universität
innsbruck

LASSO-Type Penalization in the Framework of Generalized Additive Models for Location, Scale and Shape

Nikolaus Umlauf

https://eeecon.uibk.ac.at/~umlauf/

Overview

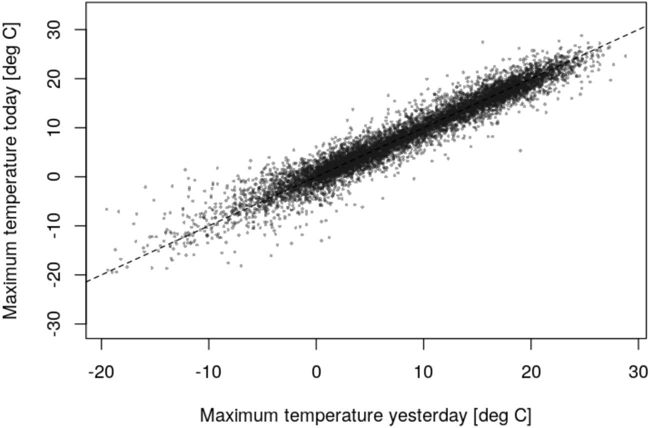
Joint work with Andreas Groll, Julien Hambuckers and Thomas Kneib.

1. Introduction
2. Model specification
3. Model fitting
4. L1-type penalization
5. A simulation study
6. An application on the Munich rent data

Introduction

Helsinki daily maximum temperature data (1993/06-2017/06)

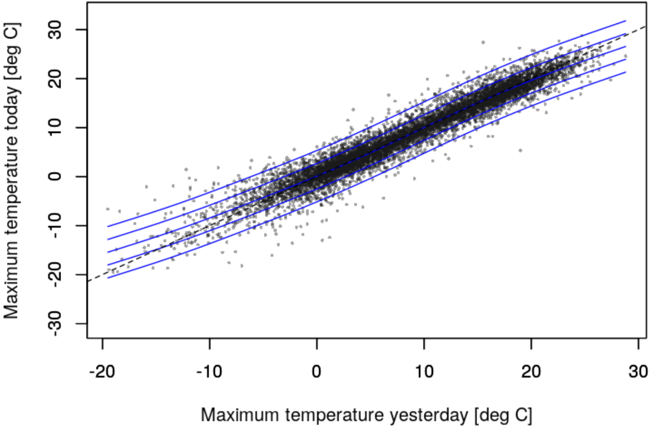
$$\mathtt{T} \sim N(\mu, \sigma^2).$$



Introduction

Helsinki daily maximum temperature data (1993/06-2017/06)

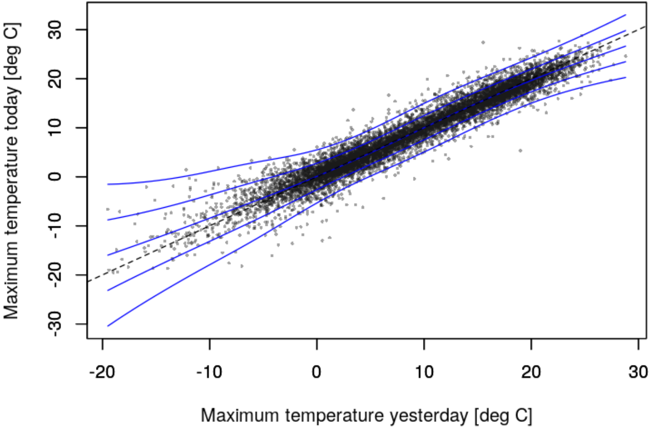
$$\mathtt{T} \sim N(\mu = f(\mathtt{T}_{t-1}), log(\sigma^2) = \beta_0).$$



Introduction

Helsinki daily maximum temperature data (1993/06-2017/06)

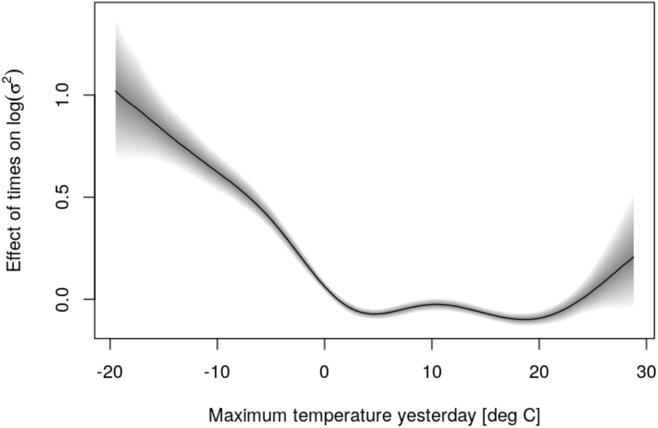
$$\mathtt{T} \sim N(\mu = f(\mathtt{T}_{t-1}),\ log(\sigma^2) = f(\mathtt{T}_{t-1})).$$



Introduction

Helsinki daily maximum temperature data (1993/06-2017/06)

$$\mathtt{T} \sim N(\mu = f(\mathtt{T}_{t-1}),\ log(\sigma^2) = f(\mathtt{T}_{t-1})).$$



Model specification

Any parameter of a population distribution $\mathcal{D}$ may be modeled by explanatory variables

$$y \sim \mathcal{D}\left(h_1(\theta_1) = \eta_1,\ h_2(\theta_2) = \eta_2, \dots,\ h_K(\theta_K) = \eta_K\right),$$

Each parameter is linked to a structured additive predictor

$$h_k(\theta_k) = \eta_k = \eta_k(\mathbf{x}; \boldsymbol{\beta}_k) = f_{1k}(\mathbf{x}; \boldsymbol{\beta}_{1k}) + \ldots + f_{J_k k}(\mathbf{x}; \boldsymbol{\beta}_{J_k k}),$$

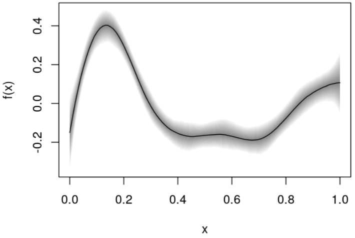
$j = 1, \dots, J_k$ and $k = 1, \dots, K$ and $h_k(\cdot)$ are known monotonic link functions.

Vector of function evaluations $\mathbf{f}_{jk} = (f_{jk}(\mathbf{x}_1; \boldsymbol{\beta}_{jk}), \dots, f_{jk}(\mathbf{x}_n; \boldsymbol{\beta}_{jk}))^\top$

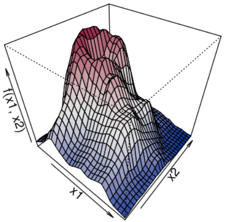
$$\mathbf{f}_{jk} = \begin{pmatrix} f_{jk}(\mathbf{x}_1; \boldsymbol{\beta}_{jk}) \\ \vdots \\ f_{jk}(\mathbf{x}_n; \boldsymbol{\beta}_{jk}) \end{pmatrix} = f_{jk}(\mathbf{X}_{jk}; \boldsymbol{\beta}_{jk}).$$

Model specification

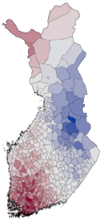
Nonlinear effects of continuous covariates



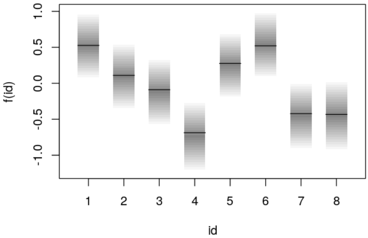
Two-dimensional surfaces



Spatially correlated effects f(x) = f(s)



Random intercepts f(x) = f(id)



Model specification

For simple linear effects $\mathbf{X}_{jk}\boldsymbol{\beta}_{jk}$: $p_{jk}(\boldsymbol{\beta}_{jk}) \propto const.$

For the smooth terms:

$$p_{jk}(\boldsymbol{\beta}_{jk}; \boldsymbol{\tau}_{jk}, \boldsymbol{\alpha}_{jk}) \propto d_{\boldsymbol{\beta}_{jk}}(\boldsymbol{\beta}_{jk} | \, \boldsymbol{\tau}_{jk}; \boldsymbol{\alpha}_{\boldsymbol{\beta}_{jk}}) \cdot d_{\boldsymbol{\tau}_{jk}}(\boldsymbol{\tau}_{jk} | \, \boldsymbol{\alpha}_{\boldsymbol{\tau}_{jk}}).$$

Using a basis function approach a common choice is

$$d_{\boldsymbol{\beta}_{jk}}(\boldsymbol{\beta}_{jk} | \, \boldsymbol{\tau}_{jk}, \boldsymbol{\alpha}_{\boldsymbol{\beta}_{jk}}) \propto |\mathbf{P}_{jk}(\boldsymbol{\tau}_{jk})|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\beta}_{jk}^{\top}\mathbf{P}_{jk}(\boldsymbol{\tau}_{jk})\boldsymbol{\beta}_{jk}\right).$$

Precision matrix $\mathbf{P}_{jk}(\boldsymbol{\tau}_{jk})$ derived from prespecified penalty matrices $\boldsymbol{\alpha}_{\boldsymbol{\beta}_{jk}} = \{\mathbf{K}_{1jk}, \ldots, \mathbf{K}_{Ljk}\}$.

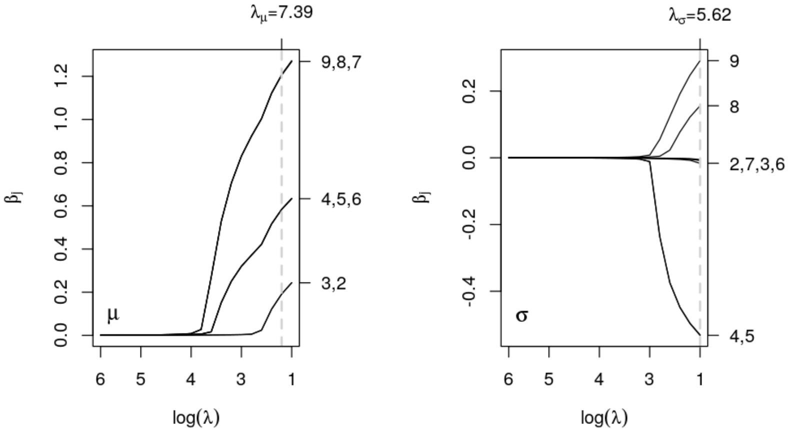
The variances parameters $\boldsymbol{\tau}_{jk}$ are equivalent to the inverse smoothing parameters in a frequentist approach.

Regularization in the GAMLSS framework

- A gradient boosting approach is provided by Mayr et al. (2012).
- Allows for variable selection within GAMLSS framework.
- Corresponding R-package *gamboostLSS* (Hofner et al., 2015).
- Provides a large number of pre-specified distributions.
- **New:** an alternative *gradient boosting* approach is implemented in the R-package *bamlss* (Umlauf et al., 2017b):
  - embeds many different approaches suggested in literature and software,
  - serves as unified conceptional "Lego toolbox" for complex regression models.

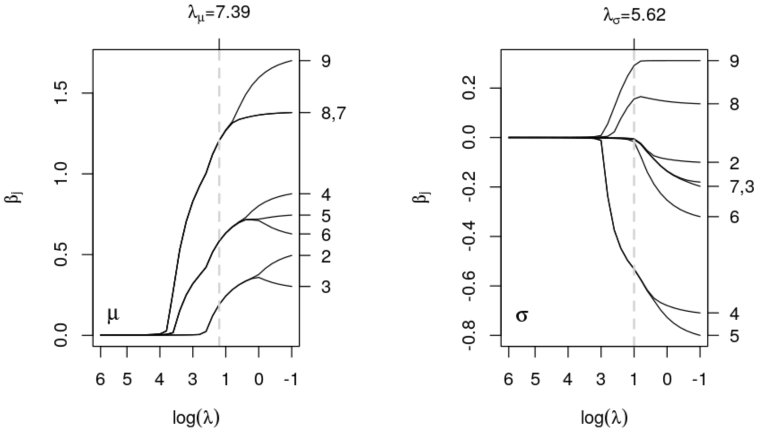
Regularization in the GAMLSS framework

New model terms $f_{jk}(\mathbf{x}; \beta_{jk})$ with LASSO-type penalties.



Regularization in the GAMLSS framework

New model terms $f_{jk}(\mathbf{x}; \beta_{jk})$ with LASSO-type penalties.



Model fitting

The main building block of regression model algorithms is the probability density function $d_y(\mathbf{y}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$.

Estimation typically requires to evaluate

$$\begin{aligned}\ell(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) &= \sum_{i=1}^{n} \log \, d_y(y_i; \theta_{i1} = h_1^{-1}(\eta_{i1}(\mathbf{x}_i, \boldsymbol{\beta}_1)), \ldots \\ &\qquad \ldots, \theta_{iK} = h_K^{-1}(\eta_{iK}(\mathbf{x}_i, \boldsymbol{\beta}_K))),\end{aligned}$$

with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top)^\top$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_K)$.

The log-posterior

$$\log \, \pi(\boldsymbol{\beta}, \boldsymbol{\tau}; \mathbf{y}, \mathbf{X}, \boldsymbol{\alpha}) \propto \ell(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) + \sum_{k=1}^{K} \sum_{j=1}^{J_k} \left[\log \, p_{jk}(\boldsymbol{\beta}_{jk}; \boldsymbol{\tau}_{jk}, \boldsymbol{\alpha}_{jk}) \right],$$

where $\boldsymbol{\tau} = (\boldsymbol{\tau}_1^\top, \ldots, \boldsymbol{\tau}_K^\top)^\top = (\boldsymbol{\tau}_{11}^\top, \ldots, \boldsymbol{\tau}_{J_1 1}^\top, \ldots, \boldsymbol{\tau}_{1K}^\top, \ldots, \boldsymbol{\tau}_{J_K K}^\top)^\top$ (frequentist, penalized log-likelihood).

Model fitting

Posterior mode estimation, fortunately, partitioned updating is possible

$$\begin{aligned}\beta_1^{(t+1)} &= U_1(\beta_1^{(t)}, \beta_2^{(t)}, \ldots, \beta_K^{(t)}) \\ \beta_2^{(t+1)} &= U_2(\beta_1^{(t+1)}, \beta_2^{(t)}, \ldots, \beta_K^{(t)}) \\ &\vdots \\ \beta_K^{(t+1)} &= U_K(\beta_1^{(t+1)}, \beta_2^{(t+1)}, \ldots, \beta_K^{(t)}),\end{aligned}$$

E.g., Newton-Raphson type updating

$$\beta_k^{(t+1)} = U_k(\beta_k^{(t)}, \cdot) = \beta_k^{(t)} - \mathbf{H}_{kk}\left(\beta_k^{(t)}\right)^{-1}\mathbf{s}\left(\beta_k^{(t)}\right).$$

Can be further partitioned for each function within parameter block k. Moreover, using a basis function approach yields IWLS updates

$$\beta_{jk}^{(t+1)} = (\mathbf{X}_{jk}^\top \mathbf{W}_{kk}\mathbf{X}_{jk} + \mathbf{G}_{jk}(\boldsymbol{\tau}_{jk}))^{-1}\mathbf{X}_{jk}^\top \mathbf{W}_{kk}(\mathbf{z}_k - \boldsymbol{\eta}_{k,-j}^{(t)}).$$

Model fitting

A simple generic algorithm for distributional regression models:

while(eps $> \varepsilon$ & t < maxit) {
  for(k in 1:K) {
    for(j in 1:J[k]) {
      Compute $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta}_{k} - \mathbf{f}_{jk}$.
      Obtain new $(\boldsymbol{\beta}_{jk}^{\star}, \boldsymbol{\tau}_{jk}^{\star})^\top = U_{jk}(\mathbf{X}_{jk}, \mathbf{y}, \tilde{\boldsymbol{\eta}}, \boldsymbol{\beta}_{jk}^{[t]}, \boldsymbol{\tau}_{jk}^{[t]})$.
      Update $\boldsymbol{\eta}_{k}$.
    }
  }
  t = t + 1
  Compute new eps.
}

Functions $U_{jk}(\cdot)$ could either return updates from an optimizing algorithm or proposals from a MCMC sampler.

L1-type penalization

Idea: depending on the type of covariate effects, subtract a combination of (parts of) the following penalty terms $\tau^{-1} J(\boldsymbol{\beta})$ from the log-likelihood.

Classical LASSO (Tibshirani, 1996): For a metric covariate x_{jk} use

$$J_m(\beta_{jk}) = |\beta_{jk}| \,.$$

Group LASSO (Meier et al., 2008): For a (dummy-encoded) categorical covariate $\mathbf{x}_{jk}$ use

$$J_g(\boldsymbol{\beta}_{jk}) = ||\boldsymbol{\beta}_{jk}||_2 \,,$$

with vector $\boldsymbol{\beta}_{jk}$ collecting all corresponding coefficients.

L1-type penalization

Alternatively, for categorical covariates often *clustering* of categories with implicit *factor selection* is desirable.

Fused LASSO (Gertheiss and Tutz, 2010): Depending on the *nominal* (left) or *ordinal* scale level (right) of the covariate, use

$$J_f(\boldsymbol{\beta}_{jk}) = \sum_{l>m} w_{lm}^{(jk)} \, |\beta_{jkl} - \beta_{jkm}| \text{ or } J_f(\boldsymbol{\beta}_{jk}) = \sum_{l=1}^{c_{jk}} w_{l}^{(jk)} \, |\beta_{jkl} - \beta_{jk,l-1}|$$

where c_{jk} is the number of levels of categorical predictor $\mathbf{x}_{jk}$ and $w_{lm}^{(jk)}, w_{l}^{(jk)}$ denote suitable weights. Choosing $l = 0$ as the reference, $\beta_{jk0} = 0$ is fixed.

L1-type penalization

Quadratic approximations of the penalties (compare Oelker & Tutz, 2017)

$$J_{jk}(\boldsymbol{\beta}_{jk}) \approx J_{jk}(\boldsymbol{\beta}_{jk}^{(t)}) + \frac{1}{2}\left(\boldsymbol{\beta}_{jk}^{\top}\mathbf{P}_{jk}(\boldsymbol{\beta}_{jk})\boldsymbol{\beta}_{jk} + (\boldsymbol{\beta}_{jk}^{(t)})^{\top}\mathbf{P}_{jk}(\boldsymbol{\beta}_{jk}^{(t)})\boldsymbol{\beta}_{jk}^{(t)}\right),$$

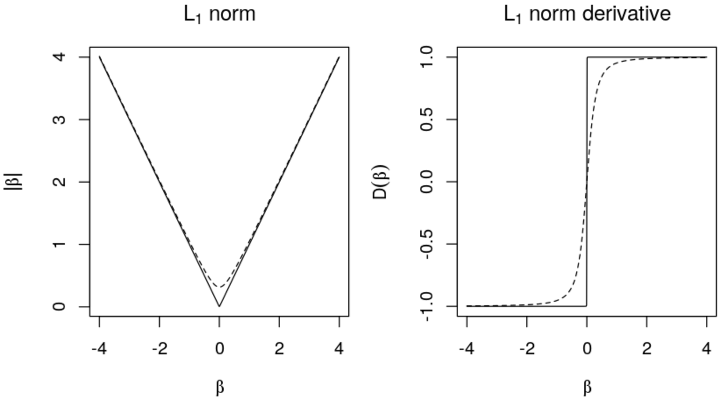
with

$$\mathbf{P}_{jk}(\boldsymbol{\beta}_{jk}^{(t)}) = q'_{jk}\left(\left\|\mathbf{a}_{jk}^{\top}\boldsymbol{\beta}_{jk}^{(t)}\right\|_{N_{jk}}\right) \cdot \frac{D_{jk}(\mathbf{a}_{jk}^{\top}\boldsymbol{\beta}_{jk}^{(t)})}{\mathbf{a}_{jk}^{\top}\boldsymbol{\beta}_{jk}^{(t)}} \cdot \mathbf{a}_{jk}\mathbf{a}_{jk}^{\top}.$$

E.g., $\|\beta\|_1 = |\beta|$ is approximated by $\sqrt{\beta^2 + c}$, hence, IWLS based updating functions $\mathtt{U_{jk}}(\cdot)$ are relatively easy to implement.

L1-type penalization

Example of the approximation of the L_1 norm.



Usually setting the constant to $c \approx 10^{-5}$ works well.

R package *bamlss*

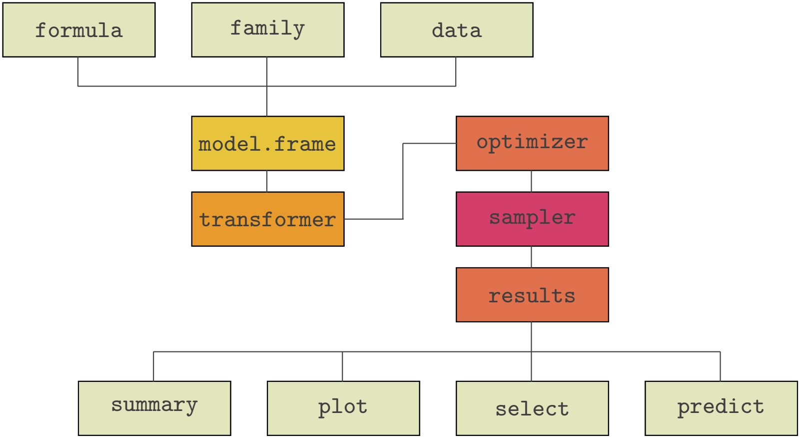
The package is available at

https://CRAN.R-project.org/package=bamlss

Development version, in R simply type

```
R> install.packages("bamlss",
+   repos = "http://R-Forge.R-project.org")
```

R package *bamlss*



In principle, the setup does not restrict to any specific type of engine (Bayesian or frequentist).

R package *bamlss*

Type	Function
Parser	`bamlss.frame()`
Transformer	`bamlss.engine.setup()`, `randomize()`
Optimizer	`bfit()`, `opt()`, `cox.mode()`, `jm.mode()`
	`boost()`, `lasso()`
Sampler	`GMCMC()`, `JAGS()`, `STAN()`, `BayesX()`,
	`cox.mcmc()`, `jm.mcmc()`
Results	`results.bamlss.default()`

To implement new engines, only the building block functions have to be exchanged.

R package *bamlss*

Work in progress ...

Function	Distribution
`beta_bamlss()`	Beta distribution
`binomial_bamlss()`	Binomial distribution
`cnorm_bamlss()`	Censored normal distribution
`cox_bamlss()`	Continuous time Cox-model
`gaussian_bamlss()`	Gaussian distribution
`gamma_bamlss()`	Gamma distribution
`gpareto_bamlss()`	Generalized Pareto distribution
`jm_bamlss()`	Continuous time joint-model
`multinomial_bamlss()`	Multinomial distribution
`mvn_bamlss()`	Multivariate normal distribution
`poisson_bamlss()`	Poisson distribution
...	

New families only require density, distribution, random number generator, quantile, score and hess functions.

R package *bamlss*

Wrapper function:

```
R> f <- list(y ~ la(id,fuse=2), sigma ~ la(id,fuse=1))
R> b <- bamlss(f, family = "gaussian", sampler = FALSE,
+    optimizer = lasso, criterion = "BIC", multiple = TRUE)
```

Standard extractor and plotting functions:

`summary()`, `plot()`, `fitted()`, `residuals()`, `predict()`, `coef()`, `logLik()`, `DIC()`, `samples()`, ...

Simulation setting

- $Y \sim N(\mu(\mathbf{x}); \sigma(\mathbf{x})^2)$
- $\mu(\mathbf{x}) = \beta_{0\mu} + \mathbf{x}_1\boldsymbol{\beta}_{1\mu} + \ldots + \mathbf{x}_4\boldsymbol{\beta}_{4\mu}\,, \quad \sigma(\mathbf{x}) = exp\,(\,\beta_{0\sigma} + \mathbf{x}_1\boldsymbol{\beta}_{1\sigma} + \ldots + \mathbf{x}_4\boldsymbol{\beta}_{4\sigma})$
- $\mathbf{x}_1, \mathbf{x}_2$ nominal factors, $\mathbf{x}_3, \mathbf{x}_4$ ordinal factors.
- Intercepts: $\beta_{0\mu} = 1, \beta_{0\sigma} = -0.5$.

$$
\begin{aligned}
\boldsymbol{\beta}_{1\mu} &= (0, 0.5, 0.5, 0.5, 0.5, -0.2, -0.2), & \boldsymbol{\beta}_{2\mu} &= (0, 1, 1) \\
\boldsymbol{\beta}_{3\mu} &= (0, 0.5, 0.5, 1, 1, 2, 2), & \boldsymbol{\beta}_{4\mu} &= (0, -0.3, -0.3)
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\beta}_{1\sigma} &= (0, -0.5, 0.4, 0, -0.5, 0.4, 0), & \boldsymbol{\beta}_{2\sigma} &= (0.4, 0, 0.4) \\
\boldsymbol{\beta}_{3\sigma} &= (0, 0, 0.4, 0.4, 0.4, 0.8, 0.8), & \boldsymbol{\beta}_{4\sigma} &= (0, -0.5, -0.5)
\end{aligned}
$$

- Additional noise variables: $\mathbf{x}_5, \mathbf{x}_6$ nominal factors, $\mathbf{x}_7, \mathbf{x}_8$ ordinal factors.
- $n_{train} = 1000$ with 100 simulation runs.

Model comparison

The following different models are compared w.r.t. goodness-of-fit:

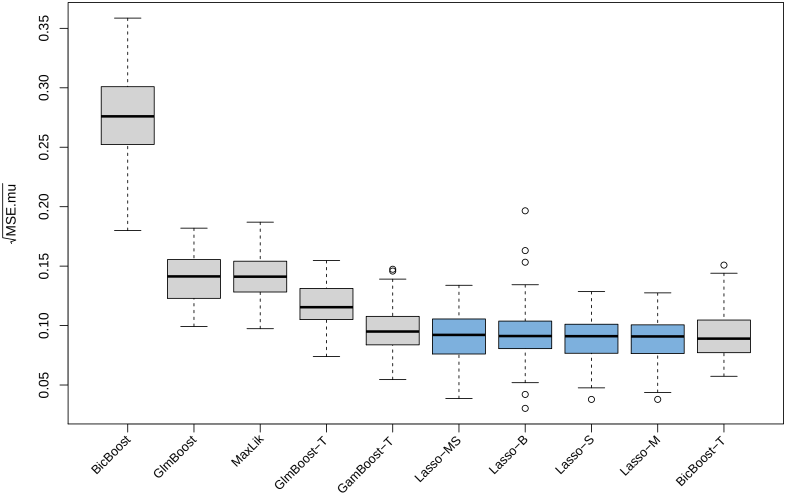
- (unregularized) Maximum-Likelihood.
- Gradient boosting with m_{stop} determined by BIC. (*bamlss*)
- Gradient boosting with m_{stop} determined by out-of-sample prediction error (*glmboostLSS* / *gamboostLSS*).
- Fused LASSO with global λ determined by BIC.
- Fused LASSO with $\lambda_\mu, \lambda_\sigma$ determined by BIC.
- Fused LASSO with separate λ's for each predictor term.
- Combination of gradient boosting and (fused) LASSO.

Goodness-of-fit measures

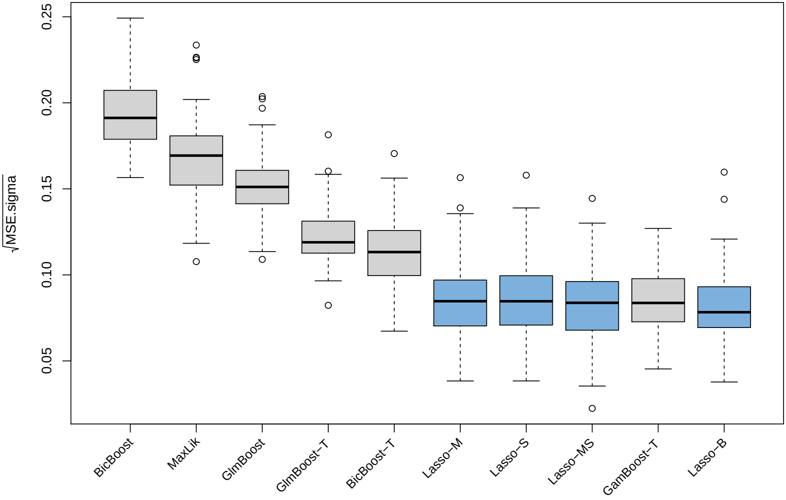
The different models are compared (amongst others) w.r.t. the following criteria:

- $\text{MSE}_{\boldsymbol{\beta}_\mu} = ||\boldsymbol{\beta}_\mu - \hat{\boldsymbol{\beta}}_\mu||^2, \quad \text{MSE}_{\boldsymbol{\beta}_\sigma} = ||\boldsymbol{\beta}_\sigma - \hat{\boldsymbol{\beta}}_\sigma||^2$
- False positives of differences.
- False positives of pure noise variables.
- False negatives of non-noise variables.

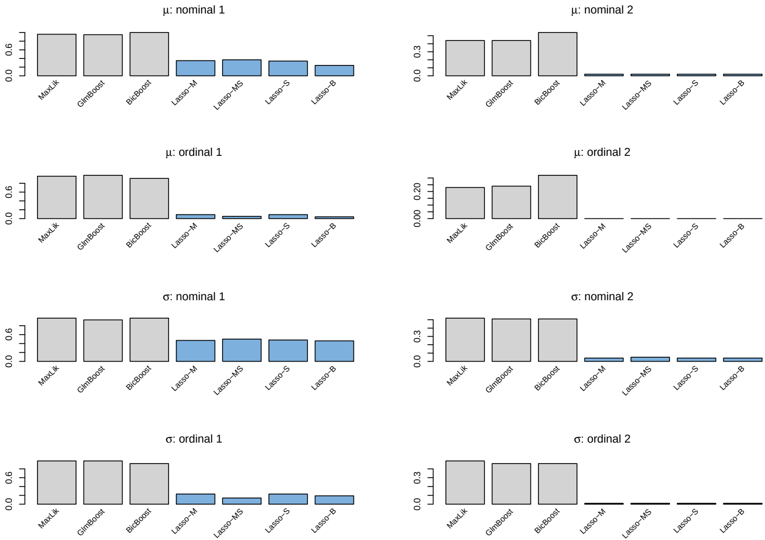
Results: MSE_{β_μ}



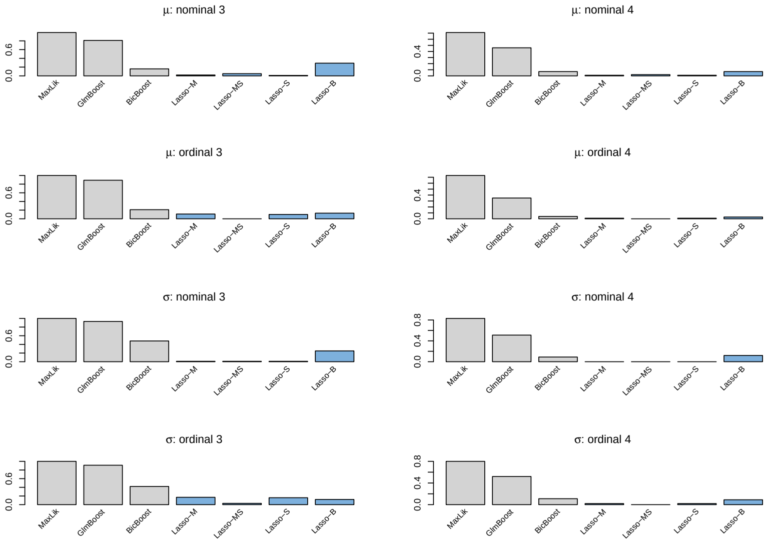
Results: $\text{MSE}_{\beta_\sigma}$



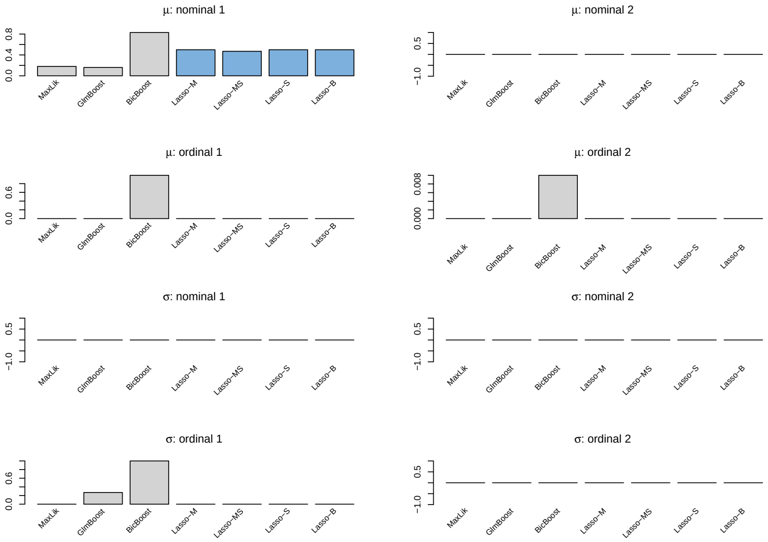
Results: FP of differences



Results: FP of pure noise variables



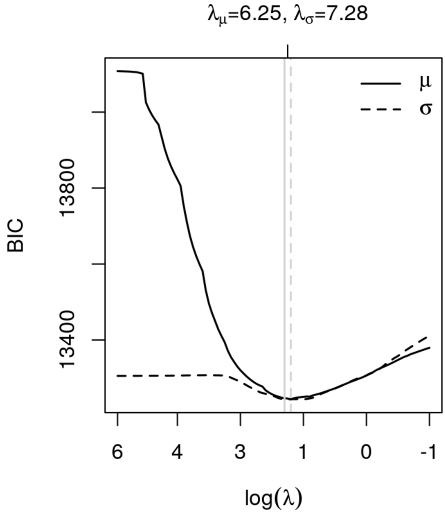
Results: FN of non-noise variables



The Munich rent data 2007

- *Munich rent standard data:* used as a reference for the average rent of a flat depending on its characteristics and spatial features.
- $n = 3015$ households.
- *Response:* monthly rent per square meter (in €).
- *Covariates:* out of a large set we incorporate a selection of 9 factors (ordered and nominal/binary; similar to Gertheiss and Tutz, 2010).
- *Model:* Gaussian GAMLSS and use for both distribution parameters, i.e. μ and σ, a combination of the two different fused LASSO penalties.
- *Optimal tuning parameters:* via BIC on a 2-dimensional grid.

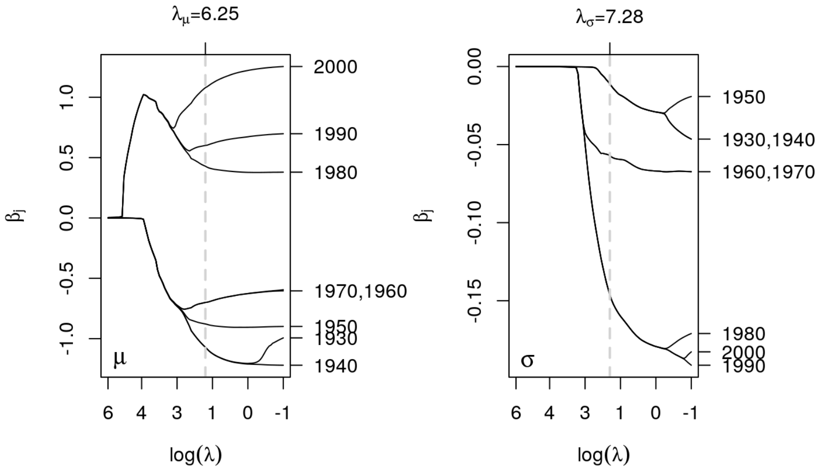
The Munich rent data 2007



Marginal BIC curves for μ & σ

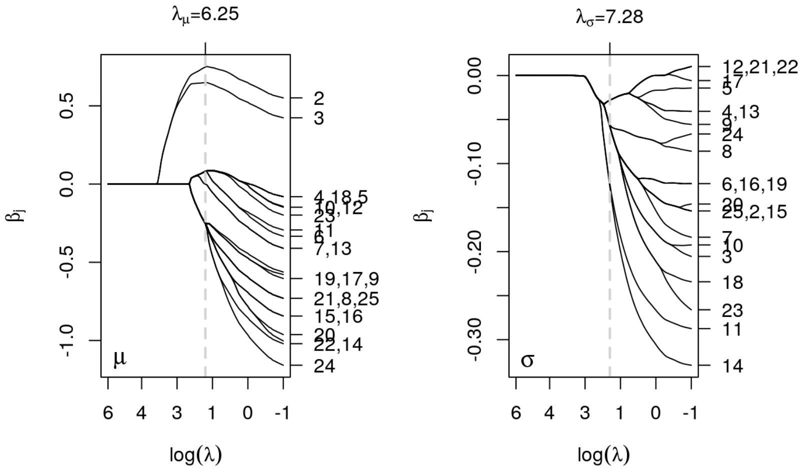
(in each case fixing the other tuning parameter at its respective BIC-minimum)

The Munich rent data 2007



Ordinal fused coefficient paths for the **year of construction** for parameters μ (left) and σ (right).

The Munich rent data 2007



Nominal fused coefficient paths for the **district** effect for parameters μ (left) and σ (right).

Summary & Conclusions

- Different LASSO-type penalties for the GAMLSS have been proposed.
- In particular, good results of the fused LASSO are obtained if clustering of categories is desirable/necessary.
- Boosting methods turned out to be problematic, if categories are clustered.
- Reasonable results for the application on the Munich rent data.
- Implementation available in the R-package *bamlss*.
- Further investigation of the combination of boosting and LASSO.

References & Software

- Gertheiss, J. & Tutz, G. (2010). Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics*, **4**(4), 2150 – 2180.
- Hofner, B., A. Mayr, & M. Schmid (2016). **gamboostLSS**: An R package for model building and variable selection in the GAMLSS framework. *Journal of Statistical Software* **74**(1), 1 – 31.
- Mayr, A., Fenske, N., Hofner, B., Kneib, T., & Schmid, M.(2012). Generalized additive models for location, scale and shape for high-dimensional data - a flexible approach based on boosting. *Journal of the Royal Statistical Society, Series C*, **61**(3), 403 – 427.
- Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society B* **70**(1), 53 – 71.
- Oelker, M.-R. & Tutz, G. (2017). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification* **11**(1), 97 – 120.

References & Software

- Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54(3)**, 507 – 554.
- Stasinopoulos, D.M. & Rigby, R.A. (2007) Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software* **23**(7).
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58* 267 – 288.
- Umlauf, N., Klein, N., & Zeileis, A. (2017a). BAMLSS: Bayesian additive models for location, scale and shape (and beyond). Working Paper, Faculty of Economics and Statistics, University of Innsbruck.
- Umlauf, N., Klein, N., Zeileis, A. & Köhler, M. (2017b). **bamlss**: Bayesian additive models for location, scale and shape (and beyond). R package version 0.1-2, URL: http://cran.r-project.org/package=bamlss.

Thank you for your attention!

Nikolaus Umlauf

https://eeecon.uibk.ac.at/~umlauf/

Weights for fused LASSO

Along the lines of Gertheiss and Tutz (2010), for the fused LASSO the following weights are used:

For **nominal** factors:

$$w_{lm}^{(jk)} = \frac{2}{|\beta_{jkl}^{ML} - \beta_{jkm}^{ML}|\,(c_{jk}+1)c_{jk}} \sqrt{\frac{n_l^{(jk)} + n_m^{(jk)}}{n}}$$

For **ordinal** factors:

$$w_{l}^{(jk)} = \frac{1}{|\beta_{jkl}^{ML} - \beta_{jk,l-1}^{ML}|\, c_{jk}} \sqrt{\frac{n_l^{(jk)} + n_{l-1}^{(jk)}}{n}}$$

Here, $n_l^{(jk)}$ denotes the number of observations on level l of predictor $\mathbf{x}_{jk}$.