimation times (lower panels)

Australian Credit Data

- from http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval)
- used for estimation:

dataset name	sample	defaults	regressors continuous	regressors discrete	regressors categorical
Australian	678	55.90%	3	1	8

- use only 7 CV subsamples for validation
- original A13 and A14 were dropped since actually multicollinear with A10, some observations were dropped because of very few categories
- A10 was transformed to $\log(1 + A10)$, nevertheless used only as a linear predictor (as half of the observations have the same value)
- important findings:
  - essentially, the estimated additive function for A2 differs between mgcv::gam and gam::gam
  - gam::gam mostly outperforms than all other estimates (recall, that however the number of CV subsamples is rather small!)
  - estimation times of mgcv::gam are around **3** to **5** times higher than for gam::gam (less than a second, though)

French Credit Data

- data were already analyzed with GPLMs in Müller and Härdle (2003), here used for estimation:

dataset name	sample	defaults	regressors continuous	regressors discrete	regressors categorical
French	8178	5.86%	5	3	15

- use the same preprocessing as in as in Müller and Härdle (2003)
- the original estimation + validation samples were merged, use 20 CV subsamples for validation instead
- continuous variables are X1, X2, X3, X4 and X6, in particular X3, X4 and X6 are known to have nonlinear form in a GAM
- important findings:
  - it is confirmed that additive functions for X3, X4 and X6 should be modelled by a nonlinear function be nonlinear
  - again observation(s) "confusing" mgcv::gam in one of the subsamples
  - all estimates show similar discriminatory power, though with a slightly better performance for both mgcv::gam and gam::gam
  - estimation times of mgcv::gam are around **15** to **24** times higher than for gam::gam (for the largest model: 20-40 sec. on a 3Ghz Intel CPU for the subsamples of about 7800 observations)

UC2005 Credit Data

- data from the 2005 UC data mining competitionwere already analyzed with GPLMs in Müller and Härdle (2003), here used for estimation:

			regressors		
dataset name	sample	defaults	continuous	discrete	categorical
UC2005	5058	23.92%	12	3	21

- the original estimation + validation + quiz samples were merged, use again 20 CV subsamples for validation
- stratified data (true default rate $\approx$ 5%)
- several of the variables have been preprocessed with a log-transform or to binary
- in general, the data haven't been very carefully analysed, it's use is rather meant a "proof-of concept"

- important findings:
  - there are again observations "confusing" mgcv::gam in one of the subsamples
  - performance of mgcv::gam and gam::gam w.r.t. is very similar and outperforms the other approaches (closest to them is the logit fit with cubic additive functions)
  - estimation times of mgcv::gam are around **8** to **40** times higher than for gam::gam (for the largest model: 5-8 min on a 3Ghz Intel CPU for up to 400 seconds for the subsamples of about 4800 observations)

Simulation Study for (G)PLM

$$E(Y|\boldsymbol{X}, T) = \beta_1 X_1 + \beta_2 X_2 + m(T)$$

which of the (G)AM estimators is preferable ...?

- ... to fit the additive component functions and/or the regression function?
- ... w.r.t. discriminatory power in the GPLM/GAM cases?
- ... from a practical point of view (comp. speed, numerical stability etc.)?

simulation setup:

$$\beta_1 = 1, \quad \beta_2 = -1, \quad m(t) = 1.5\cos(\pi t) + c$$

$$X_1, U, T \sim \text{Uniform[-1,1]}, \quad X_2 \sim m\left(\rho T + (1-\rho)U\right) \text{ (centered)}$$

$$n_{sim} = 1000, \quad n \in \{100, 1000, 10000\}, \quad \rho \in \{0.0, 0.7\}, \quad c \in \{0, -1, -2\}$$

- X_2 and T are nonlinearly dependent (if $\rho = 0.7$) or independent otherwise
- sample size n up to 10000 which is a possible size for credit data
- the intercept c controls for the default rate (15%–50%) in the GPLM

Simulation Study for (G)PLM

$$E(Y|\boldsymbol{X},T) = \beta_1 X_1 + \beta_2 X_2 + m(T)$$

which of the (G)AM estimators is preferable ...?

- ... to fit the additive component functions and/or the regression function?
- ... w.r.t. discriminatory power in the GPLM/GAM cases?
- ... from a practical point of view (comp. speed, numerical stability etc.)?

simulation setup:

$$\beta_1 = 1, \quad \beta_2 = -1, \quad m(t) = 1.5\cos(\pi t) + c$$
$$X_1, U, T \sim \text{Uniform[-1,1]}, \quad X_2 \sim m\left(\rho T + (1-\rho)U\right) \text{ (centered)}$$

$$n_{sim} = 1000, \quad n \in \{100, 1000, 10000\}, \quad \rho \in \{0.0, 0.7\}, \quad c \in \{0, -1, -2\}$$

- X_2 and T are nonlinearly dependent (if $\rho = 0.7$) or independent otherwise
- sample size n up to 10000 which is a possible size for credit data
- the intercept c controls for the default rate (15%–50%) in the GPLM

Simulation Study for (G)PLM

$$E(Y|\boldsymbol{X},T) = \beta_1 X_1 + \beta_2 X_2 + m(T)$$

which of the (G)AM estimators is preferable ...?

- ... to fit the additive component functions and/or the regression function?
- ... w.r.t. discriminatory power in the GPLM/GAM cases?
- ... from a practical point of view (comp. speed, numerical stability etc.)?

simulation setup:

$$\beta_1 = 1, \quad \beta_2 = -1, \quad m(t) = 1.5\cos(\pi t) + c$$

$$X_1, U, T \sim \text{Uniform[-1,1]}, \quad X_2 \sim m\left(\rho T + (1-\rho)U\right) \text{ (centered)}$$

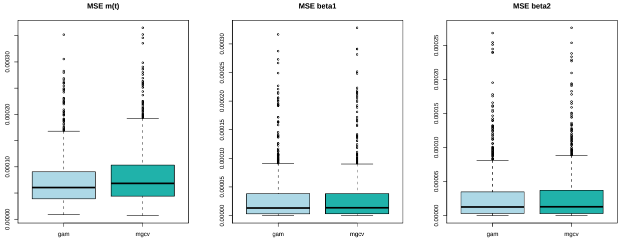
$$n_{sim} = 1000, \quad n \in \{100, 1000, 10000\}, \quad \rho \in \{0.0, 0.7\}, \quad c \in \{0, -1, -2\}$$

- X_2 and T are nonlinearly dependent (if $\rho = 0.7$) or independent otherwise
- sample size n up to 10000 which is a possible size for credit data
- the intercept c controls for the default rate (15%–50%) in the GPLM

Simulation Study: Additive Components for GPLM

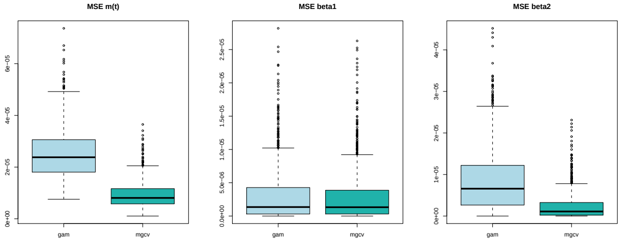
gam
mgcv

$\rho = 0.7$
$c = -2$
$n = 1000$



gam
mgcv

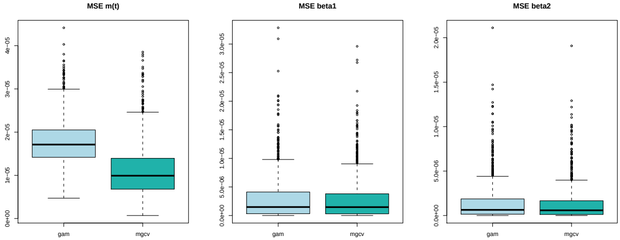
$\rho = 0.7$
$c = -2$
$n = 10000$



Simulation Study: Independent Components vs. Dependent

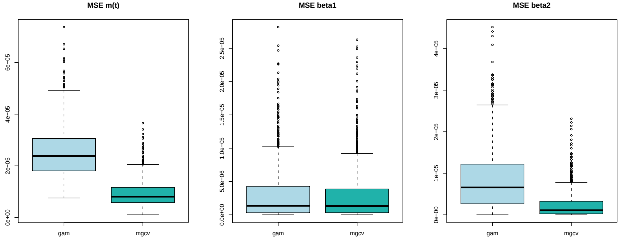
gam
mgcv

$\rho = 0.7$
$c = 0$
$n = 10000$



gam
mgcv

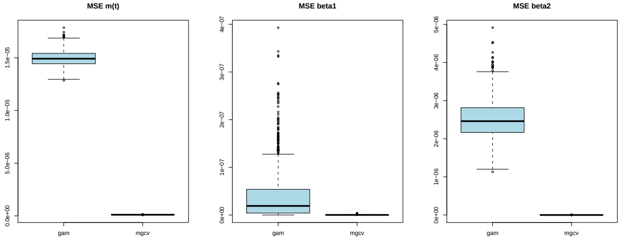
$\rho = 0.7$
$c = -2$
$n = 10000$



Simulation Study: Comparison with Components for PLM

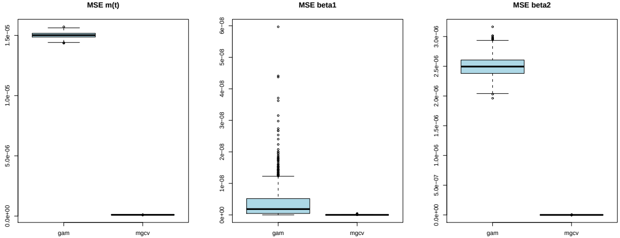
gam
mgcv

$\rho = 0.7$
$c = 0$
$n = 1000$



gam
mgcv

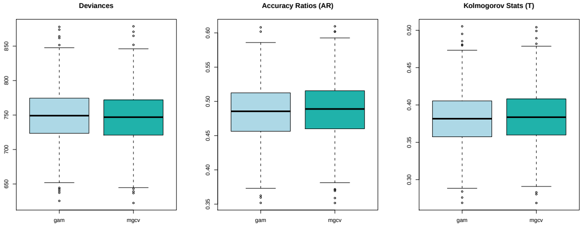
$\rho = 0.7$
$c = 0$
$n = 10000$



Simulation Study: Deviance and Discriminatory Power for GPLM

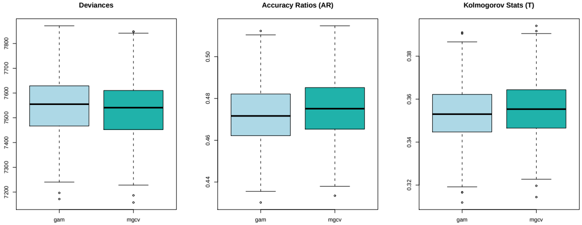
gam
mgcv

$\rho = 0.7$
$c = -2$
$n = 1000$



gam
mgcv

$\rho = 0.7$
$c = -2$
$n = 10000$

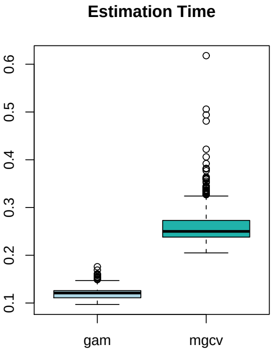


in fact, most of the gam::gam deviances are larger here than the mgcv::gam deviances and gam::gam fits have smaller discriminatory power

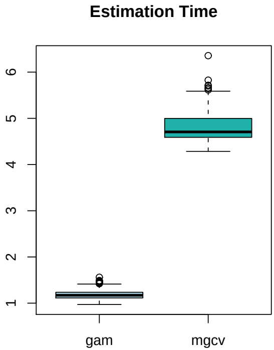
Simulation Study: Estimation Times for GPLM

gam
mgcv

$\rho = 0.7$
$c = -2$



$n = 1000$



$n = 10000$

(estimation times in sec. on a Xeon 2.50GHz)

Fraunhofer ITWM

Conclusions

- typically, categorical regressors improve fit significantly, therefore estimation methods for credit data should adequately use these
- backfitting + local scoring (gam::gam) provides fast and numerically stable results
- there is however clear indication, that penalized regression splines (mgcv::gam) may provide more precise estimates of the additive component functions; its current drawbacks are:
  - estimation time (increasing with model complexity, categorical variables)
  - effects are to be seen only in large samples
- thus: no clear recommendation, no "ultimate method"
  $\rightsquigarrow$ clearly topics for more research

References

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Modeling: An Introduction*, Springer, New York.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

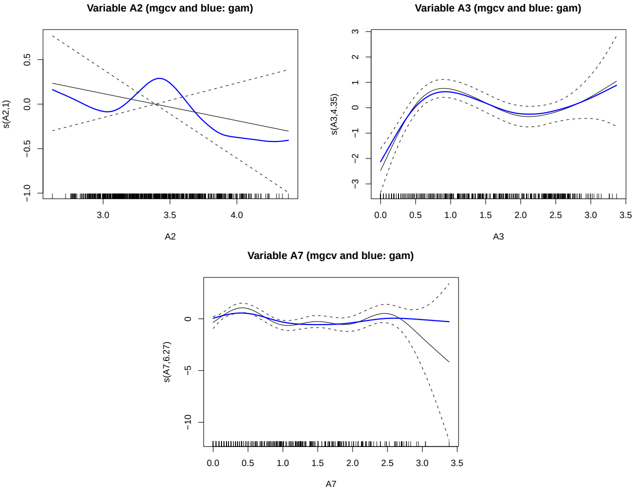
Müller, M. (2001). Estimation and testing in generalized partial linear models — a comparative study, *Statistics and Computing* **11**: 299–309.

Müller, M. and Härdle, W. (2003). Exploring credit data, *in* G. Bol, G. Nakhaeizadeh, S. Rachev, T. Ridder and K.-H. Vollmer (eds), *Credit Risk - Measurement, Evaluation and Management*, Physica-Verlag.

Speckman, P. E. (1988). Regression analysis for partially linear models, *Journal of the Royal Statistical Society, Series B* **50**: 413–436.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*, Texts in Statistical Science, Chapman and Hall, London.

Australian Credit Data: Additive Functions



Australian Credit Data: Comparison

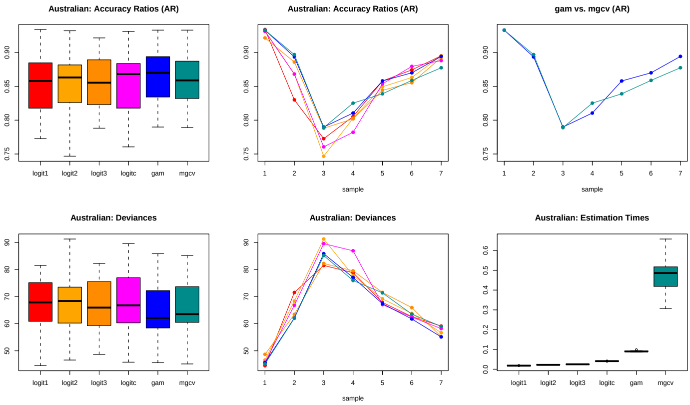


Figure: Out of sample comparison (blockwise CV with 7 blocks) for various estimators, accuracy ratios from CAP curves (upper panels), deviance values and estimation times (lower panels)

Fraunhofer ITWM

Australian Credit Data: Models with only Continuous Regressors

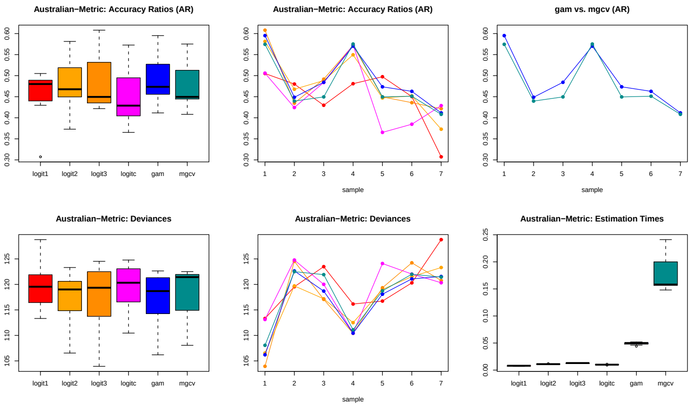
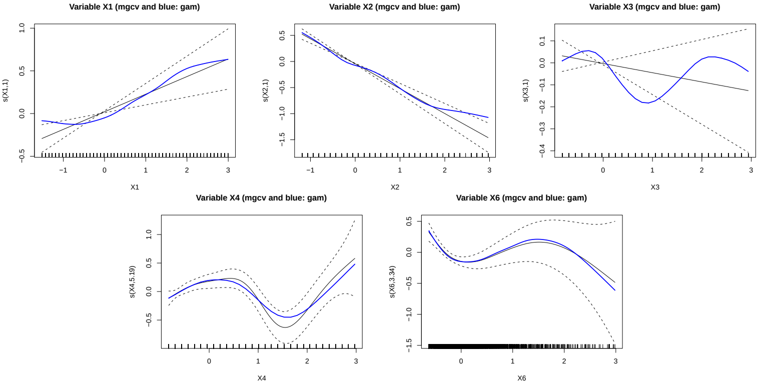


Figure: Out of sample comparison (blockwise CV with 7 blocks) for various estimators, accuracy ratios from CAP curves (upper panels), deviance values and estimation times (lower panels)

French Credit Data: Additive Functions



French Credit Data: Comparison

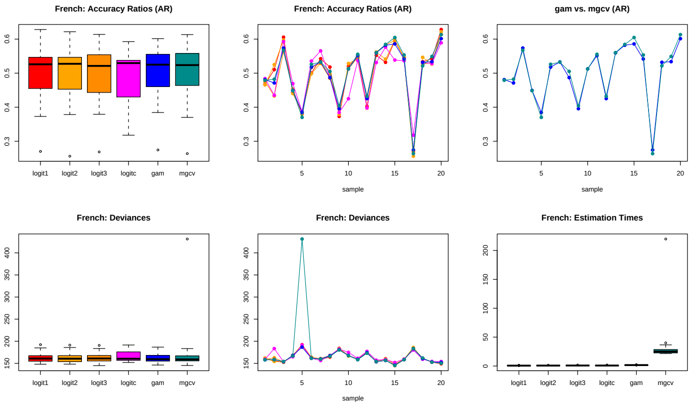


Figure: Out of sample comparison (blockwise CV with 20 blocks) for various estimators, accuracy ratios from CAP curves (upper panels), deviance values and estimation times (lower panels)

French Credit Data: Models with only Significant Regressors

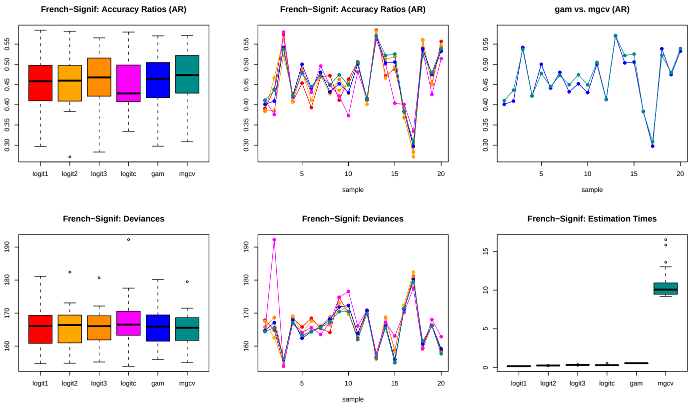


Figure: Out of sample comparison (blockwise CV with 20 blocks) for various estimators, accuracy ratios from CAP curves (upper panels), deviance values and estimation times (lower panels)

French Credit Data: Models with only Metric Regressors

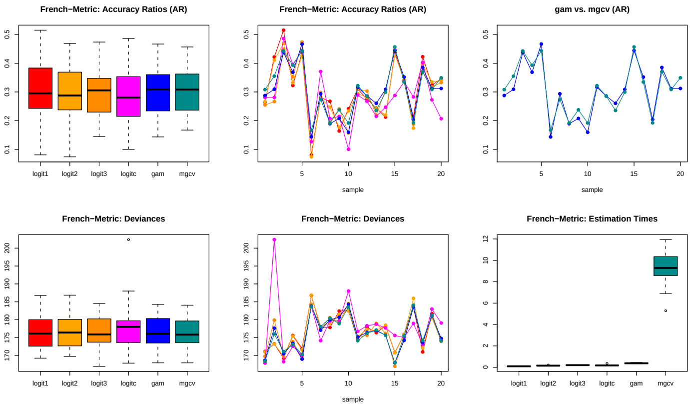
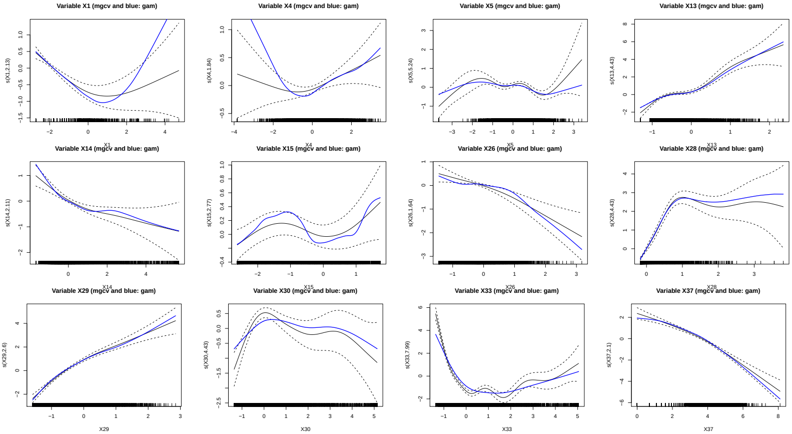


Figure: Out of sample comparison (blockwise CV with 20 blocks) for various estimators, accuracy ratios from CAP curves (upper panels), deviance values and estimation times (lower panels)

UC2005 Credit Data: Additive Functions



UC2005 Credit Data: Comparison

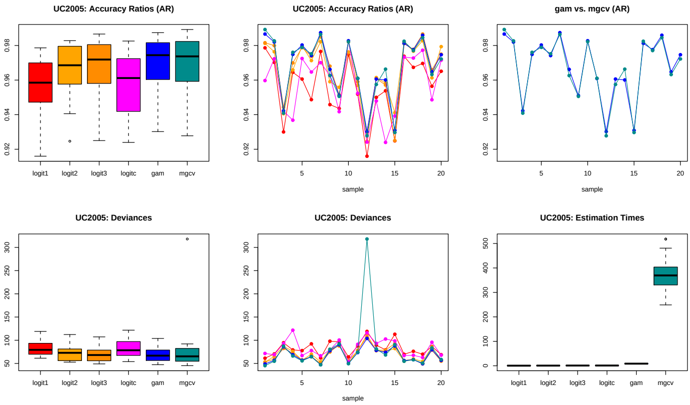


Figure: Out of sample comparison (blockwise CV with 20 blocks) for various estimators, accuracy ratios from CAP curves (upper panels), deviance values and estimation times (lower panels)

UC2005 Credit Data: Models with only Metric Regressors

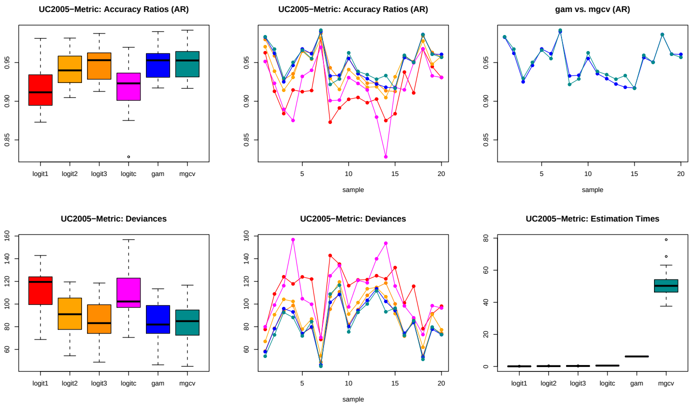


Figure: Out of sample comparison (blockwise CV with 20 blocks) for various estimators, accuracy ratios from CAP curves (upper panels), deviance values and estimation times (lower panels)