Computational Statistics and Data Analysis 140 (2019) 59–73

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

LASSO-type penalization in the framework of generalized additive models for location, scale and shape

Andreas Groll [a], Julien Hambuckers [b,c], Thomas Kneib [b], Nikolaus Umlauf [d,*]

[a] Faculty of Statistics, Technische Universität Dortmund, 44221 Dortmund, Germany
[b] Chairs of Statistics, Universität Göttingen, Humboldtallee 3, 37073 Göttingen, Germany
[c] Finance Department, HEC Management School, University of Liège, Rue Louvrex, 14, 4000 Liège, Belgium
[d] Department of Statistics, Faculty of Economics and Statistics, Universität Innsbruck, Universitätsstr. 15, 6020 Innsbruck, Austria

ARTICLE INFO

Article history:
Received 3 September 2018
Received in revised form 7 June 2019
Accepted 15 June 2019
Available online 26 June 2019

Keywords:
GAMLSS
Distributional regression
Model selection
LASSO
Fused LASSO

ABSTRACT

For numerous applications, it is of interest to provide full probabilistic forecasts, which are able to assign plausibilities to each predicted outcome. Therefore, attention is shifting constantly from conditional mean models to probabilistic distributional models capturing location, scale, shape and other aspects of the response distribution. One of the most established models for distributional regression is the generalized additive model for location, scale and shape (GAMLSS). In high-dimensional data set-ups, classical fitting procedures for GAMLSS often become rather unstable and methods for variable selection are desirable. Therefore, a regularization approach for high-dimensional data set-ups in the framework of GAMLSS is proposed. It is designed for linear covariate effects and is based on L_1-type penalties. The following three penalization options are provided: the conventional least absolute shrinkage and selection operator (LASSO) for metric covariates, and both group and fused LASSO for categorical predictors. The methods are investigated both for simulated data and for two real data examples, namely Munich rent data and data on extreme operational losses from the Italian bank UniCredit.

1. Introduction

A model class that has gained increasing attention in recent years is the class of generalized additive models for location, scale and shape (GAMLSS), introduced by Rigby and Stasinopoulos (2005). In contrast to conventional regression approaches where the mean is regressed, the GAMLSS framework allows to simultaneously model all distribution parameters (as, for example, the location, scale and shape) in terms of covariates. Within the corresponding predictors, parametric and/or additive nonparametric (smooth) functions of the explanatory variables and/or random-effects terms can be included. In general, the (non)parametric models are fitted via maximum (penalized) likelihood estimation. In particular, Newton–Raphson or Fisher scoring algorithms can be used to maximize the (penalized) likelihood.

GAMLSS represent a very general regression-type model in which both the systematic and random parts of the model are highly flexible: the distribution of the response variable does not have to belong to the exponential family, can be continuous or discrete, as well as highly skewed or feature excess kurtosis (Stasinopoulos and Rigby, 2007). However, in high-dimensional data set-ups classical fitting procedures for the GAMLSS often become rather unstable and methods for

* Corresponding author.
E-mail addresses: groll@statistik.tu-dortmund.de (A. Groll), jhambuc@uni-goettingen.de (J. Hambuckers), tkneib@uni-goettingen.de (T. Kneib), Nikolaus.Umlauf@uibk.ac.at (N. Umlauf).

https://doi.org/10.1016/j.csda.2019.06.005
0167-9473/

variable selection are desirable. In addition, the more distributional parameters are related to covariates, the further the model's complexity is increased.

The first ones who systematically addressed the issue of variable selection, i.e. the selection of a reasonably small subset of informative covariates to be included in a particular GAMLSS, were Mayr et al. (2012). They extended boosting techniques, which originated in the machine learning field, to the GAMLSS framework. The approach is called **gamboostLSS** and is based on classical gradient boosting, which they successfully adapted to the GAMLSS characteristics. Both variable selection and model choice are naturally available within their regularized regression framework. For an implementation into the statistical software R (R Core Team, 2018), see Hofner et al. (2016).

An alternative strategy for variable selection, which is mainly designed for linear covariate effects, uses L_1-type penalties. A first attempt for such a penalization-based, regularized estimation in the high-dimensional GAMLSS framework is proposed in Hambuckers et al. (2018). There, only linear effects are considered, so in fact a generalized linear model for location, scale and shape is regarded. The conventional least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) for metric covariates is then applied on Generalized Pareto distributed extreme operational loss data from the Italian bank UniCredit. For the implementation of the estimation procedure, Hambuckers et al. (2018) follow Zou and Li (2008) and Oelker and Tutz (2017) and use local quadratic approximations of the penalty terms. Relying on this approximation, the maximization problem can be linearized and solved with usual Newton methods.

If, however, some of the independent variables are categorical, some modifications to usual shrinking procedures are necessary. The present work describes a regularization approach, which is also mainly designed for linear covariate effects and is also based on L_1-type penalties, but which extends the previous approaches by including penalization strategies that are specifically designed for nominal or ordinal categorical predictors. Using adequate penalties, not only the cases of the conventional LASSO for metric covariates, but also of both the group (Meier et al., 2008) and fused LASSO (Gertheiss and Tutz, 2010) for categorical predictors are covered. The implementation of the methods is incorporated into the unified modeling architecture for distributional regression models established in Umlauf et al. (2018a), which exploits the general structure of classical generalized additive models (GAMs) and encompasses many different response distributions, estimation techniques, model terms etc. The corresponding R-package **bamlss** (Umlauf et al., 2018b) embeds many different approaches suggested in literature and software and serves as a unified conceptional "Lego toolbox" for complex regression models. Furthermore, within its framework both the implementation of algorithms for complex regression problems and the integration of already existing software are substantially facilitated.

The performances of these new methods are investigated in two extensive simulation studies and are compared to different other approaches. In the applications considered later in this work, we consider both Gaussian and generalized Pareto distributed responses. We focus on the fusion of factor levels of either nominal or ordinal factors. Different performance aspects are investigated, in particular, mean squared errors of the fitted coefficients, but also the performance with regard to factor fusion and variable selection in the presence of noise variables.

For illustration purposes, the proposed methods are also applied to two different real data sets. The first data set contains Munich rent data from the year 2007, which are used as a reference for the average rent of a flat depending on its characteristics and spatial features. We model and select the predictor effects of nine covariates describing the apartments in terms of their size, age and other characteristics related to the net rent per square meter. These data have already been analyzed in Kneib et al. (2011) and in Mayr et al. (2012), where also a more detailed description of the data can be found. The second data set is a database of 10,217 extreme operational losses from the Italian bank UniCredit, covering a period of 10 years and 7 different event types. These data have recently been analyzed in Hambuckers et al. (2018).

The article is set out as follows. In the next section, we specify the underlying fully parametric regression model framework. We then introduce different L_1-type penalties in Section 3, which are designed for different kinds of regularization. The algorithmic details related to the fitting procedures of the penalized models are presented in Section 4. Next, the performance of the different methods is investigated in simulation studies in Section 5. Then, we illustrate their applicability in the two aforementioned real data examples in Section 6. Finally, we summarize the main findings and conclude in Section 7.

2. Model specification

Along the lines of Rigby and Stasinopoulos (2005), who regard GAMLSS as a semiparametric regression-type model with both linear and smooth covariate effects, in the following we focus on the fully parametric model with solely linear effects. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be the response vector with single observations y_i, $i = 1, \ldots, n$, being conditionally independent given a set of covariates. The corresponding conditional density $f(y_i|\boldsymbol{\theta}_i)$ usually depends on several distribution parameters $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{id})^T$ that commonly represent distribution characteristics like location, scale, shape and/or kurtosis, but generally may be any of the distribution's parameters. The key feature of a GAMLSS is that each of these distribution parameters θ_k can be modeled by its own predictor η_{θ_k} for $k = 1, \ldots, d$, which, in our case, depends linearly on a set of p_k covariates together with an intercept β_{0k}. Following Mayr et al. (2012), we denote by $g_k(\cdot)$ known monotonic link functions, relating the linear predictors to their corresponding parameters θ_k. Then, a generalized linear model for location, scale and shape is given by the following set of equations

$$g_k(\theta_k) = \beta_{0k} + \sum_{j=1}^{p_k} \mathbf{x}_{jk}^T \boldsymbol{\beta}_{jk} \; = \; \eta_{\theta_k} \, . \tag{1}$$

As the covariates can be metric and/or categorical, we use the general notation $\mathbf{x}_{jk}^T\boldsymbol{\beta}_{jk}$ for a single predictor term. If the covariate is categorical, this term collects all covariate dummies and regression coefficients corresponding to the jk-th group of variables. If the covariate is metric, it reduces to a product of scalar values, i.e. $x_{jk}\beta_{jk}$. These effects are collected in the coefficient vectors $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{p_k,k})^T$, $k = 1, \ldots, d$ corresponding to the d submodels. Estimation of regression parameters can be obtained via maximization of the model's log-likelihood, given by

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\theta}_i)\,, \tag{2}$$

with vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_d^T)^T$ collecting the effects of all linear predictors $\eta_{\theta_k}, k = 1, \ldots, d$. Note that the log-likelihood (2) depends on the parameters $\boldsymbol{\beta}_{jk}$ through the relations $\theta_{ik} = g_k^{-1}(\eta_{\theta_{ik}})$.

In principle, the maximization of (2) can be carried out using Newton–Raphson or Fisher scoring algorithms. Suitable fitting schemes are implemented in the R-package **gamlss** (Stasinopoulos and Rigby, 2007) and rely on the following principle: at each iteration, backfitting steps are successively applied to all distribution parameters, using the submodel fits of previous iterations as offset values for those parameters that are not involved in the current step. However, in high-dimensional situations these fitting procedures often become highly unstable and methods for regularization variable selection are needed.

3. L_1-type penalization

In the following, different L_1-type penalties are introduced, which are designed for linear covariate effects: the conventional LASSO for metric covariates, and both group and fused LASSO for categorical covariates. The different penalization terms impose different kinds of shrinkage depending on the covariates' structure and the intentions of the modeler. In particular, the group and fused LASSO penalties are designed for nominal and ordinal categorical predictors, addressing specific characteristics of those. Altogether, a term $\lambda J(\boldsymbol{\beta})$ is subtracted from the log-likelihood (2). Here, $J(\boldsymbol{\beta})$ is a combination of (parts of) the four penalty terms introduced in this section, whereas λ is a tuning parameter that controls the overall strength of the penalties.

Classical LASSO

For (standardized) metric covariates x_{jk}, following Tibshirani (1996), the absolute value of the corresponding data (scalar) regression coefficient β_{jk} is penalized by the conventional LASSO penalty, i.e. the penalty terms have the following form

$$J_c(\beta_{jk}) = |\beta_{jk}|\,. \tag{3}$$

This penalty structure shrinks the regression coefficient towards zero. If the effect is sufficiently small, the regression coefficient can even be set exactly to zero, therefore excluding the corresponding covariate from the linear predictor η_{θ_k}. The strength of the penalization is controlled by the global penalty parameter λ: for large values of λ, only the most influential covariates are retained and all other effects are shrunken to zero. On the contrary, for lower values of λ, shrinkage is smaller and fewer coefficients are excluded from the different linear predictors $\eta_{\theta_k}, k = 1, \ldots, d$. Hence, the penalty parameter λ plays the role of a tuning parameter: it controls the number of LASSO-penalized metric covariates that are related with the distribution parameters θ_k of the response variable.

Group LASSO

For a (dummy-encoded) categorical covariate with corresponding group of dummies collected in covariate vector $\mathbf{x}_{jk}$ and vector $\boldsymbol{\beta}_{jk}$ of corresponding regression coefficients, the L_2-norm of $\boldsymbol{\beta}_{jk}$ is penalized by the group LASSO penalty (compare, e.g., Meier et al., 2008), i.e. the single penalty terms yield

$$J_g(\boldsymbol{\beta}_{jk}) = \sqrt{df_{jk}} \cdot \|\boldsymbol{\beta}_{jk}\|_2\,, \tag{4}$$

where df_{jk} is the size of the jk-th group of dummy variables. The factors $\sqrt{df_{jk}}$ are used to rescale the penalty terms with respect to the dimensionality of the parameter vectors $\boldsymbol{\beta}_{jk}$, see also Yuan and Lin (2006). They ensure that the penalty terms are of the order of the number of parameters and, hence, are comparable to the conventional LASSO-penalty (3). Consequently, if $J(\boldsymbol{\beta})$ is a combination of penalties for both metric covariates from (3) and penalties for (dummy-encoded) categorical covariates from (4), still a single overall penalty parameter λ can be used.

The effect of jointly penalizing the whole group of dummies corresponding to a categorical covariate via $\|\boldsymbol{\beta}_{jk}\|_2$ is similar to the one of the conventional LASSO penalty and we either obtain $\hat{\boldsymbol{\beta}}_{jk} = \mathbf{0}$ or $\hat{\beta}_{jkl} \neq 0$ for all $l = 1, \ldots, df_{jk}$. Consequently, a categorical predictor is either included (with all its dummies) or excluded completely from its respective linear predictor η_{θ_k}.

Fused LASSO

Alternatively, for categorical covariates, clustering of categories with implicit factor selection is desirable. Again, let $\mathbf{x}_{jk}$ be a (dummy-encoded) categorical covariate with categories $l = 0, 1, \ldots, df_{jk}$ and corresponding coefficient vector $\boldsymbol{\beta}_{jk}$. By choosing $l = 0$ as the reference category, we fix $\beta_{jk0} = 0$. Depending on the nominal or ordinal scale level of the covariate, one of the following two penalties can be used (compare Gertheiss and Tutz, 2010). For nominally scaled covariates, all possible pairwise differences of the regression effects are penalized by the fused LASSO penalty, for which the individual penalty terms are given by

$$J_{\text{fn}}(\boldsymbol{\beta}_{jk}) = \sum_{\substack{l>m \\ l,m\leq df_{jk}}} w_{lm}^{(jk)} \, |\beta_{jkl} - \beta_{jkm}|. \tag{5}$$

For ordinally scaled covariates, only the differences of neighboring regression effects are penalized. In this case, the penalty terms can be specified by

$$J_{\text{fo}}(\boldsymbol{\beta}_{jk}) = \sum_{l=1}^{df_{jk}} w_{l}^{(jk)} \, |\beta_{jkl} - \beta_{jk,l-1}| \, , \tag{6}$$

where df_{jk} is the number of (free) dummy coefficients of the categorical predictor $\mathbf{x}_{jk}$, i.e. the number of levels minus one. Both $w_{lm}^{(jk)}$ and $w_{l}^{(jk)}$ denote suitable weights that are suggested in Bondell and Reich (2009). In principle, the use of these weights can be motivated through standardization of the corresponding design matrix part, in analogy to standardization of metric predictors. For nominal covariates we use

$$w_{lm}^{(jk)} = 2(df_{jk} + 1)^{-1} \sqrt{\frac{n_l^{(jk)} + n_m^{(jk)}}{n}},$$

where $df_{jk} + 1$ is again the number of levels of the corresponding categorical predictor $\mathbf{x}_{jk}$ and $n_l^{(jk)}$ denotes the number of observations on level l. Hence, the weights account for different numbers of levels of different predictors and for different numbers of observations on different levels.

Furthermore, notice also that an adaptive version of the weights can be used. Then, they contain additionally the factors $|\hat{\beta}_{jkl}^{(ML)} - \hat{\beta}_{jkm}^{(ML)}|^{-1}$, where $\hat{\boldsymbol{\beta}}_{jk}^{(ML)}$ denotes the unconstrained maximum likelihood (ML) estimate. The factor $(df_{jk}+1)^{-1}$ ensures that the penalties from (5) are comparable to the ordinal penalty terms from (6). For ordinal predictors, since the penalty terms (6) are already of order c_{jk}, the corresponding weights $w_l^{(jk)}$ can be chosen as

$$w_{l}^{(jk)} = \sqrt{\frac{n_l^{(jk)} + n_{l+1}^{(jk)}}{n}}.$$

Similarly to the nominal case, adaptive versions of the weights are obtained by adding the factors $|\hat{\beta}_{jkl}^{(ML)} - \hat{\beta}_{jk,l-1}^{(ML)}|^{-1}$. Due to the adequately chosen weights $w_{lm}^{(jk)}$ and $w_{l}^{(jk)}$, we can combine the penalties from (5) and (6) and still use a single penalty parameter. However, if a single penalty parameter is used, these penalties cannot be combined with those given by (3) and (4), due to differences in orders and scaling procedures. If instead different tuning parameters are used for each of the four penalty types introduced above, they all can be combined. This way, also combinations of fusion and selection are possible. Notice also that for the fusion of effects, alternative weighting schemes are used in the literature, see, for example, Chiquet et al. (2017).

Finally, note that the classical and group LASSO penalties given by (3) and (4) could be extended in a similar way by choosing suitable adaptive weights.

Some technical details

All proposed penalties have the attractive property to be able to set the coefficients of single (groups of) covariates to zero and, hence, to perform variable selection. Within the estimation procedures implemented in **bamlss**, e.g. the corresponding backfitting algorithm, local quadratic approximations of all presented penalty terms are used (see Oelker and Tutz, 2017). Furthermore, note that **bamlss** also allows to assign to each linear predictor η_{θ_k}, $k = 1, \ldots, d$, its own penalty term, i.e. a term $\lambda_k J^{(k)}(\boldsymbol{\beta}_k)$, where $J^{(k)}(\boldsymbol{\beta}_k)$ denotes the penalty term corresponding to linear predictor η_{θ_k} only. This framework allows to specify highly flexible models, although it has the drawback that the grid search for the optimal tuning parameters λ_k has to be carried out on d dimensions and, hence, becomes computationally more demanding.

Moreover, it is even possible in **bamlss** to assign to each single predictor component $\boldsymbol{\beta}_{jk}$ (or β_{jk} if x_{jk} is metric) its own penalty parameter λ_{jk}. In this case, instead of searching the tuning parameters over a multi-dimensional grid, they are implicitly determined and optimized in a stepwise manner in the backfitting procedure, as explained in the next section. If this strategy is chosen, all different penalty terms from (3)–(6) can be combined, as each term is assigned to an individual amount of penalization, and no issues regarding comparability arise.

4. Estimation

To conveniently combine the penalized estimation of GAMLSS with the LASSO-type penalties introduced in the previous section, we write the kth linear predictor of Eq. (1) for n observations in matrix notation as

$$\boldsymbol{\eta}_{\theta_k} = \beta_{0k} + \mathbf{X}_k\boldsymbol{\beta}_k = \beta_{0k} + \sum_{j=1}^{p_k}\mathbf{X}_{jk}\boldsymbol{\beta}_{jk},$$

with regression coefficients $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{1k}^\top, \ldots, \boldsymbol{\beta}_{p_k,k}^\top)^\top$ and design matrices $\mathbf{X}_k = [\mathbf{X}_{1k}, \ldots, \mathbf{X}_{p_kk}]$, where the ith row of $\mathbf{X}_{jk}$ is represented by $\mathbf{x}_{ijk}$, $i = 1, \ldots, n$. Note that the intercepts are treated separately from the rest of the parameters, since they are not suspect to shrinkage. Moreover, the presentation of the kth linear predictor is split into its p_k covariates to emphasize that these can in principle have their own shrinkage parameters. More precisely, there are three possible options which will be explained in more detail in the following: First, the use of one global shrinkage parameter λ for all distribution parameters and model terms $\mathbf{X}_{jk}\boldsymbol{\beta}_{jk}$. Second, different shrinkage parameters λ_k, one for each distribution parameter, as used in the next paragraph. Third, different shrinkage parameters λ_{jk}, one for each parameter of the distribution and each model term $\mathbf{X}_{jk}\boldsymbol{\beta}_{jk}$.

For the estimation of the coefficients $\boldsymbol{\beta}_k$ we apply a partitioned updating scheme as presented in Umlauf et al. (2018a), which maximizes the penalized log-likelihood

$$\begin{aligned}\ell_{\text{pen}}(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\log f(y_i|\boldsymbol{\theta}_i) - \sum_{k=1}^{d}\lambda_k J^{(k)}(\boldsymbol{\beta}_k)\\ &= \ell(\boldsymbol{\beta}) - \sum_{k=1}^{d}\lambda_k\boldsymbol{\beta}_k^\top\mathbf{J}_k(\boldsymbol{\beta}_k)\boldsymbol{\beta}_k\,.\end{aligned} \tag{7}$$

For clarity, $J^{(k)}(\boldsymbol{\beta}_k)$ can be rewritten as $\sum_{j=1}^{p_k}J^{(jk)}(\boldsymbol{\beta}_{jk})$, with $J^{(jk)}(\cdot)$ being one of the penalties given by Eqs. (3)–(6). Because we follow Oelker and Tutz (2017) and use local quadratic approximations in $J^{(jk)}(\cdot)$, the LASSO penalty can be written as a quadratic form with a block-diagonal *penalty matrix*:

$$\mathbf{J}_k(\boldsymbol{\beta}_k) = \text{diag}(\mathbf{J}_{1k}(\boldsymbol{\beta}_{1k}), \ldots, \mathbf{J}_{p_kk}(\boldsymbol{\beta}_{p_kk})).$$

The $\mathbf{J}_{jk}(\cdot)$ are penalty matrices determined by (the approximations of) the penalties from Eqs. (3)–(6). In this setting, for each distribution parameter θ_k we begin by penalizing the contribution of the jk-th covariate with the corresponding shrinkage parameter λ_k (note again that this is possible if the various LASSO-type penalties are appropriately scaled. Applying d different penalties instead of a single one for all distribution parameters is reasonable, since different parameters θ_k are associated with different scalings with respect to the response, and may be dependent on (possibly) different sets of covariates and/or fused categories).

Then, for fixed values of λ_k, the algorithm cycles over each model component with a Newton–Raphson-type updating step. For iteration $t+1$, the updating step for the penalized coefficients is given by

$$\boldsymbol{\beta}_k^{(t+1)} = (\mathbf{X}_k^\top\mathbf{W}_k\mathbf{X}_k + \lambda_k\mathbf{J}_k(\boldsymbol{\beta}_k))^{-1}\mathbf{X}_k^\top\mathbf{W}_k(\mathbf{z}_k - \tilde{\boldsymbol{\eta}}_{\theta_k}), \tag{8}$$

with working observations $\mathbf{z}_k = \boldsymbol{\eta}_{\theta_k}^{(t)} + \mathbf{W}_k^{-1}\mathbf{u}_k^{(t)}$, derivatives $\mathbf{u}_k = \partial\ell_{\text{pen}}(\boldsymbol{\beta})/\partial\boldsymbol{\eta}_{\theta_k}$, weights $\mathbf{W}_k = -\text{diag}(\partial^2\ell_{\text{pen}}(\boldsymbol{\beta})/\partial\boldsymbol{\eta}_{\theta_k}\partial\boldsymbol{\eta}_{\theta_k}^\top)$ and partial predictor $\tilde{\boldsymbol{\eta}}_{\theta_k} = \boldsymbol{\eta}_{\theta_k}^{(t+1)} - \beta_{0k}^{(t+1)}$ (see Umlauf et al., 2018a for a detailed description of the algorithm). The intercepts β_{0k} are updated similarly, but without the penalty terms $\lambda_k\mathbf{J}_k$ in (8) and $\tilde{\boldsymbol{\eta}}_{\theta_k} = \boldsymbol{\eta}_{\theta_k}^{(t+1)} - \mathbf{X}_k\boldsymbol{\beta}_k^{(t+1)}$. To estimate the optimal values for each λ_k, a simple grid search based on minimizing an information criterion (e.g., the BIC) is carried out, where the model complexity (i.e. the amount of shrinkage) is measured by the effective degrees of freedom for each model term. Effective degrees of freedom (edfs) can be approximated by

$$\text{edf}_k(\lambda_k) := \text{trace}\left[\mathbf{X}_k^\top\mathbf{W}_k\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{W}_k\mathbf{X}_k + \lambda_k\mathbf{J}_k(\boldsymbol{\beta}_k))^{-1}\right],$$

such that the total effective degrees of freedom can be approximated by $d + \sum_{k=1}^{d}\text{edf}_k(\lambda_k)$. In practice, the algorithm starts by initializing the intercepts and setting λ_k to very large values, such that $\boldsymbol{\beta}_k \approx \mathbf{0}$, $k = 1, \ldots, d$. Then, the coefficients $\boldsymbol{\beta}_k$, as well as β_{0k}, are re-estimated by slightly decreasing λ_k and using $\hat{\boldsymbol{\beta}}_k$ from the previous λ_k as starting values. This procedure is usually relatively fast and numerically stable, even for complicated GAMLSS models.

However, this approach has two drawbacks. First, grid search estimation for λ_k when $d > 2$ can still be time and computer memory intensive. Second, for complex combinations of penalty terms (3)–(6), a single λ_k for each distributional parameter is most likely not sufficient or even improper, since the order and scaling procedures of the different penalty terms are not comparable. In such cases we extend the penalty in (7) for the kth distribution parameter to $\sum_{j=1}^{p_k}\lambda_{jk}J^{(jk)}(\boldsymbol{\beta}_{jk})$ and estimate each λ_{jk} using a stepwise procedure in each updating iteration (8), see (Umlauf et al., 2018a; Algorithm A2). Besides, by further partitioning the updating scheme (8) for each model component $\mathbf{X}_{jk}\boldsymbol{\beta}_{jk}$, sparse matrix structures are exploited within (8), which lead to significant runtime improvements for large data sets (see also Lang et al., 2014 for more details on highly efficient updating schemes).

Table 1
True nonzero dummy coefficient vectors used in the simulations.

Distribution	Parameter	Type of factor	Values
Gaussian	μ	Nominal (1)	$(0, 0.5, 0.5, 0.5, 0.5, -0.2, -0.2)^\top$
		Nominal (2)	$(0, 1, 1)^\top$
		Ordinal (1)	$(0, 0.5, 0.5, 1, 1, 2, 2)^\top$
		Ordinal (2)	$(0, -0.3, -0.3)^\top$
	σ	Nominal (1)	$(0, -0.5, 0.4, 0, -0.5, 0.4, 0)^\top$
		Nominal (2)	$(0.4, 0, 0.4)^\top$
		Ordinal (1)	$(0, 0, 0.4, 0.4, 0.4, 0.8, 0.8)^\top$
		Ordinal (2)	$(0, -0.5, -0.5)^\top$
Generalized Pareto	ξ	Nominal (1)	$(0, 0.3, 0.3, 0.3, 0.3, -0.5, -0.5)^\top$
		Nominal (2)	$(0, -0.4, -0.4)^\top$
		Ordinal (1)	$(0, -0.4, -0.4, -0.8, -0.8, -1.1, -1.1)^\top$
		Ordinal (2)	$(0, -0.5, -0.5)^\top$
	σ	Nominal (1)	$(0, -0.6, 0.3, 0, -0.6, 0.3, 0)^\top$
		Nominal (2)	$(0.4, 0, 0.4)^\top$
		Ordinal (1)	$(0, 0, -0.4, -0.4, -0.4, -0.9, -0.9)^\top$
		Ordinal (2)	$(0, -0.3, -0.3)^\top$

Note that updating scheme (8) can also be used as a weak base learner to build a component-wise gradient boosting algorithm (Mayr et al., 2012; Thomas et al., 2018) for the fusion penalties presented in Section 3. This technique has the advantage that each model term $\mathbf{X}_{jk}\boldsymbol{\beta}_{jk}$ can have a different amount of shrinkage and that gradient boosting does not need to compute the weights $\mathbf{W}_k$, since linear models are only fitted on the negative gradient $-\partial\ell_{\text{pen}}(\boldsymbol{\beta})/\partial\boldsymbol{\eta}_{\theta_k}$. Therefore, such an approach works even faster when the choice of the optimal stopping iteration is based on, e.g., the BIC with the total edfs computed from the active set (Zou et al., 2007), i.e. the number of non-zero coefficients.

Eventually, a last practical issue can arise when computing the adaptive weights in $J^{(jk)}(\cdot)$. Indeed, in a high-dimensional GAMLSS setting, the unregularized maximum likelihood (ML) estimator might simply not exist. In this situation, we suggest to use gradient boosting with ridge-type penalties $J_{jk}(\boldsymbol{\beta}_{jk}) = \|\boldsymbol{\beta}_{jk}\|_2^2$ to obtain $\hat{\boldsymbol{\beta}}_{jk}^{(ML)}$, since gradient boosting is one of the most stable algorithms in complex modeling problems.

Altogether, we focus on the following four approaches and compare them in the simulation study of the next section:

1. `MaxLik`: Unpenalized maximum likelihood estimation.
2. `Lasso-S`: Backfitting algorithm with LASSO penalties (see Section 4) with one single shrinkage parameter λ for all parameters of the distribution.
3. `Lasso-M`: Backfitting algorithm with LASSO penalties (see Section 4) with two shrinkage parameters λ_k, one for each parameter of the distribution. Optimal λ_k-s are selected using a two-dimensional grid search.
4. `Lasso-MS`: Backfitting algorithm with LASSO penalties (see Section 4) with single shrinkage parameters λ_{jk}, one for each model term. Optimal λ_{jk}-s are selected using a stepwise selection algorithm (see Umlauf et al., 2018a).

5. Simulation

For the investigation of the fusion LASSO penalties within the framework of GAMLSS we follow the application data from Section 6 and consider two scenarios: The first simulation setting is based on simulated Gaussian responses, the second on the generalized Pareto distribution. For both settings we model different covariate effects on all distributional parameters, i.e. for μ and σ in the Gaussian setting and for ξ (shape parameter) and σ (scale parameter) in the generalized Pareto setting. In total 150 replications for each distribution are simulated.

Similar to Gertheiss and Tutz (2010), for each setting, we use 4 informative covariates and 4 non-informative covariates for each distributional parameter θ_k, i.e., in total 16 covariates for parameters μ and σ in the Gaussian case, and the same for parameters ξ and σ in the generalized Pareto simulation. For each parameter θ_k the informative variables are split into 2 nominal and 2 ordinal factor variables. The same setting is used for the noise variables. Table 1 summarizes the true nonzero dummy coefficients used in the simulations. Here, all predictors have several levels with equal effects that actually could be fused. The performance of the proposed LASSO-type estimation method is now compared to several competing methods, which were introduced in Section 4.

For the estimation of the optimal tuning parameters we use the Akaike information criterion (AIC; Sakamoto et al., 1986) and the Bayesian information criterion (BIC; Schwarz, 1978). We additionally compare the LASSO models with several gradient boosting methods (Mayr et al., 2012; Thomas et al., 2018), using either the original categorical variables including all levels, only, or together with the true fused categories to see whether boosting algorithms can select the true fused effects. The performance of the boosting algorithms is generally inferior compared to the LASSO in this setting, therefore the results are not shown. However, we also implemented the fusion penalty with adaptive weights for boosting which in principle showed good performance, but selecting the stopping iteration turned out to be problematic. Selecting

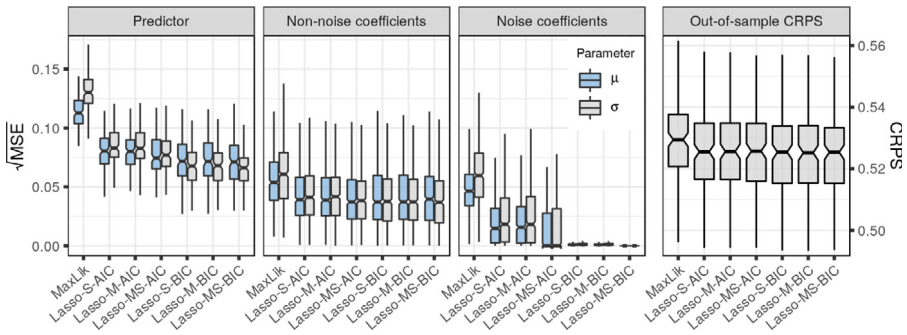

Fig. 1. Gaussian simulation study, performance of the applied algorithms in terms of $\sqrt{\text{MSE}}$ and out-of-sample CRPS. The first three plots (from left to right) show the $\sqrt{\text{MSE}}$ of the linear predictor, the non-noise coefficients and the noise coefficients, respectively. The right plot shows the out-of-sample CRPS.

the stopping iteration from out-of-sample data (e.g., using proper scoring rules, Gneiting et al., 2007) seems to be inferior compared to selecting the stopping iteration using, e.g., the BIC. Since the computation of the BIC involves calculation of the equivalent degrees of freedom, one can rely on the active set (the number of non-zero coefficients, as proposed by Zou et al., 2007) or approximate these using the trace of recursively defined Hessians. The active set leads to very sparse models with poor performance in the distributional regression setting, the second option indicates quite good model performance, but the computation time is extremely long and therefore unfeasible in most applications (already with 2000 observations computations are cumbersome). In summary, although gradient boosting is a quite stable optimization algorithm, in our findings the estimation of the stopping criterion for high-dimensional data sets in a distributional regression setting remains difficult and needs to be investigated more thoroughly in future research. For reproducibility we provide all R-scripts within the supplemental materials.

For the Gaussian simulation, we use 500 observations for training the models. In the generalized Pareto simulation, we use 1500, 3000, 6000 and 15,000 observations for estimation. The second setting uses different numbers of observations since the generalized Pareto is not an easy distribution to model. For example, Hambuckers et al. (2018) encountered problems if the number of observations is small. We chose this setup in order to investigate how sensible the estimation of the generalized Pareto model is, if the sample size becomes small. We evaluate performance of the different settings using the root mean squared error ($\sqrt{\text{MSE}}$) of the true and estimated linear predictors η_k. The out-of-sample predictive performance is evaluated using the continuous rank probability score (CRPS; Gneiting et al., 2007), a score especially suited for evaluating probabilistic forecasts accounting for calibration and sharpness simultaneously. In addition, to compare the performance of the different fused LASSO penalties, we compute false positive rates of true zero differences between coefficients $\boldsymbol{\beta}_{jk}$, i.e., for true fused categories we calculate the percentage rate of nonzero differences over all replications. To investigate the variable selection performance, we calculate false positive rates of the noise variable coefficients. Similarly, we calculate false negative rates of true non-noise coefficients.

The results for the Gaussian simulation generally show the best performance for the fused LASSO penalties using the BIC for tuning parameter selection. According to the $\sqrt{\text{MSE}}$ displayed in Fig. 1 in the first two panels, especially for the scale parameter σ the LASSO with single tuning parameters for each model term (`Lasso-MS-BIC`) seems to be the best method. In the third panel, the $\sqrt{\text{MSE}}$ of the noise coefficients is clearly the smallest for all BIC-tuned models. The out-of-sample CRPS in the right panel also indicates better performance for the LASSO models. However, the improvement compared to ML estimation is relatively small in this simulation setting.

False positive rates of truly zero differences are shown in Fig. 2 top row. Clearly, unpenalized ML (`MaxLik`) cannot anticipate the fused categories, but also LASSO using the AIC seems to approximate the true fused categories inferiorly. Similarly, concerning false positive rates of true noise coefficients, Fig. 2 in the second row shows that `MaxLik` and AIC-based LASSO are not performing as good as the LASSO with BIC tuning. In the bottom row of Fig. 2, false negative rates of non-noise coefficients show relatively equal results for all methods with slight indication of higher shrinkage of the LASSO when looking at the first nominal fused covariate for μ.

The results of the simulation with the generalized Pareto are in principle similar. Fig. 3, top row, shows that the LASSO using the BIC for tuning parameter selection has the smallest $\sqrt{\text{MSE}}$ according to the predictors for ξ and σ. Except for models using more than 3000 observations, here the differences in $\sqrt{\text{MSE}}$ are not that prominent for all methods and distributional parameters. The $\sqrt{\text{MSE}}$'s of non-noise and noise coefficients in Fig. 3, middle and bottom row, are principally similar to the ones observed in the Gaussian simulation. Moreover, results are rather similar for all numbers of observations, except that for large sample sizes ($n = 6000, n = 15{,}000$) also `MaxLik` yields satisfactory results for the non-noise coefficients. Concerning the out-of-sample predictive performance evaluated using the CRPS shown in Fig. 4, the LASSO methods using BIC tuning clearly outperform the other methods.

The false positive rates in Fig. 5, top row, are again the lowest for all LASSO-type penalties using the BIC, although it seems that in the generalized Pareto case it is more difficult to fuse categories, especially for nominal variables with

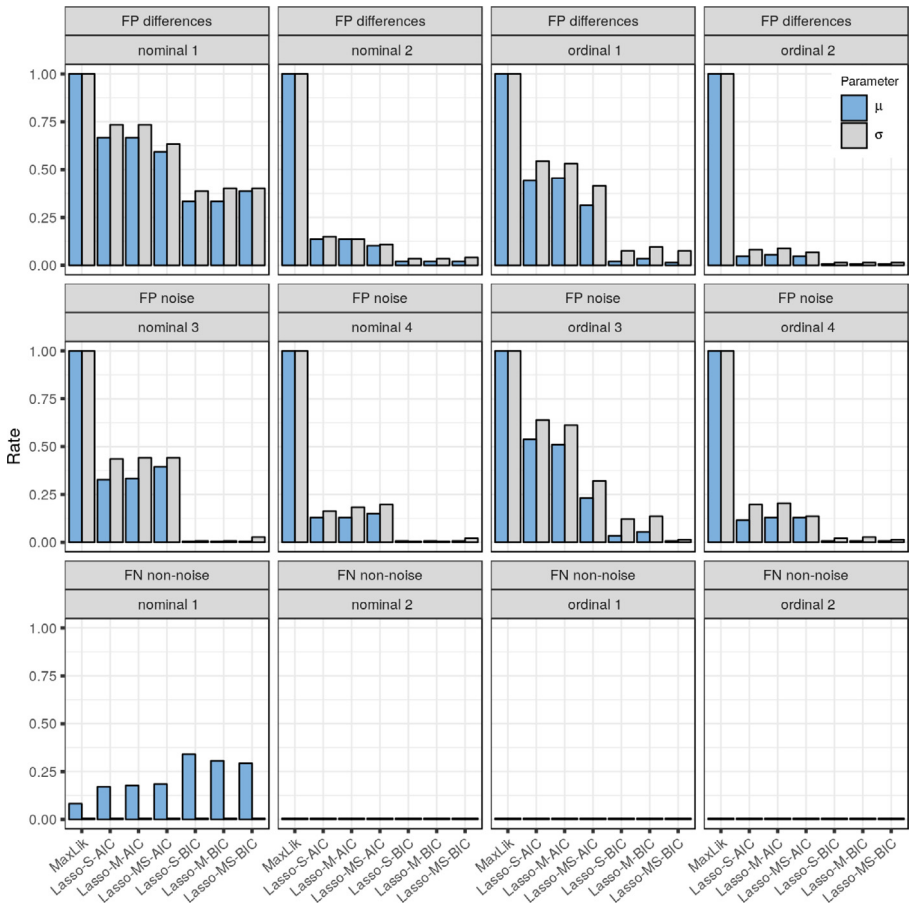

Fig. 2. Gaussian simulation study. The top row presents false positive (FP) rates of truly zero differences, the middle row false positive rates of pure noise coefficients, the bottom row false negative (FN) rates of non-noise coefficients.

many (fused) levels. The false positive rates of pure noise coefficients in Fig. 5, middle row, again show that all shrinkage methods using BIC tuning detect correctly the fused categories, whereas for parameter σ shrinkage seems to be more difficult in this setting.

In Fig. 5, bottom row, the false negative rates of non-noise coefficients indicate that in the generalized Pareto model all shrinkage methods show a good performance despite being more conservative for parameter ξ.

To sum up, for both regarded response scenarios the BIC-based LASSO outperforms the other approaches with respect to all performance criteria except that it is more conservative.

6. Applications

In this section we apply the proposed penalization approaches to two different real data sets, namely to Munich rental guide data and to data on extreme operational losses of the Italian bank UniCredit. Due to the categorical covariate structure of both data sets, we focus on the approach that turned out to be the most suitable one for this setting in the simulation studies from the previous section (in particular in terms of $\sqrt{MSE}$ and the out-of-sample predictive performance using the CRPS, see Figs. 3 and 4), which is the fused LASSO penalty approach using the BIC for selection of optimal tuning parameters.

6.1. Munich rental guide data

We now apply the proposed penalization approaches to the Munich rent data, which stem from 3015 households interviewed for the Munich rent standard 2007. The response is the monthly rent per square meter in Euro. From a large set of covariates, we incorporate a selection of nine factors describing certain characteristics of the flats, such as e.g. the quality of the bathroom equipment or the number of rooms, similar to Gertheiss and Tutz (2010). All of those covariates are considered in the form of categorical factors, which are both ordered and nominal, as well as binary, and are

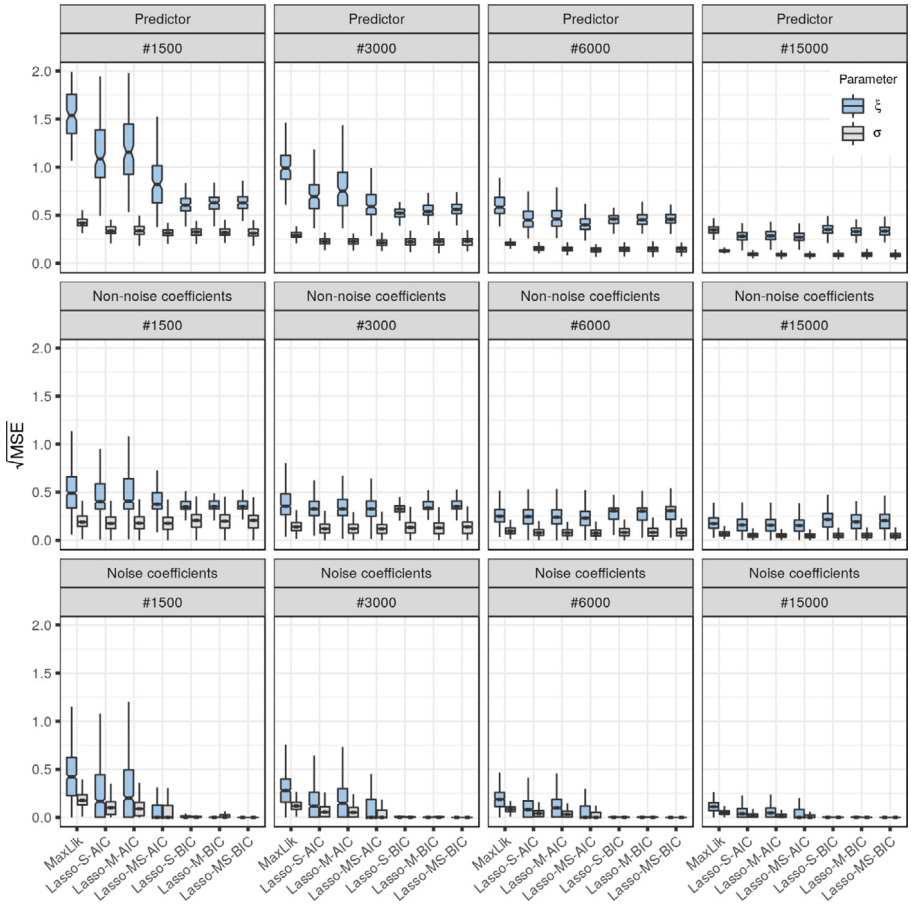

Fig. 3. Generalized Pareto simulation study, performance of the applied algorithms in terms of $\sqrt{\text{MSE}}$. The top row shows the $\sqrt{\text{MSE}}$ of the linear predictor, the middle row of the non-noise coefficients and the bottom row of the noise coefficients, respectively. The columns show results for the different numbers of observations (`#nobs`).

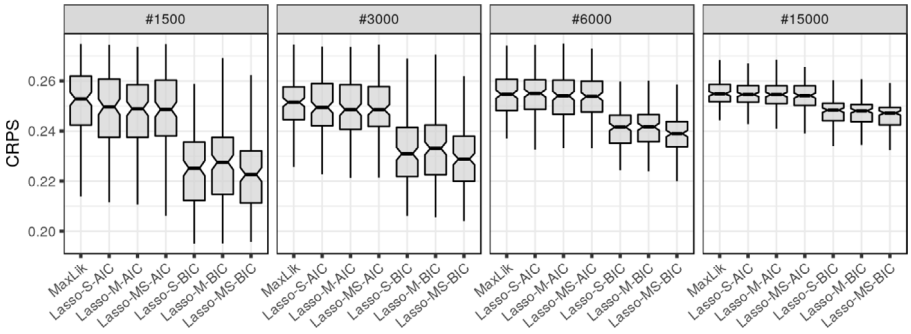

Fig. 4. Generalized Pareto simulation study, performance of the applied algorithms in terms of the out-of-sample CRPS for different numbers of observations (`#nobs`).

standardized as explained in Section 3. The two continuous covariates *size of the flat* and *year of the building's construction* were categorized. A short overview of the data set is found in Table 2, while a more detailed description can be found in Kneib et al. (2011) and Mayr et al. (2012).

We fit a Gaussian GAMLSS and use for both distribution parameters, i.e. μ and σ, a combination of the two different fused LASSO penalties introduced above. In particular, the use of the `Lasso-M` approach allows for a flexible fit, as the penalty terms of both corresponding linear predictors are assigned with separate tuning parameters λ_μ and λ_σ, respectively.

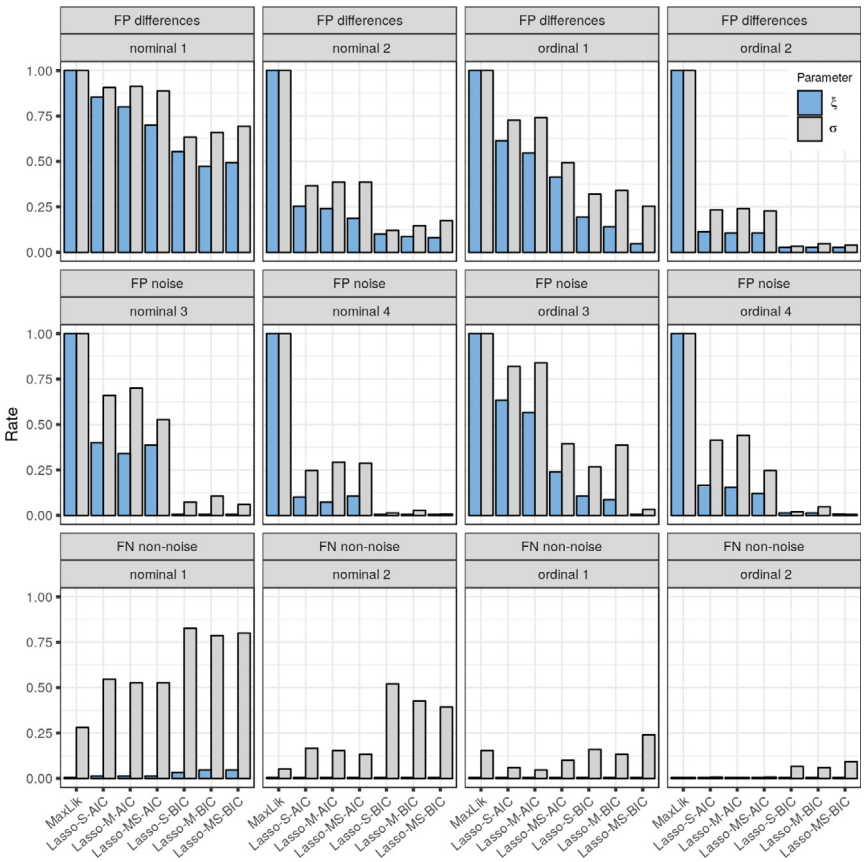

Fig. 5. Generalized Pareto simulation study, aggregating over all sample sizes $n \in \{1500, 3000, 6000, 15{,}000\}$. The top row presents false positive (FP) rates of truly zero differences, the middle row false positive rates of pure noise coefficients, the bottom row false negative (FN) rates of non-noise coefficients.

Table 2
Response variable and selection of covariates from the Munich rental guide data.

Variable	Description
`rentsqm`	Rent per square meters (continuous; response variable)
`district`	Id number of district (categorical; 25 levels)
`yoc`	Building's construction year (categorical; $\in \{[1920, 1930), \ldots, [2000, 2010)\}$)
`rooms`	Number of rooms of the flat (categorical; $\in \{1, \ldots, 7\}$)
`rarea`	Rent area (categorical; $\in \{\textit{fair}, \textit{good}, \textit{excellent}\}$)
`fspace`	Flat size in m^2 (categorical; $\in \{[0, 30), [30, 40), \ldots, [130, 140), [140, \text{inf})\}$)
`water`	Warm water supply (binary; $\in \{\textit{yes}, \textit{no}\}$)
`cheating`	Central heating (binary; $\in \{\textit{yes}, \textit{no}\}$)
`tbath`	Separate bathroom (binary; $\in \{\textit{yes}, \textit{no}\}$)
`kitchen`	Quality of kitchen (binary; $\in \{\textit{normal}, \textit{good}\}$)

The optimal tuning parameters are selected by BIC on a 2-dimensional grid. The left panel of Fig. 6 shows the corresponding marginal BIC curves for both μ and σ, in each case holding the other tuning parameter fixed at the respective minimum of the BIC. For the final model, the middle and right panels indicate a relatively good model fit according to the resulting randomized quantile residuals (Dunn and Smyth, 1996). Only for rents below the 1% and above the 99% quantile the model seems to be less appropriate and could possibly be further improved, e.g., by selecting other categories or response distributions (this would go beyond the scope and is therefore not presented). Figs. 7 and 8 show the paths of the dummy coefficients of both the ordinal covariate *year of construction* and the nominal *district*, which are penalized by the two different fused LASSO penalties from above. It is seen that with increasing tuning parameters λ_μ and λ_σ, respectively, categories are successively fused, i.e. the coefficients are set equal. In addition, it can be seen that for the ordinal covariate *year of construction* in Fig. 7 only neighboring coefficients are fused, while for the nominal factor *district* in Fig. 8 any groups of coefficients can be aggregated.

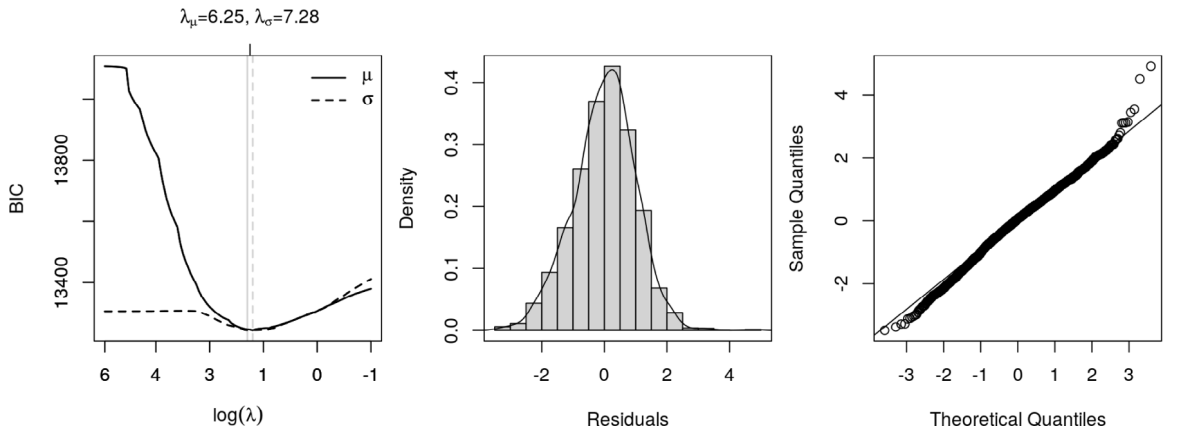

Fig. 6. The left panel shows marginal BIC curves for parameters μ and σ, holding the other tuning parameter fixed at the respective minimum of the BIC. The middle panel shows a histogram together with a kernel density estimate of the resulting randomized quantile residuals, the right panel the corresponding quantile–quantile plot.

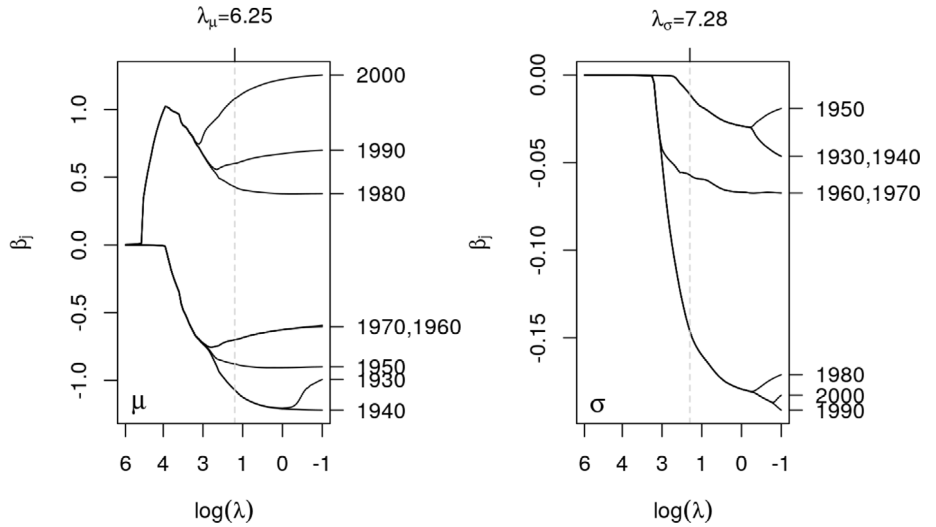

Fig. 7. Ordinal fused coefficient paths for the year of construction for parameters μ (left) and σ (right); vertical dashed lines: optimal tuning parameters.

At the optimal values of the tuning parameters, both several (neighboring) years of construction and several districts are fused and a much less complex model is obtained compared to the (unregularized) ML estimator. Similar fusion could also be observed on the seven remaining categorical predictors (not shown here). Altogether, the fused LASSO approach detects the decisive number of different categories per predictor and yields a sparse model that is much easier to interpret in comparison to the unrestricted model. Potentially, it can even exclude irrelevant factors completely from the model.

6.2. UniCredit loss data

Here, we study the same data as in Hambuckers et al. (2018). This data set consists of 10,217 extreme operational losses registered by the Italian bank UniCredit, between January 2005 and June 2014. Operational losses in the banking industry are defined as "losses resulting from inadequate or failed internal processes, people and systems or from external events" (Basel Committee on Banking Supervision (BCBS), 2004). Examples include losses related to unauthorized trading, legal disputes with employees, sales malpractices or cyber attacks. For regulatory and risk management purposes, banks have an interest in adequately modeling the density of these losses, so that they can compute appropriate risk indicators (e.g. quantiles or moments). These risk indicators are used later on to determine the requested operational risk capital (Basel Committee on Banking Supervision (BCBS), 2004). To reflect properly the probability of tail events, a generalized Pareto distribution is usually assumed, in the framework of Extreme Value Theory (EVT, see, e.g., Embrechts et al., 1997; Chapelle et al., 2008; Chavez-Demoulin et al., 2016 and Hambuckers et al., 2018). Recently, researchers have

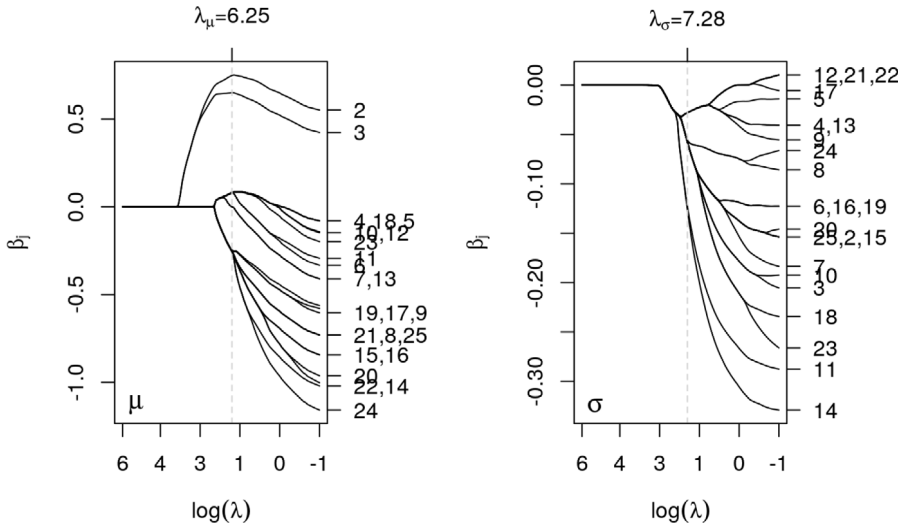

Fig. 8. Nominal fused coefficient paths for the district effect for parameters μ (left) and σ (right); vertical dashed lines: optimal tuning parameters.

Table 3
Levels of nominal factor event type (`event`).

Event type	Description
`dpa`	Damages to physical assets (DPA)
`ifraud`	Internal frauds (IFRAUD)
`efraud`	External frauds, related to payments and others (EFRAUD)
`epws`	Employment practices and workplace safety (EPWS)
`cpbp`	Clients, products and business practices (CPBP)
`bdfs`	Business disruptions and system failures (BDFS)
`edpm`	Execution, delivery and process management (EDPM)

started to investigate the effect of changing economic conditions on the distribution of these losses (see Cope et al., 2012 and Chernobai et al., 2011). In particular, Hambuckers et al. (2018) used a generalized Pareto regression model similar to the one considered in Section 5, with up to 292 explanatory variables. Regressors consisted of a nominal categorical variable with seven levels (called *event types*, referring to the physical process of the losses) and 20 lagged economic indicators related to the macroeconomic, financial and internal contexts of UniCredit (see Tables 3 and 4 for additional details). The model was estimated using a traditional LASSO estimator and a narrow set of variables was identified as relevant predictors.

However, this approach suffers from two drawbacks: on the one hand, Hambuckers et al. (2018) only used L_1-penalties, neglecting potential fusion effects among event types (see, e.g., Tables 10 and 11 of their article, where several regression coefficients are quite similar). On the other hand, they treated the various economic factors as continuous. Practically speaking, this assumption (however correct) implies that any change in one of the explanatory variables is associated with a change in ξ (shape parameter) or/and σ (scale parameter). From the point of view of a risk manager, such changes might lead to frequent updates of the requested capital. This additional variability is particularly inconvenient since it creates additional liquidity risks (see, e.g., the discussion in Distinguin et al., 2013).

In light of these considerations, we reconsider the data analysis performed in Hambuckers et al. (2018). To overcome the variability issue, each economic factor is categorized into ordered categories, defined *ex ante* by a range of values. We use both basic economic reasoning (e.g., implication of the sign of the covariate) and descriptive statistics like quantiles to define the categories. Table 4 provides the detailed categorization for each variable. This framework implies that the distribution parameters stay constant when the covariates' values stay inside a given interval, which in turn lowers the variability of the requested capital. Notice that the categorization might influence the results since we have some uncertainty surrounding that stage. In particular, if a category is too wide and the true effect would vary inside that category, the model will be misspecified. However, using a relatively large number of categories in combination with a fused LASSO approach would limit this issue. Following Hambuckers et al. (2018), *event type* is kept as a nominal predictor, however subject to regularization. Our dependent variable is the *excess loss* amount (in Euro), i.e. the difference between the loss and the threshold selected in the EVT procedure (see Hambuckers et al. (2018), Section 2.2 for details). For anonymity reasons, losses have been scaled by an unknown factor, preventing us from any reasoning on the level itself. Then, we fit a generalized Pareto GAMLSS, and use for both ξ and σ the fused LASSO penalties described previously to control for the number of parameters. As for the first application, the optimal tuning parameters are chosen over a

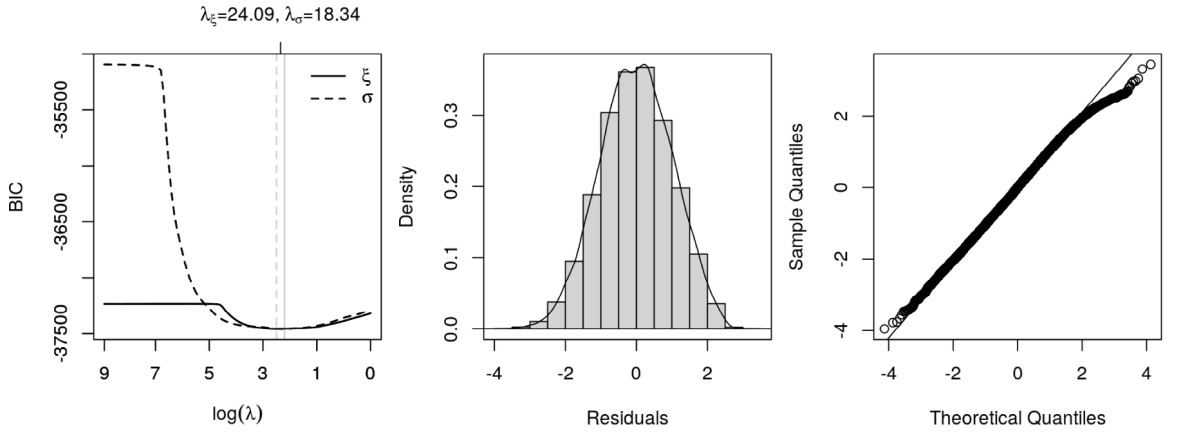

Fig. 9. UniCredit model, the left panel shows marginal BIC curves for parameters ξ and σ, holding the other tuning parameter fixed at the respective minimum of the BIC. The middle panel shows a histogram together with a kernel density estimate of the resulting randomized quantile residuals, whereas the right panel shows the corresponding quantile–quantile plot.

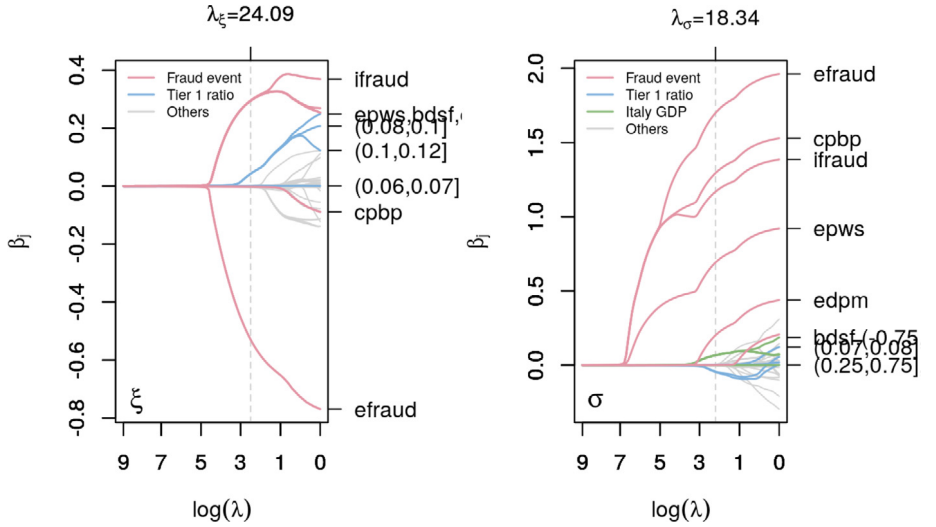

Fig. 10. UniCredit model, coefficient paths using ordinal fused LASSO.

two-dimensional grid by BIC. The left panel of Fig. 9 displays marginal BIC curves. Similar to the first application, the middle and right panels indicate a relatively good model fit according to the resulting randomized quantile residuals. Only very high losses above the 98% quantile seem to be less appropriately modeled and could be further investigated. We see that clear values for λ are chosen. Fig. 10 shows coefficients' paths for both distribution parameters. The dotted lines indicate the level of the selected penalization parameters, and of the different regression coefficients.

Our results are the following: for ξ (Fig. 10, left panel), the variable *event type* is fused in a smaller number of categories: EPWS, BDSF and EDPM form a single category, EFRAUD and IFRAUD stand alone, whereas CPBP and DPA have their associated coefficients set to zero. Regarding economic covariates, only the Tier-I capital ratio (TCR) is selected. The three upper categories (.07, .08], (.08, .1] and (.1, .12] have been fused, whereas the two lower categories have their regression coefficients set to zero. For σ (Fig. 10, right panel) we do not observe any fusion of categories regarding the *event types*. Only the BDSF event type has its coefficient set to zero. Regarding the economic factors, we observe an effect of the following variables: Italian unemployment rate, Italian GDP growth rate (GDP IT), monetary aggregate M1, S&P 500 log-returns, VIX and TCR. With the exception of the GDP IT and the TCR, all other variables exhibit extremely small regression coefficients, suggesting that these variables should be completely excluded from the final model. For GDP IT, the four upper categories (ranging from $-.75\%$ to 1.25%) have been fused. For TCR, only the two upper categories have non-zero coefficients but have not been fused.

We draw several economic interpretations from these results. First, the signs of the regression coefficients indicate that an increase in the Italian GDP growth rate above $-.75\%$ is associated with a relative increase in σ. It suggests that in relatively good economic times, the likelihood of large losses increases. It can be explained by the fact that, in a booming

Table 4
Summary of the explanatory variables in the UniCredit analysis. The third column shows for all (categorized) variables the interval limits (except for variable `event`, which is a true factor variable) together with the frequencies of the so created categories, as indicated by the height of the black bars.

Type	Variable	Categories	Description
Firm-specific	`event`	dpa ifraud efraud epws cpbp bdsf edpm	Event type
	`leveragelag`	12 15 18 21 24 32	Leverage ratio (LR)
	`tier1ratiolag`	0.05 0.06 0.07 0.08 0.1 0.12	Tier-I capital ratio (TCR)
	`prflag`	0.23 0.27 0.31 0.35 0.39 0.43	% revenue coming from fees (PRF)
	`depositgrowthlag`	−0.35 −0.2 −0.05 0.1 0.25 0.4 0.55	Deposit growth rate (DGR)
	`logreturnslag`	−0.315 −0.225 −0.135 −0.045 0.045 0.135 0.225	UniCredit stock returns (SR)
Macro-economic	`unempitlag`	5.1 6.1 7.1 8.1 10.1 12.1 14.1	Italian unemployment rate (UR IT)
	`unempeulag`	6.3 7.3 8.3 9.3 10.3 11.3 12.3	EU unemployment rate (UR EU)
	`gdpitlag`	−Inf −1.25 −0.75 −0.25 0.25 0.75 1.25	Italian GDP growth rate (GDP IT)
	`gdpeulag`	−Inf −1.25 −0.5 0 0.5 1 1.5	EU GDP growth rate (GDP EU)
	`rpi.eu`	−6 −4 −2 0 2 4 6 8	EU housing price growth rate (HPI)
	`m1`	0 3 6 9 12 15	Monetary aggregate M1 growth rate (M1)
	`lfc.italy`	5 7 9 11 13	Consumer loans rate <1 year in Italy (LOR IT)
	`lfc.eu`	5 6 7 8 9	Consumer loans rate <1 year in EU (LOR EU)
Financial	`splogreturns`	−0.315 −0.225 −0.135 −0.045 0.045 0.135 0.225	S&P 500 returns
	`trlogreturns`	−0.3 −0.18 −0.06 0.06 0.18 0.3	TR EU Stock Index returns (TRSI)
	`miblogreturns`	−0.275 −0.165 −0.055 0.055 0.165 0.275	FTSE MIB index returns (MIB)
	`vixlag`	8 14 20 26 32 38 44 50	VIX
	`vftselag`	8 14 20 26 32 38 44	VFTSE
	`itinterbank.rate`	0 1.5 3 4.5 6	3-month Italian interbank rate
	`italtbr`	2 3.5 5 6.5 8	10-year Italian government bond yield

economy, the sizes of the transactions increase, letting mechanically the potential amount of money to be lost increase as well. The same effect is observed for fines and compensation claims in lawsuits, whereas better economic conditions may also create more incentives to commit frauds (Povel et al., 2007), increasing the likelihood of large losses related to fraud events. Similar findings were obtained by Hambuckers et al. (2018) and Cope et al. (2012). However, here, our results suggest that a small recession or a positive growth rate does not lead to significant differences in terms of risk. Second, regarding the TCR, we find contradictory effects on ξ and σ: an increase up to 7% and above leads to an increase in ξ, whereas an increase of the TCR above 8% leads to a decrease in σ. One explanation would be the following: it has been shown that banks suffering from a huge degree of uncertainty regarding future losses tend to self-insure by holding more capital (Valencia, 2016). Hence, an increase in TCR seems to be indicative of a higher probability of large losses, which is consistent with the positive regression coefficient observed for ξ and findings in Hambuckers et al. (2018). On the other hand, a high TCR can be indicative of a bank with strong internal controls, as suggested in Chernobai et al. (2011) and Cope et al. (2012). Improved management practices would therefore explain a decrease in the scale of large losses, reflected in the negative regression coefficients for σ. Nevertheless, the present analysis suggests that, in term of tail risk, the former effect dominates: an increase in TCR above 7% is synonym of a heavier tail of the density.

Overall, our procedure selects a sparse model and enforces the fusion of several categories (adjacent ones for ordered predictors). We start from an unrestricted model with 232 regression coefficients to obtain a final model with only 18 parameters. The selected set of predictors, as well as the signs and magnitudes of the coefficients, provide a model theoretically coherent and easy to interpret. Lastly, this model limits strongly the variability of associated risk measures.

7. Conclusion

We presented a regularization approach for high-dimensional data set-ups for GAMLSS. The framework is based on LASSO-type penalties for metric covariates, and both group and fused LASSO for categorical predictors. Estimation is performed using a backfitting algorithm with different types of shrinkage parameter selection. Moreover, we showed that the fused LASSO can even be implemented using a gradient boosting algorithm.

We investigated the performance of the novel fused LASSO-type penalties for GAMLSS compared to unpenalized and commonly used boosting methods in an intensive simulation study. The performance of the LASSO-type penalties was shown to be superior over the other methods, even if the true fused categories are supplied as covariates in the model. In particular, it turned out that the fused boosted LASSO models have a very good performance. However, the estimation

of the effective degrees of freedom within the boosting algorithm is to some extent critical. We used the active covariate set as proposed in Zou et al. (2007), which has considerable computational advantages, but it turned out that we partly overestimated the true number of parameters. For this reason, the performance was inferior for some settings, when selecting the stopping iteration based on BIC. Therefore, a more elaborate estimation of the effective degrees of freedom in gradient boosting will be a topic of future research.

The proposed methods were also applied to two different real data sets, namely to Munich rental guide data from the year 2007 and to data on extreme operational losses of the Italian bank UniCredit. In the first data set, a selection of nine factors describing certain characteristics of apartments in Munich was related to the monthly net rent per square meter. The fusion behavior of the fused LASSO was illustrated by the help of both nominal and categorial factor covariates. In particular, it was shown that the method detects the decisive number of different categories per predictor and yields a sparse model, which facilitates interpretation of the estimated regression effects. In the second data set, the severity distribution of operational losses was related to 21 economic variables, mapped into 232 ordered categorical predictors. With the help of the proposed approach, we excluded numerous non-informative predictors from our final model. In addition, thanks to the fused LASSO penalty, we identified the levels of the covariates that have a similar effect on the distribution of the losses. Consequently, we were able to obtain a final model sparse and theoretically sound, producing stable financial risk indicators.

In future research the fusion penalties presented here could also be extended for regularization and model selection purposes when categorical effect modifiers are considered. More precisely, similar to Gertheiss and Tutz (2012), the fusion penalties could be complemented by an additional LASSO term.

Appendix A. Supplementary data

Supplementary material, including the R-scripts and the UniCredit loss data, can be found online at https://doi.org/10.1016/j.csda.2019.06.005.

References

Basel Committee on Banking Supervision (BCBS), 2004. Basel II: International convergence of capital measurement and capital standards. A revised framework. Technical Report, Bank of International Settlements, Basel, Switzerland.

Bondell, H.D., Reich, B.J., 2009. Simultaneous factor selection and collapsing levels in ANOVA. Biometrics 65 (1), 169–177.

Chapelle, A., Crama, Y., Hübner, G., Peters, J.P., 2008. Practical methods for measuring and managing operational risk in the financial sector: A clinical study. J. Bank. Financ. 32 (6), 1049–1061.

Chavez-Demoulin, V., Embrechts, P., Hofert, M., 2016. An extreme value approach for modeling operational risk losses depending on covariates. J. Risk Insur. 83 (3), 735–776.

Chernobai, A., Jorion, P., Yu, F., 2011. The derminants of operational risk in U.S. financial institutions. J. Financ. Quant. Anal. 46 (8), 1683–1725.

Chiquet, J., Gutierrez, P., Rigaill, G., 2017. Fast tree inference with weighted fusion penalties. J. Comput. Graph. Statist. 26 (1), 205–216.

Cope, E., Piche, M., Walter, J., 2012. Macroenvironmental determinants of operational loss severity. J. Bank. Financ. 36 (5), 1362–1380.

Distinguin, I., Roulet, C., Tarazi, A., 2013. Bank regulatory capital and liquidity: Evidence from US and European Publicly traded banks. J. Bank. Financ. 37 (9), 3295–3317.

Dunn, P.K., Smyth, G.K., 1996. Randomized quantile residuals. J. Comput. Graph. Statist. 5, 236–245. http://dx.doi.org/10.2307/1390802.

Embrechts, P., Klupperlberg, C., Mikosch, T., 1997. Modelling Extremal Events for Insurance and Finance. Springer - Verlag, Berlin.

Gertheiss, J., Tutz, G., 2010. Sparse modeling of categorial explanatory variables. Ann. Appl. Stat. 4 (4), 2150–2180.

Gertheiss, J., Tutz, G., 2012. Regularization and model selection with categorial effect modifiers. Statist. Sinica 957–982.

Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (2), 243–268.

Hambuckers, J., Groll, A., Kneib, T., 2018. Understanding the economic determinants of the severity of operational losses: A regularized generalized pareto regression approach. J. Appl. Econometrics 33 (6), 898–2180.

Hofner, B., Mayr, A., Schmid, M., 2016. **gamboostLSS**: An r package for model building and variable selection in the GAMLSS framework. J. Stat. Softw. 74 (1), 1–31.

Kneib, T., Konrath, S., Fahrmeir, L., 2011. High dimensional structured additive regression models: Bayesian regularization, smoothing and predictive performance. J. R. Stat. Soc. Ser. C. Appl. Stat. 60 (1), 51–70.

Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K., Kneib, T., 2014. Multilevel structured additive regression. Stat. Comput. 24 (2), 223–238.

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. J. R. Stat. Soc. Ser. C. Appl. Stat. 61 (3), 403–427.

Meier, L., Van de Geer, S., Bühlmann, P., 2008. The group lasso for logistic regression. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (1), 53–71.

Oelker, M.R., Tutz, G., 2017. A uniform framework for the combination of penalties in generalized structured models. Adv. Data Anal. Classif. 11 (1), 97–120.

Povel, P., Singh, R., Winton, A., 2007. Booms, busts, and fraud. Rev. Financ. Stud. 20 (4), 1219–1254.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, ISBN 3-900051-07-0.

Rigby, R.A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape. J. R. Stat. Soc. C (Appl. Stat.) 54 (3), 507–554.

Sakamoto, Y., Ishiguro, M., G., K. (Eds.), 1986. Akaike Information Criterion Statistics. D. Reidel Publishing Company.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464. http://dx.doi.org/10.1214/aos/1176344136.

Stasinopoulos, D.M., Rigby, R.A., 2007. Generalized additive models for location scale and shape (gamlss) in r. J. Stat. Softw. 23 (7), 1–46.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., Hofner, B., 2018. Gradient boosting for distributional regression: Faster tuning and improved variable selection via noncyclical updates. Stat. Comput. 28 (3), 673–687.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

Umlauf, N., Klein, N., Zeileis, A., 2018a. BAMLSS: Bayesian additive models for location, scale and shape (and beyond). J. Comput. Graph. Statist. 27 (3), 612–627.

Umlauf, N., Klein, N., Zeileis, A., Köhler, M., Simon, T., 2018b. bamlss: Bayesian additive models for location scale and shape (and beyond). http://CRAN.R-project.org/package=bamlss, R package version 1.0-1.
Valencia, F., 2016. Bank capital and uncertainty. J. Bank. Financ. 69 (S1), S1–S9.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (1), 49–67.
Zou, H., Hastie, T., Tibshirani, R., 2007. On the 'degrees of freedom' of the lasso. Ann. Statist. 35 (5), 2173–2192.
Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models. Ann. Statist. 36 (4), 1509–1533.