KernGPLM – A Package for Kernel-Based Fitting of Generalized Partial Linear and Additive Models

June 8, 2006

Marlene Müller

Fraunhofer Institut
Techno- und
Wirtschaftsmathematik

Aim of this Talk

analysis of highdimensional data by semiparametric (generalized) regression models

- compare different approaches to additive models (AM) and generalized additive models (GAM)
- include categorical variables $\Longrightarrow$ *partial linear* terms (combination of AM/PLM and GAM/GPLM)
- provide software $\Rightarrow$ R package **KernGPLM**
- focus on kernel-based techniques for **high-dimensional data**

Financial application: Credit Rating

- new interest in this field because of **Basel II**:
  capital requirements of a bank are adapted to the individual credit portfolio
- key problems: determine **rating score** and subsequently **default probabilities (PDs)** as a function of some explanatory variables

$\rightarrow$ classical **logit/probit-type models** to estimate linear predictors (scores) and probabilities (PDs)

Two objectives:

- study **single factors**
- find the **best model**

Binary choice model

$\rightarrow$ credit rating: estimate scores + PDs

$$P(Y=1|\boldsymbol{X}) = E(Y|\boldsymbol{X}) = G(\boldsymbol{\beta}^\top \boldsymbol{X})$$

$\rightarrow$ parametric binary choice models

logit	$P(Y=1\|\boldsymbol{X}) = F(\boldsymbol{X}^\top\boldsymbol{\beta})$	$F(\bullet) = \frac{1}{1+e^{-\bullet}}$
probit	$P(Y=1\|\boldsymbol{X}) = \Phi(\boldsymbol{X}^\top\boldsymbol{\beta})$	$\Phi(\bullet)$ standard normal cdf

Generalized linear model (GLM)

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

Data Example: Credit Data

References: Fahrmeir/Hamerle (1984); Fahrmeir & Tutz (1995)

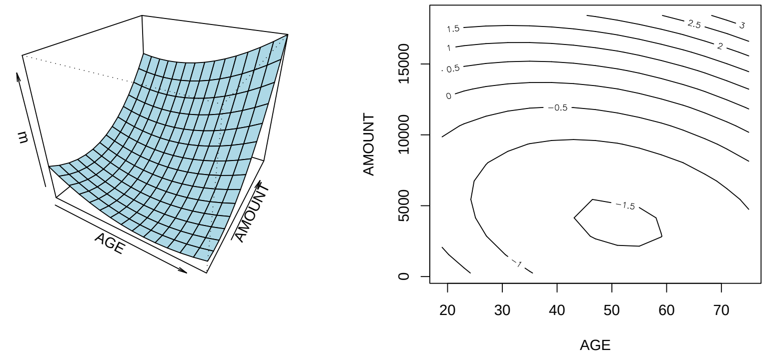
- default indicator: $Y \in \{0,1\}$, where $1 =$ default
- explanatory variables:
  personal characteristics, credit history, credit characteristics
- sample size: 1000 (stratified sample with 300 defaults)

Estimated (Logit) Scores

$$\begin{aligned}\text{Score} \quad = \quad & 1.334 - 0.763^{***}\cdot \text{previous} - 0.310\cdot \text{employed} + 0.566^{**}\cdot (\text{d9-12}) \\ & +0.898^{**}\cdot (\text{d12-18}) + 0.981^{***}\cdot (\text{d18-24}) + 1.550^{***}\cdot (\text{d>24}) \\ & -0.984^{***}\cdot \text{savings} - 0.363^{**}\cdot \text{purpose} + 0.660^{***}\cdot \text{house} \\ & -0.000251^{**}\cdot \text{amount} - 0.0942^{**}\cdot \text{age} + 0.0000000173^{**}\cdot \text{amount}^2 \\ & +0.000833^{*}\cdot \text{age}^2 + 0.00000236\cdot (\text{amount} \cdot \text{age})\end{aligned}$$

*, **, *** denote significant coefficients at the 10%, 5%, 1% level, respectively

Data Example: Logit (with interaction)



credit default on AGE and AMOUNT using quadratic and interaction terms, left: surface and right: contours of the fitted score function

Semiparametric Models

- local regression

$$E(Y|\boldsymbol{T}) = G\{m(\boldsymbol{T})\}, \quad m \text{ nonparametric}$$

- generalized partial linear model (GPLM)

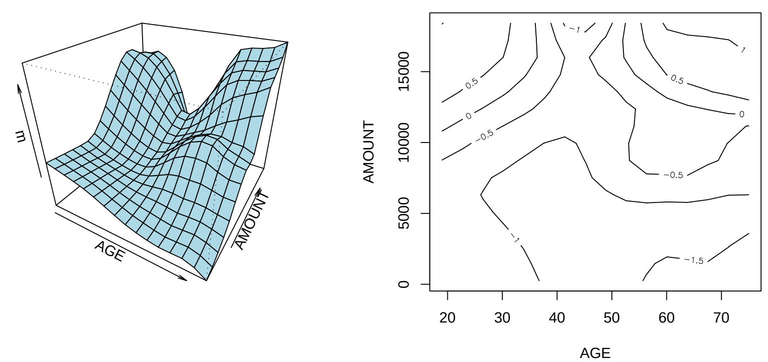
$$E(Y|\boldsymbol{X},\boldsymbol{T}) = G\left\{\boldsymbol{X}^\top\boldsymbol{\beta} + m(\boldsymbol{T})\right\} \quad m \text{ nonparametric}$$

- generalized additive partial linear model (semiparametric GAM)

$$E(Y|\boldsymbol{X},\boldsymbol{T}) = G\left\{\beta_0 + \boldsymbol{X}^\top\boldsymbol{\beta} + \sum_{j=1}^{p} m_j(T_j)\right\} \quad m_j \text{ nonparametric}$$

Some references:
Loader (1999), Hastie and Tibshirani (1990), Härdle et al. (2004), Green and Silverman (1994)

Data Example: GPLM



credit default on AGE and AMOUNT using a nonparametric function, left: surface and right: contours of the fitted score function on AGE and AMOUNT

Estimation Approaches for GPLM/GAM

- GPLM:
  - ⋆ generalization of Speckman's estimator (type of profile likelihood)
  - ⋆ backfitting for two additive components and local scoring

References:
(PLM) Speckman (1988), Robinson (1988); (PLM/splines) Schimek (2000), Eubank et al. (1998), Schimek (2002); (GPLM) Severini and Staniswalis (1994), Müller (2001)

- semiparametric GAM:
  - ⋆ [modified | smooth] backfitting and local scoring
  - ⋆ marginal [internalized] integration

References:
(marginal integraton) Tjøstheim and Auestad (1994), Chen et al. (1996), Hengartner et al. (1999), Hengartner and Sperlich (2005);
(backfitting) Buja et al. (1989), Mammen et al. (1999), Nielsen and Sperlich (2005)

Estimation of the GPLM: generalized Speckman estimator

- partial linear model (identity G)

$$E(Y|\boldsymbol{X},\boldsymbol{T}) = \boldsymbol{X}^T\boldsymbol{\beta} + m(\boldsymbol{T})$$

$$\Longrightarrow \quad \begin{aligned} \boldsymbol{m}^{new} &= \mathbf{S}(\boldsymbol{Y} - \mathcal{X}\boldsymbol{\beta}) \\ \boldsymbol{\beta}^{new} &= (\widetilde{\mathcal{X}}^T\widetilde{\mathcal{X}})^{-1}\widetilde{\mathcal{X}}^T\widetilde{\boldsymbol{Y}} \end{aligned}$$

- generalized partial linear model

$$E(Y|\boldsymbol{X},\boldsymbol{T}) = G\{\boldsymbol{X}^T\boldsymbol{\beta} + m(\boldsymbol{T})\}$$

$\Longrightarrow$ above for adjusted dependent variable

$$\boldsymbol{Z} = \mathcal{X}\boldsymbol{\beta} + \boldsymbol{m} - \mathcal{W}^{-1}\boldsymbol{v},$$

$\boldsymbol{v} = (\ell_i')$, $\mathcal{W} = \mathrm{diag}(\ell_i'')$

References: Severini and Staniswalis (1994)

Comparison of Algorithms

	parametric step	nonparametric step	est. matrix
Speckman	$\boldsymbol{\beta}^{new} = (\widetilde{\mathcal{X}}^T\mathcal{W}\widetilde{\mathcal{X}})^{-1}\widetilde{\mathcal{X}}^T\mathcal{W}\widetilde{\boldsymbol{Z}}$	$\boldsymbol{m}^{new} = \mathbf{S}(\boldsymbol{Z} - \mathcal{X}\boldsymbol{\beta})$	$\boldsymbol{\eta} = \mathcal{R}^S\boldsymbol{Z}$
Backfitting	$\boldsymbol{\beta}^{new} = (\mathcal{X}^T\mathcal{W}\widetilde{\mathcal{X}})^{-1}\mathcal{X}^T\mathcal{W}\widetilde{\boldsymbol{Z}}$	$\boldsymbol{m}^{new} = \mathbf{S}(\boldsymbol{Z} - \mathcal{X}\boldsymbol{\beta})$	$\boldsymbol{\eta} = \mathcal{R}^B\boldsymbol{Z}$
Profile	$\boldsymbol{\beta}^{new} = (\mathcal{X}^T\mathcal{W}\widetilde{\mathcal{X}})^{-1}\mathcal{X}^T\mathcal{W}\widetilde{\boldsymbol{Z}}$	$\boldsymbol{m}^{new} = \ldots$	$\boldsymbol{\eta} = \mathcal{R}^P\boldsymbol{Z}$

Speckman/Backfitting:
$\widetilde{\mathcal{X}} = (\mathbf{I} - \mathbf{S})\mathcal{X}$, $\widetilde{\boldsymbol{Z}} = (\mathbf{I} - \mathbf{S})\boldsymbol{Z}$, $\mathbf{S}$ weighted smoother matrix

Profile Likelihood:
$\widetilde{\mathcal{X}} = (\mathbf{I} - \mathbf{S}^P)\mathcal{X}$, $\widetilde{\boldsymbol{Z}} = (\mathbf{I} - \mathbf{S}^P)\boldsymbol{Z}$, $\mathbf{S}^P$ weighted (different) smoother matrix

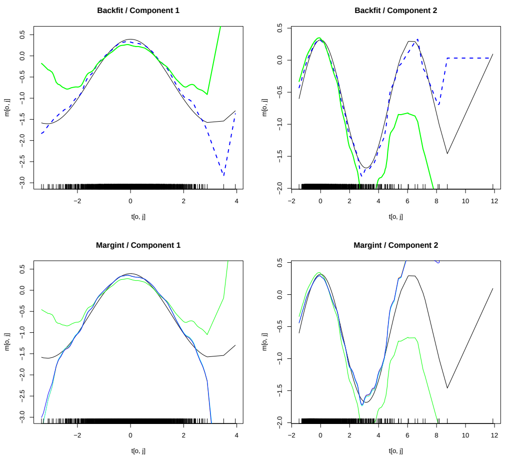
References: Severini and Staniswalis (1994), Müller (2001)

Estimation of the GAM

$$E(Y|\boldsymbol{X},\boldsymbol{T}) = G\left\{\beta_0 + \boldsymbol{X}^\top\boldsymbol{\beta} + \sum_{j=1}^{p} m_j(T_j)\right\} \qquad m_j \text{ nonparametric}$$

- **classical backfitting:** fit single components by regression on the residuals w.r.t. the other components
- **modified backfitting:** first project on the linear space spanned by all regressors and then nonparametrically fit the partial residuals
- **marginal (internalized) integration:** estimate the marginal effect by integrating a full dimensional nonparametric regression estimate
  $\Longrightarrow$ original proposal is computationally intractable: $O(n^3)$
  $\Longrightarrow$ choice of nonparametric estimate is essential: **marginal internalized integration**

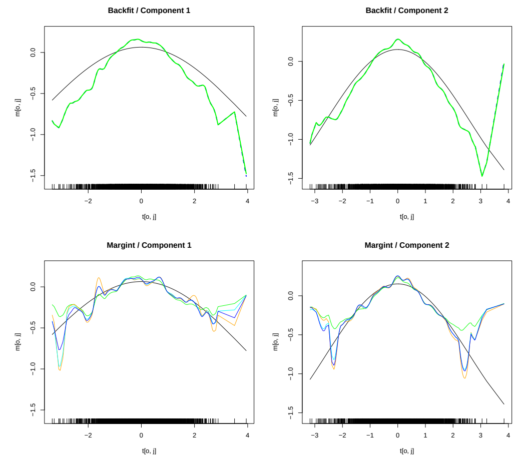
Simulation Example: True Additive Function



- B – classical
- B – modified
- M – classical
- M – pdf estimate 1
- M – pdf estimate 2
- M – normal pdfs

Marginal integration – as initialization for backfitting

Simulation Example: True Non-Additive Function



- B – classical
- B – modified
- M – classical
- M – pdf estimate 1
- M – pdf estimate 2
- M – normal pdfs

Marginal integration – estimate of marginal effects

Comparison of Algorithms

- consistency of marginal integration:
  - ⇒ if underlying function is truly additive, backfitting outperforms marginal integration
  - ⇒ consider marginal integration to initialize backfitting (replacing the usual zero-functions)
- comparison of backfitting and marginal integration:
  - ⇒ marginal integration indeed estimates marginal effects, but large number of observations is needed
  - ⇒ estimation method of the instruments is essential, dimension reduction techniques are required

Summary

- GPLM and semiparametric GAM are natural extensions of the GLM
- *large amount of data* is needed for estimating marginal effects

⇒ R package **KernGPLM** with routines for
  - ⋆ (kernel based) generalized partial linear and additive models
  - ⋆ additive components by [modified] backfitting + local scoring
  - ⋆ additive components by marginal [internalized] integration

- possible extensions:
  - ⋆ smooth backfitting
  - ⋆ externalized marginal integration

References

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Annals of Statistics*, 17:453–555.

Chen, R., Härdle, W., Linton, O., and Severance-Lossin, E. (1996). Estimation and variable selection in additive nonparametric regression models. In Härdle, W. and Schimek, M., editors, *Proceedings of the COMPSTAT Satellite Meeting Semmering 1994*, Heidelberg. Physica Verlag.

Eubank, R. L., Kambour, E. L., Kim, J. T., Klipple, K., Reese, C. S., and Schimek, M. G. (1998). Estimation in partially linear models. *Computational Statistics & Data Analysis*, 29:27–34.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, volume 58 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

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*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

Hengartner, N., Kim, W., and Linton, O. (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, 8:1–20.

Hengartner, N. and Sperlich, S. (2005). Rate-optimal estimation with the integration method in the presence of many covariates. *Journal of Multivariate Analysis*, 95:246–272.

Loader, C. (1999). *Local Regression and Likelihood*. Springer, New York.

Mammen, E., Linton, O., and Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, 27:1443–1490.

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

Nielsen, J. and Sperlich, S. (2005). Smooth backfitting in practice. *Journal of the Royal Statistical Society, Series B*, 67:43–61.

Robinson, P. M. (1988). Root n–consistent semiparametric regression. *Econometrica*, 56:931–954.

Schimek, M. G. (2000). Estimation and inference in partially linear models with smoothing splines. *Journal of Statistical Planning and Inference*, 91:525–540.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89:501–511.

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

Tjøstheim, D. and Auestad, B. (1994). Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association*, 89:1398–1409.