

---

# **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 4, 2017  
by Stanislaus Stadlmann,  
born in Vienna, Austria



GEORG-AUGUST-UNIVERSITÄT  
GÖTTINGEN

# Contents

1	Introduction	1
2	Motivating Bayesian Additive Models for Location, Scale and Shape	2
2.1	Additive Models . . . . .	3
2.2	Structured Additive Regression Models . . . . .	4
2.2.1	Spatial Effects . . . . .	4
2.2.2	Interaction Terms . . . . .	5
3	bamlss.vis	6
4	Conclusion	6
	Bibliography	8

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).

In the ability to additively model multiple parameters of one distribution, bamlss bear many similarities with Generalized Additive Models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2001). Disparities lie in the estimation of model parameters, where bamlss utilize Markov Chain Monte Carlo (MCMC) simulations which provide credible intervals in situations where asymptotic maximum likelihood confidence intervals often fail, as well as bamlss' ability to model multivariate parametric distributions (Umlauf et al., 2017).

## 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  $y_1, \dots, y_n$  given the explanatory variables  $z_1, \dots, z_K$ , we obtain the following model equation:

$$y_i = f_1(z_{i1}) + f_2(z_{i2}) + \dots + f_k(z_{ik}) + \epsilon_i \quad (1)$$

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  $x_1, \dots, x_Q$ , we can extend (1) to a semiparametric regression model with the following specification:

$$y_i = \sum_{j=1}^K f_j(z_{ij}) + \underbrace{\sum_{l=1}^Q \beta_l x_{il}}_{\beta_0 + \beta_1 x_{i1} + \dots + \beta_Q x_{iQ}} + \epsilon_i \quad (2)$$

Eq. (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

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

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:

$$\mathbf{y} = \sum_{j=1}^K \mathbf{Z}_j \boldsymbol{\gamma}_j + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (3)$$

Here, the design matrices  $\mathbf{Z}_1, \dots, \mathbf{Z}_K$  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  $\epsilon_i$  are identically and independently (i.i.d) normally distributed with  $E(\epsilon_i) = 0$  and  $Var(\epsilon_i) = \sigma^2$  (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 (STAR) (Fahrmeir et al., 2013). 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  $i$ th 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), a geospatial effect is now added:

$$\begin{aligned} y_i &= \sum_{j=1}^K f_j(z_{ij}) + \sum_{l=1}^Q \beta_l x_{il} + f_{geo}(s_i) + \epsilon_i \\ &= \kappa^{add} + f_{geo}(s_i) + \epsilon_i \end{aligned} \tag{4}$$

$\kappa^{add}$  includes the non-spatial effects from (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 geoaddivitive components  $\mathbf{f}_{geo}$  can be written as

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

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

$$\mathbf{y} = \sum_{j=1}^K \mathbf{Z}_j \boldsymbol{\gamma}_j + \mathbf{X} \boldsymbol{\beta} + \mathbf{Z}_{geo} \boldsymbol{\gamma}_{geo} + \boldsymbol{\epsilon} \quad (5)$$

which bears similarities to the basis function approach in (3) (Fahrmeir et al., 2013, chap. 9.2).

### 2.2.2 Interaction Terms

The regression equation (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) in the following way:

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

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).



3 bamlss.vis

4 Conclusion

## Appendix

## Bibliography

- 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>.
- Jerome H Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981.
- 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>.
- 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>.
- Georges Matheron. Principles of geostatistics. *Economic geology*, 58(8):1246–1266, 1963.
- Chris O’Connor. Data: the key measure of relevance in a digital revolution. <https://www.ibm.com/blogs/internet-of-things/data-revolution/>, 2016.
- 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. 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.

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>.