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

Müller, Marlene
Beuth University of Applied Sciences Berlin, Department II
Luxemburger Str. 10
D–13353 Berlin, Germany
E-mail: marlene.mueller@beuth-hochschule.de

We consider the estimation of credit scores by means of semiparametric logit models. In credit scoring, the fitted rating score shall not only provide an optimal classification result but serves also as a modular component of a (typically quite complex) rating system. This means in particular that a rating score should be given by a linearly weighted sum of rating factors. That way the rating procedure can be easily interpreted and understood also by non-statisticians.

For that reason the logit model or the logistic regression approach is one of the most popular models for estimating credit rating scores. The first step in fitting the rating model is usually a nonlinear transformation of the raw variables in order to obtain a linear predictor (rating score) in the final estimation. As an alternative to this two-step approach, generalized additive models (GAM) would allow for a simultaneous estimation of both the initial transformation and final logit fit. In this study we compare GAM estimating approaches with a focus on the specific structure of credit data: small default rates, mixed discrete and continuous explanatory variables, possibly nonlinear dependencies between the regressors.

Credit Rating

The statistical aspects of credit scoring have gained new importance with the implementation of current the Basle II (Basel Committee on Banking Supervision; 2004) and the upcoming Basle III capital accords on minimal capital requirements for banks. Core terms for banks in the development of an internal rating system are the estimation of a rating score and the subsequent assignment of default probabilities (PDs). Both terms are typically functions of the given explanatory variables (rating factors). In practice, often classical logit/probit-type models are used to estimate linear predictors (rating scores) and PDs simultaneously. From a statistical perspective, we consider two-group classification problem which can be analyzed using binary regression methods. However, there are additional risk management issues that should be taken into account:

- credit risk is only one part of a bank's total risk, meaning credit risk will be aggregated with other risks later on,
- the estimates obtained historical data have to allow for: stress-tests to simulate future extreme situations, easy adaptation of the rating system to possible future changes, and the possibility to extrapolate to segments without observations.

The development of a rating score and the default probability often consists of the following steps: We start from the raw data, i.e. risk factors X_j, which are measurements of several explanatory variables. A first step is the (nonlinear) transformation $X_j \to \widetilde{X}_j = m_j(X_j)$ to handle outliers and in particular to allow for nonlinear dependence on raw risk factors. The rating score is thus given by

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

Finally, the default probability is then estimated by a binary regression, implementing the model

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

where G is e.g. the logistic or Gaussian cdf (logit or probit model).

The aim of this paper is to provide a case study on (cross-sectional) rating data in order to compare different approaches to generalized additive models (GAM). We consider in particular models that allow for additional categorical variables (partial linear terms). Our interest is to simultaneously fit the transformations from the raw data, the linear rating score and the default probabilities.

Generalized Additive Models

Binary regression models, in particular logit and probit are special cases of the generalized linear model (GLM):

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

The classical generalized additive model modifies this in the way that the linear additive components are generalized to nonparametrically estimates functions:

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

This paper consider a further development, the generalized additive partial linear model (which is often also quoted as semiparametric GAM). This model allows for additional linear components:

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

This additional linear part allows us to use pre-known transformation functions for some of the risk factors as wells to add or control for additional categorical regressors.

The statistical programming environment R (R Development Core Team; 2010) comprises two standard tools to estimate generalized additive models: the function `gam::gam` implements backfitting with local scoring (Hastie and Tibshirani; 1990) and the function `mgcv::gam` implements penalized regression splines (Wood; 2006). This study compares these two procedures under their default settings.

Case Study Setup

Altogether, we consider the following competing estimators here: With logit we denote a binary GLM logit fit using the logistic cdf $G(u) = 1/\{1 + \exp(-u)\}$ as the (inverse) link function. This fit is complemented by logit2 and logit3 which denote logit fits with second and third order polynomial terms for the continuous regressors. For a further comparison we consider a logit fit where the continuous regressors categorized (4–5 factor levels) denoted by logitc. The notations gam and mgcv are used for the binary GAM fits from the R packages `gam::gam` and `mgcv::gam` with spline terms for the continuous regressors. The case study considers four credit datasets:

dataset	sample	defaults	regressors continuous	discrete	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

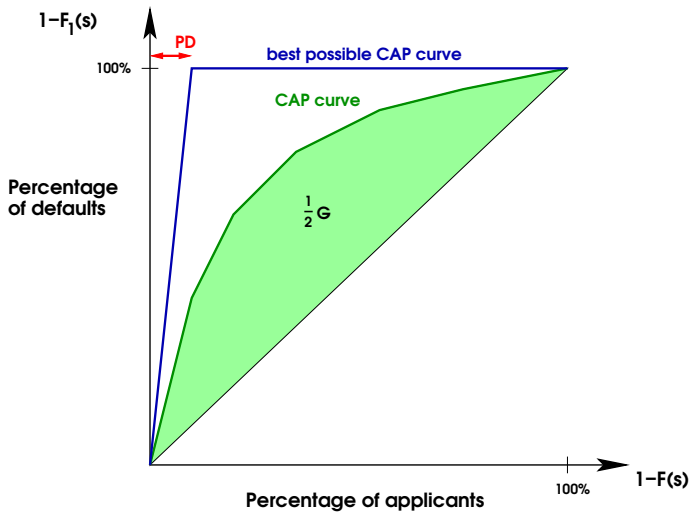
How to Compare the Binary GLM and GAM Fits?

Typically, the assessment of a credit ratings systems focuses on two aspects: discriminatory power of the rating scores and calibration (goodness of fit) of the default probabilities. Discriminatory power is commonly measured by the CAP or Lorenz curve (a variant of the ROC curve) and the accuracy ratio AR derived from this curve. Figure 1 shows the construction of the CAP curve. The difference in comparison with the ROC curve consists in plotting the cumulative distribution function of all scores against that of the default score (instead of plotting the cumulative distribution functions of the non-default and default scores against each other). The accuracy ratio calculated from the CAP curve is however linearly related to the area under curve AUC of the ROC curve:

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

The AR is the area between the CAP curve and diagonal (no separation) in relation to the corresponding area for the best possible CAP curve (perfect separation). In practice, the AR values thus vary between 0% (diagonal) and 100% (best possible).

Figure 1: Lorenz Curve (Cumulated Accuracy Profile)



The AR values that we report in the following are obtained by an out-of-sample validation. We use a block cross-validation approach, that leaves out subsamples of x% from the fitting procedure, while the remaining (100-x)% are used for estimation of the model. The AR is then calculated for the x% left-out observations. The percentage x% is differently chosen for the data cases (depending on the

default rate). Additionally to AR, we also compute deviance values

$$\mathrm{D} = -\frac{2}{n}\sum_{i=1}^{n}\left\{y_i \log(\widehat{PD}_i) + (1-y_i)\log(1-\widehat{PD}_i)\right\}$$

to assess the calibration of the fitted PDs. The deviances are obtained using the same block cross-validation approach.

Data case: German Credit Data

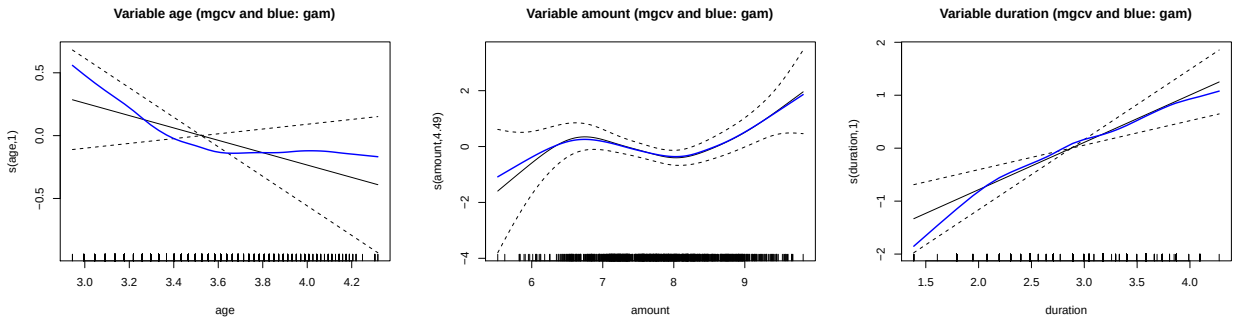
The data set is a stratified sample which oversampled the default group (30% in the sample, while the true default rate is about 5%). It contains only three continuous variables (age of the credit applicant, amount and maturity of the loan) which are complemented by numerous categorical risk factors. It is the only data set in the study where the meaning of the variables is known. For that reason, this data set is of particular interest as the results can also be interpreted from an economic point of view.

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

Data source: `http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html`

Figure 2 shows the estimated additive component functions using both `gam::gam` (in blue) and `mgcv::gam` (in black). We used the default settings for both estimators. The indicated confidence bands (dashed lines) are those of `mgcv::gam`.

Figure 2: German Credit Data: Additive component functions for continuous regressors



The following Figure 3 shows the out-of-sample comparison (blockwise validation with 10 blocks) for the various estimators, accuracy ratios from CAP curves (upper panels), deviance values and estimation times (lower panels) The most important findings are: Some observation(s) that seem to confuse `mgcv::gam` in one CV subsample, which causes the peak in the middle lower panel Figure 3. However, `mgcv::gam` seems to improve deviance and discriminatory power w.r.t. `gam::gam` in all other cases. If we only use the continuous regressors, both GAM estimators are comparable to logit estimates with cubic additive functions (Figure 4).

Figure 3: German Credit Data: Comparison of models

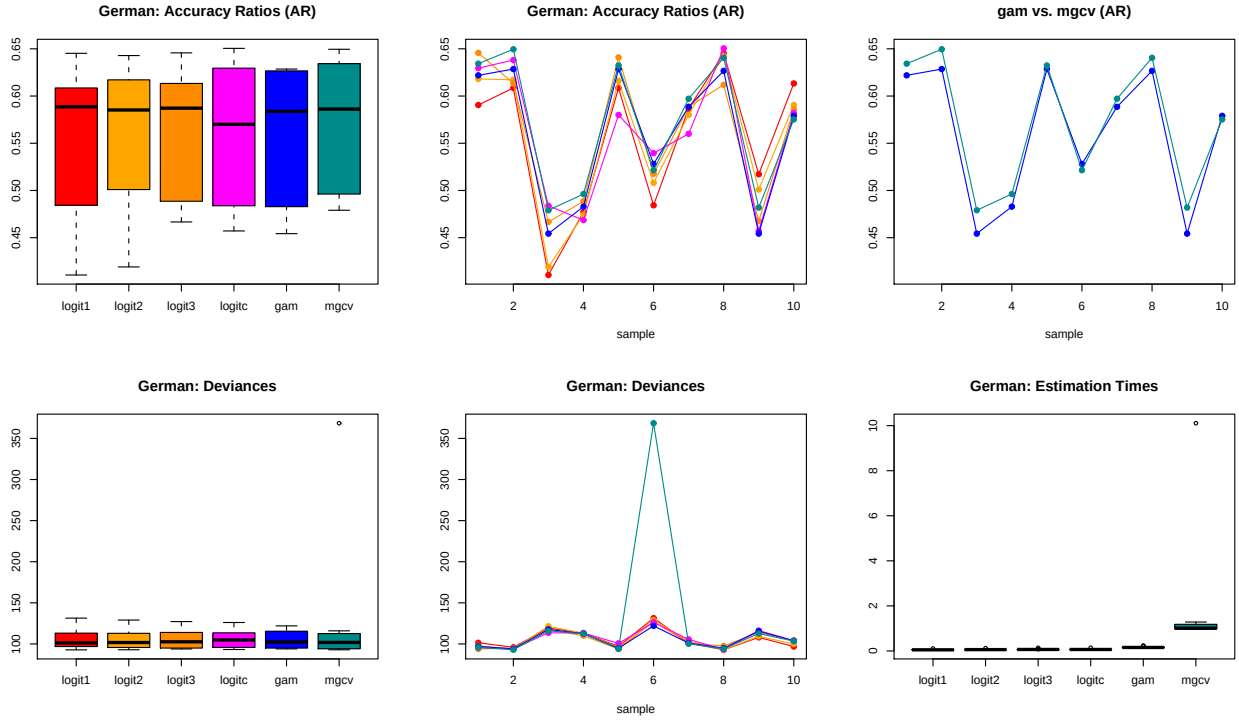
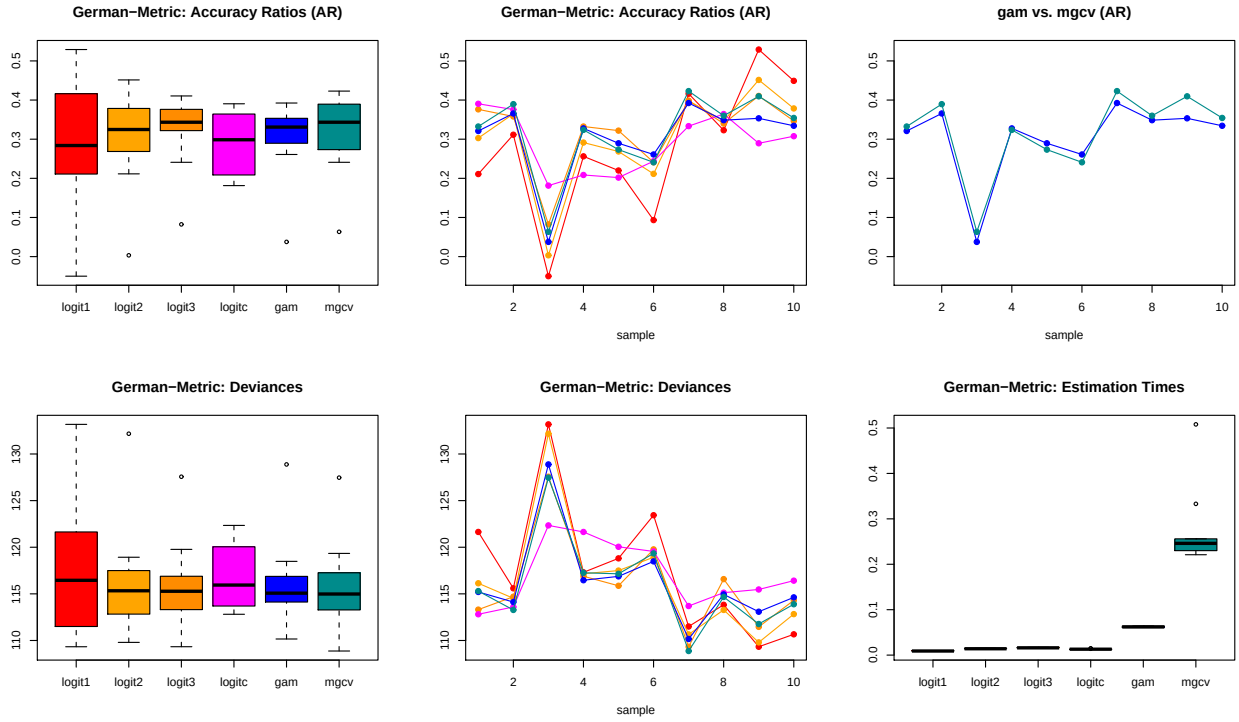


Figure 4: German Credit Data: Models with only Continuous Regressors



Conclusions

This study focuses on the semiparametric GAM estimation of ratings scores and default probabilities. As we experience that typically, categorical regressors improve fit significantly, estimation methods for credit data should adequately use these as well. The classical backfitting with local scoring approach for GAM (in R: `gam::gam`) provides fast and numerically stable results. There is however clear indication, that penalized regression splines (in R: `mgcv::gam`) may provide more precise estimates of the additive component functions. Issues for further study are the estimation time (that is increasing with model complexity and the inclusion of categorical variables) and that the effect of a higher precision seems to be seen only in large samples.

REFERENCES (RÉFERENCES)

Basel Committee on Banking Supervision (2004). *Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework*, Bank for International Settlements (BIS), Basel, Switzerland.
URL: *http://www.bis.org*

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.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

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