Georg-August Universität Göttingen

bamlss.vis: An R Package to Interactively Analyze and Visualize Bayesian Additive Models for Location, Scale and Shape (bamlss) Using the Shiny Framework

20 week Master thesis as part of the
Master of Science (M.Sc.) course “Applied Statistics”
at the University of Göttingen

Author:
Stanislaus Stadlmann,
Student ID: 21144637

Supervisors
Prof. Dr. Thomas Kneib
Dr. Nadja Klein

Submitted on November 14, 2017
by Stanislaus Stadlmann,
born in Vienna, Austria

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Contents

List of Figures

List of Tables

1 Introduction

Since the commercialization of the personal computer and the smartphone about two decades later the overwhelming majority of modern life in developing nations has greatly been revolutionized. To name a few advancements, the period stretching from the late 20th century until today has seen changes in the way modern human beings communicate, listen to music, work and are entertained. The common denominator of these changes is the switch from analogue to digital processes, which saw the creation of entire industries, such as Digital Image Processing. The digital revolution also started a significant growth in the number of data collection possibilities and -techniques, with the newest breakthrough, the Internet of Things (IoT), being right around the corner (O'Connor, 2016).

The exponential increase in available datapoints, paired with dramatic improvements in computing power, gave rise to numerous advancements in statistical sciences. Many computation-heavy models were able to be applied on a broader basis and new methods, such as Neural Nets or Generalized Additive Models could finally be realistically used (The Economist, 2015). With the increase in number of new methods and improvements in data availability, the recent past also saw a significant rise in employed statisticians. In the United States alone, the number of jobs classified as statisticians has increased by more than 120% in the years from 1997 to 2016 (Bureau of Labor Statistics, 2016).

One of the new fields that has emerged is distributional regression, where not only the mean, but each parameter of a response distribution can be modeled using a set of predictors (Klein et al., 2015). Notable frameworks called Generalized Additive Models for Location, Scale and Shape (gamlss) and Bayesian Additive Models for Location, Scale and Shape (bamlss) were invented by Rigby and Stasinopoulos (2001) in the form of a frequentist perspective and Umlauf et al. (2017) with a Bayesian approach, respectively.

Because methods have become increasingly more complex and capable over the years, it is important to make them accessible and understandable to the growing number of statistical users. In the case of distributional regression models, the interpretation of covariate effects on response moments and the expected conditional response distribution is harder than with traditional methods such as Ordinary Least Squares or Generalized Linear Models, since the moments of a distribution do not directly equate the modeled parameters, but are rather a combination of them with a varying degree of complexity.

This thesis will introduce a framework for the visualisation of distributional regression models fitted using the **bamlss** R package (Umlauf et al., 2017) as well as display an implementation as an R extension titled **bamlss.vis**. The goal of this framework is the ability to:

- See and compare the expected distribution for chosen sets of covariates and
- View the direct relationship between moments of the response distribution and a chosen explanatory variable, given a set of covariates.

Additionally, the user can obtain the code which created the graphs to potentially reproduce them later. The implementation will be done using the statistical software R (R Core Team, 2017) in the form of a Shiny application (Chang et al., 2017).

2 Motivating Bayesian Additive Models for Location, Scale and Shape

Bayesian Additive Models for Location, Scale and Shape (bamlss) are a form of Bayesian regression models in which every parameter of a parametric distribution with K parameters is related to a set of additive predictors. The distribution does not have to follow the exponential family, which extends the distributions available for modeling beyond the ones used in Generalized Linear Models (GLM). In similar fashion to Generalized Additive Models (GAM, Hastie and Tibshirani, 1990), the additive predictors can assume different shapes, including non-linear, fixed, random and spatial effects (Umlauf et al., 2017).

To give a sufficient depiction of this model class, this section will start with explaining Additive Models and then gradually generalize the broader frameworks to finally arrive at bamlss. Furthermore, a brief overview of the different estimation techniques for the covered model frameworks will be given.

2.1 Additive Models

Bamlss can be seen as a generalization of Structured Additive Regression, which are in turn a generalization of Additive Models. Additive Models, first proposed by Friedman and Stuetzle (1981) represent a model type in which a dependent variable y is related to a set of non-parametric predictors in an additive way. Assuming conditional independence of $y_1, \ldots, y_n$ given the explanatory variables $\mathbf{z}_1, \ldots, \mathbf{z}_K$, we obtain the following model equation:

$$y_i = f_1(z_{i1}) + f_2(z_{i2}) + \ldots + f_k(z_{ik}) + \epsilon_i \tag{2.1}$$

where $f_j(\cdot)$ depict unspecified non-parametric functions of covariate z_j, which can include smoothing splines or local regression approaches. This makes additive models more flexible compared to standard linear regression, while still being more interpretable than non-additive models (Buja et al., 1989).

Fahrmeir et al. (2013) suggest that an Additive Model can also include parametric components. Given covariates $\mathbf{x}_1, \ldots, \mathbf{x}_Q$, we can extend (2.1) to a semiparametric regression model with the following specification:

$$y_i = \sum_{j=1}^{K} f_j(z_{ij}) + \underbrace{\sum_{l=1}^{Q} \beta_l x_{il}}_{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_Q x_{iQ}} + \epsilon_i \tag{2.2}$$

Eq. (2.2) combines non-parametric and parametric components. Because the model would otherwise not be identified, functions $f_j(\cdot)$ now have to be centered around zero, such that

$$\sum_{i=1}^{n} f_1(x_{i1}) = \ldots = \sum_{i=1}^{n} f_K(x_{iK}) = 0$$

holds. The functions $f_j(\cdot)$ are approximated using basis functions in the following scheme:

$$f_j(z_j) = \sum_{m=1}^{d_j} \mathbf{B}_m(z_j)\gamma_{jm}$$

This allows to write the Additive Model in a matrix form, indifferent of the chosen basis:

$$\mathbf{y} = \sum_{j=1}^{K} \mathbf{Z}_j \boldsymbol{\gamma}_j + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.3}$$

Here, the design matrices $\mathbf{Z}_1, \ldots, \mathbf{Z}_K$ represent the basis functions assessed at different covariates. $\mathbf{X}$ is constructed in equivalence to the standard linear regression model. Assumptions about the error term of a semiparametric Additive

Model are also similar to the classic linear model, where ϵ_i are identically and independently (i.i.d) normally distributed with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. These properties are then also valid for the response variable, so that $y_i \overset{i.i.d.}{\sim} N(\hat{y}, \sigma_y^2)$ (Fahrmeir et al., 2013, chap. 9.1).

2.2 Structured Additive Regression Models

The nonparametric components in additive models open the possibility for more flexible relationships between the dependent variable and single explanatory variables, which standard linear regression methods might not capture correctly. However, sometimes the area of model application requires even more flexibility, e.g. by including spatial covariates, fixed/random effects or interaction terms. These specific types of effects extend the Additive Model to a Structured Additive Regression Model (Fahrmeir et al., 2003, STAR). This chapter will briefly describe its different components.

2.2.1 Spatial Effects

Similarly to Section 2.1, observations $(y_i, \mathbf{z}_i, \mathbf{x}_i)$ are given, where $\mathbf{z}_i$ and $\mathbf{x}_i$ represent vectors of covariate values for the ith observation. Additionally, a geographic location index s is known with observations s_i, which can be either discrete (e.g. region or country) as well as continuous (e.g. longitude/latitude). Extending the semiparametric Additive Model as specified in (2.2), a geospatial effect is now added:

$$\begin{aligned} y_i &= \sum_{j=1}^{K} f_j(z_{ij}) + \sum_{l=1}^{Q} \beta_l x_{il} + f_{geo}(s_i) + \epsilon_i \\ &= \kappa^{add} + f_{geo}(s_i) + \epsilon_i \end{aligned} \tag{2.4}$$

κ^{add} includes the non-spatial effects from (2.2). The spatial effect, $f_{geo}(\cdot)$, is often viewed as a proxy for unknown covariates, such as altitude or climate data. If the geographic location index s is tracked using discrete values, $f_{geo}(\cdot)$ could represent a Markov random field. For continuous values, smoothing techniques such as Kriging (Matheron, 1963) or a multivariate tensor product spline are available. In both the discrete and the continuous case, the vector of geoadditive components $\boldsymbol{f}_{geo}$ can be written as

$$\boldsymbol{f}_{geo} = \mathbf{Z}_{geo}\boldsymbol{\gamma}_{geo}$$

so that it can be incorporated into the geoadditive model in matrix notation in the following way

$$\mathbf{y} = \sum_{j=1}^{K} \mathbf{Z}_j\boldsymbol{\gamma}_j + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{geo}\boldsymbol{\gamma}_{geo} + \boldsymbol{\epsilon} \tag{2.5}$$

which bears similarities to the basis function approach in (2.3) (Fahrmeir et al., 2013, chap. 9.2).

2.2.2 Interaction Terms

The regression equation (2.2) of Additive Models included main nonparametric and parametric effects, but no interactions between covariates. When incorporating interaction effects, one has to differentiate between an interaction between a continuous and a categorical variable, as well as one where two continuous variables share a common effect (Fahrmeir et al., 2013, chap. 9.3).

To illustrate the first case, it is assumed that z_1 and x_1 are continuous and binary ($x_i \in (0,1)$) covariates, respectively. Then, the interaction term $f_{z_1|x_1}(z_1) \cdot x_1$ can be included in the Additive Model from (2.2) in the following way:

$$y_i = \sum_{j=1}^{K} f_j(z_{ij}) + \sum_{l=1}^{Q} \beta_l x_{il} + \underbrace{f_{z_1|x_1}(z_{i1})x_{i1}}_{\substack{0 \qquad\qquad \text{if } x_{i1}=0 \\ f_{z_1|x_1}(z_{i1}) \text{ if } x_{i1}=1}} + \epsilon_i$$

If $x_1 = 0$, the non-linear effects of z_1 are now

$$\begin{aligned} f_1(z_1) &\quad \text{if } x_1 = 0 \\ f_1(z_1) + f_{z_1|x_1}(z_1) + \beta_1 &\quad \text{if } x_1 = 1 \end{aligned}$$

This framework can also incorporate spatially covarying terms, where the interaction term $f_{geo|x_1}(s)$ represents an interaction between the location variable s and a categorical variable x_1 (Fahrmeir et al., 2013).

Using a Basis function approach, the vector of interaction effects

$$\boldsymbol{f}_{int} = (f_{z_1|x_1}(z_{11})x_{11}, \ldots, f_{z_1|x_1}(z_{n1})x_{n1})$$

can also be described in matrix notation to extend (2.3) in the following way:

$$\mathbf{y} = \sum_{j=1}^{K} \mathbf{Z}_j \boldsymbol{\gamma}_j + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{int} \boldsymbol{\gamma}_{int} + \boldsymbol{\epsilon}$$

Here, the design matrix $\boldsymbol{Z}_{int}$ represents the Basis function values multiplied with x_1 observations (Fahrmeir et al., 2013, chap. 9.3).

The possibility of interactions between two continuous covariates is also given. In this case, the interaction between z_1 and z_2 is modeled using a two-dimensional nonparametric function $f_{z_1|z_2}(z_1, z_2)$. Common two-dimensional functions include bi-variate smooth splines and Kriging techniques. When only the two-dimensional functions without main effects ($f_1(z_1)$, $f_2(z_2)$) should be included, the model equation assumes the following form:

$$y_i = f_{z_1|z_2}(z_{i1}, z_{i2}) + f_3(z_{i3}) + \ldots + f_K(z_{iK}) + \sum_{l=1}^{Q} \beta_l x_{il} + \epsilon_i \tag{2.6}$$

For reasons of identifiability, $f_{z_1|z_2}(z_1, z_2)$ also needs to be centered around zero. Fahrmeir et al. (2013, chap. 9.3.2) warn that for estimation of models with two-dimensional surfaces a high sample size with combinations of z_1 and z_2 is required. In cases where this requirement is not fulfilled, a simple main effects model as in (2.2) is preferred.

It is also possible to model the interaction effect of z_1 and z_2 using the two-dimensional surface $f_{z_1|z_2}(z_1, z_2)$ while still including the main effects. In this scenario, the model is specified as follows:

$$y_i = f_{z_1|z_2}(z_{i1}, z_{i2}) + f_1(z_{i1}) + f_2(z_{i2}) + \sum_{j=3}^{K} f_j(z_{ij}) + \sum_{l=1}^{Q} \beta_l x_{il} + \epsilon_i \tag{2.7}$$

The identifiability problem in this model is more complex than before. To solve it, Fahrmeir et al. (2013, chap. 9.3) state that not only all included functions have to be centered around zero, but also "all slices of the interaction $f_{z_1|z_2}(z_1, z_2)$, i.e. all one-dimensional smooths with fixed value of z_1 or z_2". Using the basis function approach, the matrix representation of the model can be obtained:

$$\mathbf{y} = \sum_{j=1}^{K} \mathbf{Z}_j \boldsymbol{\gamma}_j + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{z_1|z_2} \boldsymbol{\gamma}_{z_1|z_2} + \boldsymbol{\epsilon} \tag{2.8}$$

with interaction term design matrix $\mathbf{Z}_{z_1|z_2}$ (Fahrmeir et al., 2013, chap. 9.3).

2.2.3 Random Effects

When dealing with repeated measures or other longitudinal datasets it is often necessary to model cluster-specific similarities using Random Effects (Laird and Ware, 1982). Additive Models can also be extended with Random Effects to arrive at so called Additive Mixed Models. Assuming a longitudinal data structure with subjects $j = 1, ..., n_i$ in clusters $i = 1, ..., m$ and covariates $\mathbf{x}_k$, a parametric random coefficient model possesses the following structure:

$$y_{ij} = (\beta_0 + \nu_{0i}) + (\beta_1 + \nu_{1i})x_{ij1} + \ldots + (\beta_Q + \nu_{Qi})x_{ijQ} + \epsilon_i$$

The "random" coefficients ν_{0i} (intercept) and $\nu_{1i}, \ldots, \nu_{Qi}$ (slopes) represent the cluster-specific deviations from the main effects. To obtain Additive Mixed Models, the main effects are then replaced with nonparametric functions:

$$y_{ij} = f_1(x_{ij1}) + \ldots + f_Q(x_{ijQ}) + \nu_{0i} + \nu_{1i}x_{ij1} + \ldots + \nu_{Qi}x_{ijQ} + \epsilon_i \tag{2.9}$$

Like non-parametric main effects, Random Effects also have a matrix notation. In the case where every main effect is also modeled with cluster-specific effects, the matrix form of Additive Mixed Models is as follows:

$$\mathbf{y} = \sum_{j=1}^{K} \mathbf{Z}_j \boldsymbol{\gamma}_j + \mathbf{R}_0 \boldsymbol{\nu}_0 + \sum_{j=1}^{K} \mathbf{R}_j \boldsymbol{\nu}_j + \boldsymbol{\epsilon}$$

Here, $\boldsymbol{\nu}_0 = (\nu_{01}, \ldots, \nu_{0m})'$ and $\boldsymbol{\nu}_j = (\nu_{j1}, \ldots, \nu_{jm})'$ represent the Random Effects coefficients. A more in-depth look at the structure of the design matrices is given by Fahrmeir et al. (2013, chap. 9.4, p. 550)

2.3 Generalized Structured Additive Regression Models

Structured Additive Regression (STAR) models extend simple Additive Models with special model terms briefly introduced in the previous sections. These effects include:

- Nonlinear effects of z_1

- Spatial effects of location index s
- Interactions between continuous covariate z_1 and a categorical variable x_1
- Nonlinear interactions between two continuous covariates $z1$, z_2
- Random Effects with intercept ν_0 and slope ν_j deviations from main effects

All of the aforementioned model terms can be included in a STAR interchangeably, including simple linear predictors $\mathbf{x}'\boldsymbol{\beta}$ (Fahrmeir et al., 2013, chap 9.5).

STAR models provide very flexible ways of modeling the influence of explanatory variables on a given response variable y_i. Note that while the components can be nonparametric, the direct modeling of y_i assumes that the response variable follows a Gaussian distribution. However, when dealing with e.g. binary or categorical responses, this assumption is violated. Then, a type of model specification is needed that directly upholds the dependent variables' support (Olsson, 2002, chap. 2). To solve this challenge, STAR models are merged with Generalized Linear Models to Generalized STAR models.

Generalized Linear Models (GLM), first coined by Nelder and Wedderburn (1972), introduce a framework where the expectation of response y is related to a linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}$ via a link function $\eta = g(E(y)) = g(\mu)$ or a response function $h = g^{-1}$ to arrive at the following model specification:

$$\begin{aligned} \mu_i &= h(\mathbf{x}_i'\boldsymbol{\beta}) \quad \text{or} \\ g(\mu_i) &= \mathbf{x}_i'\boldsymbol{\beta} \end{aligned} \tag{2.10}$$

When modeling a binomially distributed response the probability parameter π, which has a support of $\pi \in [0, 1]$, is related to predictors $\mathbf{x}'\boldsymbol{\beta}$. Using a logit link function, we obtain a Logistic Regression Model:

$$\begin{aligned} \eta_i &= \mathbf{x}_i'\boldsymbol{\beta} \\ E(y_i) = \pi_i &= \frac{exp(\eta_i)}{1 + exp(\eta_i)} \end{aligned}$$

Here, the response function ensures the correct support of π (Fahrmeir et al., 2013, chap. 5). Using a link function, the expectation of y can be the first moment of many different continuous or discrete distributions, which includes the Poisson, Binomial and Gamma distribution. However, all possible distributions need to be part of the exponential family (Rigby and Stasinopoulos, 2005).

Note in (2.10) that the effects of covariates $\mathbf{x}_1, \ldots, \mathbf{x}_K$ are modeled parametrically. Generalized Additive Models (GAM), as suggested by Hastie and Tibshirani (1990), extend the class of Generalized Linear Models to allow for non-parametric effects. In particular, the linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}$ is interchanged by smooth non-parametric functions $f_j(x_j)$. Given response variable y and covariates $\mathbf{z}_1, \ldots, \mathbf{z}_K$, the following model specification is obtained:

$$\begin{aligned} \eta_i &= \sum_{j=1}^{K} f_j(x_{ij}) \\ \mu_i &= E(y_i) = h(\eta_i) \end{aligned} \tag{2.11}$$

Now, many different response distributions as well as flexible effects for explanatory variables are supported to create a highly flexible model framework. In (2.11), only non-parametric effects are linked to η_i. However, given response y and covariates $(\mathbf{x}_i, \mathbf{z}_i)$, all specific effects of STAR models (spatial effects $f_{geo}(\cdot)$, interactions $f_{int}(\cdot)$, etc.) as well as parametric coefficients can be combined to form a Generalized Structured Additive Regression Model (Generalized STAR):

$$\begin{aligned} \eta_i &= f_1(z_{i1}) + \ldots + f_K(z_{iK}) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_Q x_{iQ} \\ \mu_i &= h(\eta_i) \end{aligned} \tag{2.12}$$

In semiparametric Generalized STAR models, $f_j(\cdot)$ can have any of the structural forms described in Chapter 2.2. Modeled response variables also have to follow an exponential family distribution (Fahrmeir et al., 2013, chap. 9.5).

2.4 Structured Additive Distributional Regression

Generalized STAR models provide a framework to flexibly estimate the expected value of a previously specified distributional parameter. However, in many cases not only the first moment, but also higher-order moments are of special interest. In modeling income, for example, not only the expected income but also the shape of the overall distribution is important. A common measure for income inequality is the Gini coefficient, which can be calculated using the cumulative distribution function (cdf) (Lerman and Yitzhaki, 1984).

2.4.1 GAMLSS

First modeling approaches which go beyond the mean of a distribution were suggested by Nelder and Pregibon (1987) using parametric functions of explanatory covariates related to the dispersion parameter ϕ of en exponential family distribution. Building upon this approach, Generalized Additive Models for Location, Scale and Shape (gamlss) were introduced by Rigby and Stasinopoulos (2001). Gamlss combine the flexibility of being able to model multiple distributions with parametric or nonparametric explanatory effects and extend them for multiple response distribution parameters such that not only the location, but also the scale and shape of a distribution can be modeled simultaneously. Furthermore, gamlss relax the assumption of y following an exponential family distribution, which significantly increases the number of response modeling possibilities.

Assuming a dependent variable from a distribution with parameters $\theta_1, \ldots, \theta_L$ and observations $y_1, \ldots, y_n$, given covariates $\mathbf{z}_1, \ldots, \mathbf{z}_K$ and $\mathbf{x}_1, \ldots, \mathbf{x}_Q$, a gamlss can be described with the following model specification:

$$g_l(\theta_{il}) = \eta_{il} = \mathbf{x}'_{il}\boldsymbol{\beta} + \sum_{j=1}^{K_l} f_{jl}(z_{ijl}) \tag{2.13}$$

In Equation (2.13), $g_l(\cdot)$ represents a known monotonic link function, which can be different for each parameter. $\mathbf{x}'_{il}$ depicts the subset of x variables used to model parameter θ_l in observation i, while $f_{jl}(z_{ijl})$ serves as a non-parametric effect of covariate z_j on parameter θ_l, taken from a subset of the K z variables, evaluated for the ith observation. The specific subset of covariates z with non-parametric effects on parameter θ_k has a length of K_l variables (Stasinopoulos et al., 2007).

As shown above, gamlss can utilize different combinations of parametric and non-parametric effects to model each distributional parameter. Equation 2.13 displays a case in which every parameter is modeled using a non-empty subset of variables x and z. However, some parameters can also be set to a constant and not be dependent on covariates. For example, when assuming the Gaussian distribution for the dependent variable and connecting μ to parametric effects $\mathbf{x}_j$ using the identity link function ($g(\mu) = \mu$) and the variance parameter σ^2 to a constant, we arrive at a linear model specification (Stasinopoulos et al., 2007).

2.4.2 BAMLSS

As mentioned in the introduction of this thesis, not always do the modeled parameters directly equate the moments (location, scale and shape) of a distribution, but rather a combination of them. For this reason, approaches to simultaneously model the parameters of a distribution are often referred to as distributional regression, which includes gamlss. However, as seen in (2.13), gamlss in its normal form only incorporate main effect modeling. To further integrate structured additive terms, such as spatial effects, random effects and interaction terms (Brezger and Lang, 2006), distributional regression is further extended to Structured Additive Distributional Regression (Klein et al., 2015).

In 2013, Klein et al. introduced Bayesian Additive Distributional Regression, which is a model type extending gamlss to include structured additive predictors for modeling parameters of a specified distribution. It represents a fully Bayesian approach, in which coefficients are obtained by drawing samples from the approximate posterior effect distributions using using Markov Chain Monte Carlo (MCMC) simulations.

An implementation of Bayesian Additive Distributional Regression, called Bayesian Additive Models for Location, Scale and Shape (bamlss) was since created by Umlauf et al. (2017). As the authors point out, the name bears resemblance to gamlss, because of many similarities in its modeling approach. However, extensions of bamlss over gamlss are manifold. First, parallel to the proposed framework of Klein et al. (2013), MCMC simulations are utilized for estimation of coefficients. This is done in contrast to gamlss, where predictor coefficient estimates are retrieved via penalised likelihood maximisation techniques. Advantages of using MCMC simulations over likelihood-based approaches include the sample-based inference, which yields more reliable confidence intervals than the intervals of gamlss estimates based on asymptotic properties. Second, bamlss offer more flexibility of specifying covariate effects with the support of structured additive predictors, like spatial effects or two-dimensional splines. Third, bamlss also support multivariate response distributions, which enhances gamlss' univariate response framework. Furthermore, the implementation of bamlss is designed in a way that allows for the usage of external estimation algorithms and software packages like JAGS or BayesX.

The model specification of bamlss is similar to the gamlss class. The parameters

$\theta_1, \ldots, \theta_L$ of a parametric distribution $\mathbf{y}$ with observations $y_1, \ldots, y_n$ are linked to structured additive predictors using monotonic and twice-differentiable link functions $g_l(\theta_l)$ (note that the paper uses $h_l(\theta_l)$). Based on covariates $\mathbf{x}_1, \ldots, \mathbf{x}_Q$, the following model equation can be obtained:

$$g_l(\theta_l) = f_{1l}(\mathbf{x}_{1l}; \boldsymbol{\beta}_{1l}) + \ldots + f_{Q_l l}(\mathbf{x}_{Q_l l}; \boldsymbol{\beta}_{Q_l l}) \tag{2.14}$$

Here, $f_{j_l l}(\cdot)$ represent unspecified functions that can attain any structured additive predictor forms, including nonparametric effects. It is also possible to describe the effects in vector form:

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

with $\mathbf{X}_{jl}$ $(n \times m_{jl})$ specifying the design matrix for effect $f_{jl}(\cdot)$ so that they integrate themselves into the following model equation

$$g_l(\boldsymbol{\theta}_l) = \boldsymbol{\eta}_l = \boldsymbol{f}_{1k} + \ldots + \boldsymbol{f}_{J_l l} \tag{2.15}$$

where $\boldsymbol{f}_{jl}$ represents the jth effect of $\mathbf{x}_{jl}$ (subvector of $\mathbf{x}$) on parameter θ_l. Similar to Chapters 2.1 and 2.2, effects in bamlss can also be derived through a basis function approach, such that it can be written as $\boldsymbol{f}_{jl} = \mathbf{X}_{jl}\boldsymbol{\beta}_{jl}$. The structure of the design matrix depends on the types of covariates and prior assumptions about $f_{jl}(\cdot)$ (Umlauf et al., 2017). As mentioned earlier in this chapter, bamlss offer very flexible ways of specifying covariate effects. Breaking through the framework of basis function approaches, bamlss also allow covariate functions $f_{jl}(\cdot)$ which are nonlinear in its parameters $\boldsymbol{\beta}_{jl}$. An example of this is the Gompertz growth curve

$$\boldsymbol{f}_{jl} = \beta_1 \cdot \exp(-\exp(\beta_2 + \mathbf{X}_{jl}\beta_3))$$

with nonlinear parameters $\boldsymbol{\beta}_{jl}$(Umlauf et al., 2017).

2.5 Estimation

3 bamlss.vis

The previous Sections 2.1 to 2.5 gave a description of Bayesian Additive Models for Location, Scale and Shape (bamlss) and the underlying sub-models on which they are based. This section will introduce a framework to interactively visualize covariate effects and distributional predictions of fitted bamlss models and feature its implementation as an R package. Because of the visual component, the tool will be called **bamlss.vis**. A small case-study based on wages of male workers in the Mid-Atlantic region will be presented to feature most of bamlss.vis' abilities.

3.1 Motivation

As discussed in previous sections, distributional regression is concerned with modeling the parameters of a known parametric distribution. After estimation of the model, the user obtains coefficients which measure the influence of an explanatory variable on η_l, which represents the transformed parameter θ_l. However, in most cases the user is not interested in specific distributional parameters but more in the moments, which often do not directly equate the parameters but are rather a combination of them.

This problem can be well illustrated using the censored normal distribution. Assume a normally distributed variable, $y \sim N(\mu, \sigma^2)$. Then, the probability density function (pdf) of a left-censored normal distribution y^* with cut-off point $a = 0$ can be obtained by

$$f(y^* = x) = \begin{cases} f(y = x) & x > 0 \\ F(y = \frac{-\mu}{\sigma}) & x \leq 0 \end{cases}$$

where $f(y)$ and $F(y)$ are the probability density functions (pdf) and the cumulative distribution function cdf() of normally distributed variable y, respectively. It is visible in the above equation that the censored normal distribution is both discrete and continuous. While y^* shares the density with y above the cut-off point, the full remaining density in the censored normal distribution is assigned

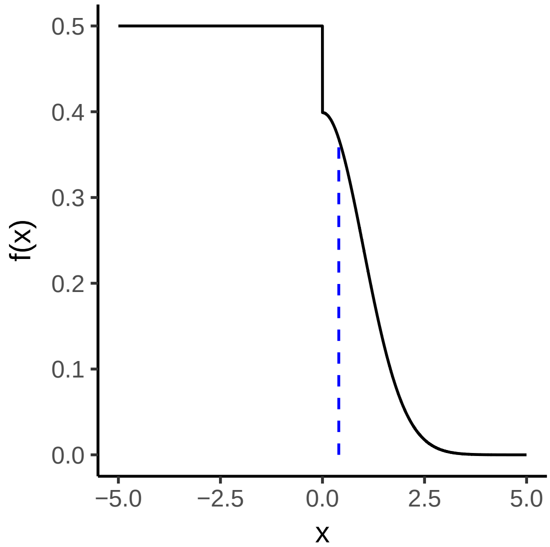

Figure 1: Probability Density Function of a left-censored normal distribution with the expected value drawn as a blue line.

to the cut-off point a. Figure 1 shows a sample left-censored normal distribution y^* created from $y \sim N(0,1)$ with $a = 0$ (Greene, 2012).

As visible in Figure 1, the moments of the standard normal distribution do not carry over to the censored normal distribution. In fact, while $E(y) = 0$, the expected value of y^* is $E(y^*) \approx 0.399$. To be exact, the censored normal distributions first two moments with cut-off $a = 0$ can be calculated as follows:

$$\begin{aligned} E(y^*) &= (1-\alpha)\cdot(\mu+\sigma\beta) \quad \text{and} \\ Var(y^*) &= \sigma^2(1-\alpha)\cdot[(1-\gamma)+(\frac{-\mu}{\sigma}-\beta)^2\cdot\alpha] \\ \\ \text{while:} \quad \alpha &= \Phi(\frac{-\mu}{\sigma}) \\ \beta &= \frac{\phi(\frac{\mu}{\sigma})}{1-\alpha} \\ \gamma &= \beta^2 - \beta\cdot(\frac{-\mu}{\sigma}) \end{aligned} \tag{3.1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function (pdf) and cumulative distribution function (cdf) of the standard normal distribution and μ and σ^2 are the parameters of y, respectively (Greene, 2012). Equation (3.1) shows that both the expected value and the variance of y^* are computed by a combination of the parameters of the original variable y, μ and σ^2, and are not equal. Thus, an explanatory variable that has a positive effect on μ has both an impact on

$E(y^*)$ and $Var(y^*)$. Therefore, coefficients for measuring covariate influences on those parameters are not directly translateable to the moments of the modeled distribution and might even have critically different estimates.

Furthermore, even in cases where the desired moments directly equate the modeled parameters (e.g. in gaussian or poisson-distributed responses), different link functions for their transformation $g_l(\theta_l)$ and possibly highly complex nonparametric effects of explanatory variables can lead to coefficient estimates that are hard to interpret. In this case, a visual comparison of predicted distributions would be helpful.

To tackle both of the aforementioned interpreting problems with fitted bamlss models, this thesis will introduce a framework with two main objectives:

- Visually compare the predicted distributions (pdf or cdf) based on interactively selected covariates
- View the changes of distribution moments over the whole range of a selected variable, based on chosen explanatory covariates.

Using **bamlss.vis**, one can then observe the influence of a covariate on the distribution by 1. its cdf or pdf and 2. its moments.

3.2 Case-Study

While automatic testing of **bamlss.vis**' main functions relies on artifical data for each supported distribution in order to prove correct behaviour, presenting the apps' abilities is best done with a dataset of real observations. This chapter will focus on fitting a bamlss based on "real" data for further use in **bamlss.vis**. The objective for a suitable dataset was that its response variable and explanatory variables are easy to understand for people without a specific scientific background. The chosen dataset, "Wage" from the ISLR R package (James et al., 2017), perfectly encompasses these requirements. "Wage", collected by the United States Census Bureau (2011), includes 3000 male individuals with records of the following variables:

- **wage**: Workers raw wage (in 1000 $)
- **age**: Age of worker
- **year**: Year that wage information was recorded

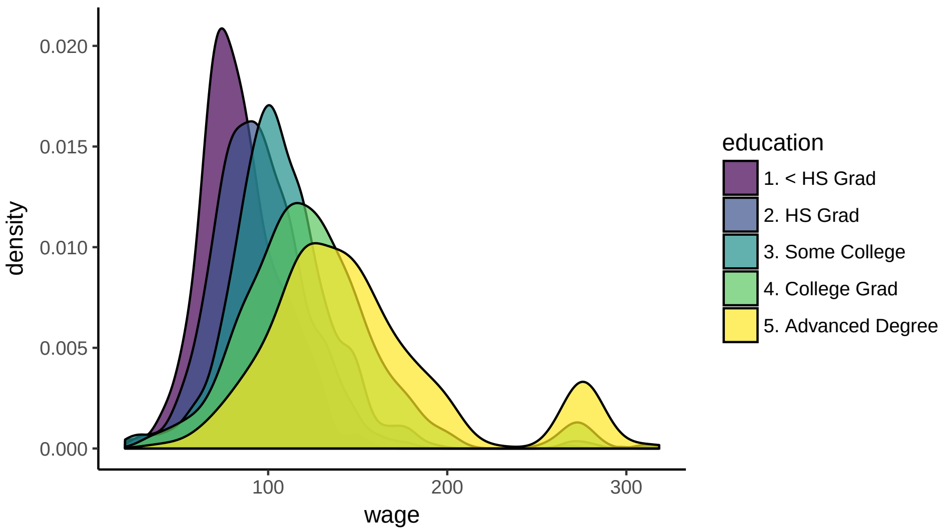

Figure 2: Gaussian kernel density estimates for wages split up by education level

- **race**: A factor with levels `1. White`, `2. Black`, `3. Asian` and `4. Other`
- **education**: A factor with levels `1. < HS Grad`, `2. HS Grad`, `3. Some College`, `4. College Grad` and `5. Advanced Degree`
- **health**: A factor with levels `1. <= Good` and `2. >= Very Good` indicating health level of worker

Naturally, the variable of interest and response variable will be the male workers wage. While doing first analyses, it is clear that the wage is highly dependent on the given variables. Figure 2 shows kernel density estimates (Gaussian) for the wage distribution depending on education level.

As visible in Figure 2, the kernel density estimates are critically different for each education level. In general, we can observe the trend that a higher education level leads to a higher expected income, but also to an increased variance. Therefore, both location and shape will be modeled when fitting the bamlss. Because income cannot be smaller than zero but does otherwise not have upper limits, the censored normal distribution with cut-off $a = 0$ will be chosen as the response family. After some data preparation, model estimation can then be achieved with the bamlss R package (Umlauf et al., 2017):

```
model <- bamlss(
  list(wage ~ s(age) + race + year + education + health,
       sigma ~ s(age) + race + year + education + health),
  data = wage_sub,
```

```
  family = cnorm_bamlss()
)
```

Code-Chunk 1: R code for fitting the bamlss based on Wage dataset

As visible in Code-Chunk 1, both μ and σ are modeled such that they relate to explanatory variables additively. Both parameters are connected to parametric effects `race`, `year`, `education` and `health`. The influence of `age` is specified with a thin-plate smooth spline.

3.3 Application Structure & Guide

As previously mentioned, **bamlss.vis** is implemented in the form of an R extension. For building and maintaining the package, GitHub is used. This allows users to easily install the package with the following R commands:

```
if (!require(devtools))
  install.packages("devtools")
devtools::install_github("Stan125/bamlss.vis")
```

Furthermore, **bamlss.vis** is strongly based on the Shiny framework (Chang et al., 2017), which is an R package designed to create interactive visualisations with HTML code and R functions. In the words of the author, Shiny combines “the computational power of R with the interactivity of the modern web” CITATION HERE.

In its core, a Shiny application is built using R functions and can therefore be called similarly. In the case of **bamlss.vis**, one can start the application with `bamlss.vis::vis()`.

4 Conclusion

Appendix

Bibliography

Andreas Brezger and Stefan Lang. Generalized structured additive regression based on bayesian p-splines. 50:967–991, 02 2006.

Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989.

Bureau of Labor Statistics. OES Data, 2016. URL `https://www.bls.gov/oes/tables.htm`.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2017. URL `https://CRAN.R-project.org/package=shiny`. R package version 1.0.5.

L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: Models, Methods and Applications*. Springer Berlin Heidelberg, 2013. ISBN 9783642343339. URL `https://books.google.de/books?id=EQxU9iJtipAC`.

Ludwig Fahrmeir, Thomas Kneib, and S. Lang. Penalized additive regression for space-time data: a bayesian perspective, 2003. URL `http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1687-9`.

Jerome H Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981.

W.H. Greene. *Econometric Analysis*. Pearson International Edition. Pearson Education, Limited, 2012. ISBN 9780273753568. URL `https://books.google.de/books?id=-WFPYgEACAAJ`.

T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1990. ISBN 9780412343902. URL `https://books.google.de/books?id=qa29r1Ze1coC`.

Gareth James, Daniela Witten, Trevor Hastie, and Rob Tibshirani. *ISLR: Data for an Introduction to Statistical Learning with Applications in R*, 2017. URL `https://CRAN.R-project.org/package=ISLR`. R package version 1.2.

Nadja Klein, Thomas Kneib, and Stefan Lang. Bayesian structured additive distributional regression. Working Papers in Economics and Statistics 2013-23, Innsbruck, 2013. URL `http://hdl.handle.net/10419/101101`.

Nadja Klein, Thomas Kneib, Stefan Lang, and Alexander Sohn. Bayesian structured additive distributional regression with an application to regional income inequality in germany. *Ann. Appl. Stat.*, 9(2):1024–1052, June 2015. doi: 10.1214/15-AOAS823. URL `https://doi.org/10.1214/15-AOAS823`.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

Robert I Lerman and Shlomo Yitzhaki. A note on the calculation and interpretation of the gini index. *Economics Letters*, 15(3-4):363–368, 1984.

Georges Matheron. Principles of geostatistics. *Economic geology*, 58(8):1246–1266, 1963.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

John A Nelder and Daryl Pregibon. An extended quasi-likelihood function. *Biometrika*, 74(2):221–232, 1987.

Chris O'Connor. Data: the key measure of relevance in a digital revolution. `https://www.ibm.com/blogs/internet-of-things/data-revolution/`, 2016.

U. Olsson. *Generalized Linear Models: An Applied Approach*. Lightning Source, 2002. ISBN 9789144041551. URL `https://books.google.de/books?id=SPljHQAACAAJ`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.

R.A. Rigby and D.M. Stasinopoulos. The gamlss project: a flexible approach to statistical modelling. In *New trends in statistical modelling: Proceedings of the 16th international workshop on statistical modelling*, volume 337, page 345. University of Southern Denmark, June 2001.

D Mikis Stasinopoulos, Robert A Rigby, et al. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23 (7):1–46, 2007.

The Economist. Rise of the machines. https://www.economist.com/news/briefing/21650526-artificial-intelligence-scares-peopleexcessively-so-rise-mach 2015.

Nikolaus Umlauf, Nadja Klein, and Achim Zeileis. Bamlss: Bayesian additive models for location, scale and shape (and beyond). Working papers, Working Papers in Economics and Statistics, 2017. URL https://EconPapers.repec.org/RePEc:inn:wpaper:2017-05.

United States Census Bureau. Supplement to current population survey. http://www.nber.org/cps/cpsmar11.pdf, March 2011.