A conceptional Lego toolbox for Bayesian distributional regression models

Nikolaus Umlauf[1], Reto Stauffer[1], Jakob W. Messner[1], Georg J. Mayr[2], Achim Zeileis[1]

[1] Department of Statistics, University of Innsbruck, Austria
[2] Institute of Meteorology and Geophysics, University of Innsbruck, Austria

E-mail for correspondence: Nikolaus.Umlauf@uibk.ac.at

Abstract: Bayesian analysis provides a convenient setting for the estimation of complex generalized additive regression models (GAM). Because of the very general structure of the additive predictor in GAMs, we propose an unified modeling architecture that can deal with a wide range of types of model terms and can benefit from different algorithms in order to estimate Bayesian distributional regression models.

Keywords: additive models; GAMLSS; distributional regression; MCMC.

1 Introduction

Bayesian estimation based on Markov chain Monte Carlo (MCMC) simulation is particularly attractive since it provides valid inference that does not rely on asymptotic properties and allows extensions such as variable selection or multilevel models. Existing estimation engines already provide infrastructures for a number of regression problems exceeding univariate responses, e.g., for multinomial, multivariate normal or mixed discrete-continuous distributed variables. In addition, most of the engines support random effect estimation that can be utilized for setting up complex models with additive predictors (see, e.g., Fahrmeir et al. 2013).
In order to ease the usage of already existing implementations and code, as well as to facilitate the development of new algorithms and extensions, we present an unified and entirely modular architecture for models with additive predictors which does not restrict to any type of regression problem. The approach follows the model class of generalized additive model for

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

location, scale and shape (GAMLSS, Rigby and Stasinopoulos 2005) but is more flexible and is sometimes referred to as distributional regression.

2 Model structure

The models discussed assume conditional independence of the response variable $y_1, \ldots, y_n$ given covariates. Within distributional regression, all parameters of the response distribution can be modeled by explanatory variables such that

$$\mathbf{y} \sim \mathcal{D}\left(h_1(\boldsymbol{\theta}_1) = \boldsymbol{\eta}_1,\ h_2(\boldsymbol{\theta}_2) = \boldsymbol{\eta}_2, \ldots,\ h_K(\boldsymbol{\theta}_K) = \boldsymbol{\eta}_K\right), \tag{1}$$

where $\mathcal{D}$ denotes any distribution available for the response variable and $\boldsymbol{\theta}_k$ are parameters that are linked to an additive predictor using known monotonic link functions $h_k(\cdot)$. The k-th additive predictor is given by

$$\eta = f_1(\mathbf{x}) + \ldots + f_p(\mathbf{x}), \tag{2}$$

where $\mathbf{x}$ represents a generic vector of all linear and nonlinear modeled covariates. The functions f_j are possibly smooth functions encompassing various types of effects, e.g., linear and nonlinear effects of continuous covariates, two-dimensional surfaces, spatially correlated effects, varying coefficients, random intercepts and random slopes, etc. Using a basis function approach, the vector of function evaluations can be written in matrix notation $\mathbf{f}_j = \mathbf{X}_j\boldsymbol{\beta}_j$ and can also be represented as a mixed model with $\mathbf{f}_j = \tilde{\mathbf{X}}_j\tilde{\boldsymbol{\gamma}}_j + \mathbf{U}_j\tilde{\boldsymbol{\beta}}_j$, where $\tilde{\boldsymbol{\gamma}}_j$ represents the fixed effects parameters and $\tilde{\boldsymbol{\beta}}_j \sim N(\mathbf{0}, \tau_j^2\mathbf{I})$ independent and i.i.d. random effects (see, e.g., Fahrmeir et al. 2013).

3 A conceptional Lego toolbox

For Bayesian inference, prior distributions need to be assigned to the regression coefficients. A general setup is obtained by using normal priors for $\boldsymbol{\beta}_j$ of the form

$$p(\boldsymbol{\beta}_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta_j}^\top\mathbf{K}_j\boldsymbol{\beta}_j\right), \tag{3}$$

where $\mathbf{K}_j$ is the so called penalty matrix that depends on the functional type chosen for f_j. The variance parameter τ_j^2 is equivalent to the inverse smoothing parameter in a frequentist approach and controls the trade off between flexibility and smoothness. A common choice of prior for the variance parameter is a weakly informative inverse Gamma hyperprior.

The main building block of all estimation engines is the logarithm of the posterior given by

$$ln\, p(\boldsymbol{\vartheta}|\mathbf{y}) = \ell(\boldsymbol{\vartheta}|\mathbf{y}) + \sum_{k=1}^{K}\sum_{j=1}^{p_k} \left\{ln\, p(\boldsymbol{\beta}_{jk}|\tau_{jk}^2) + ln\, p(\tau_{jk}^2)\right\}, \qquad (4)$$

with log-likelihood $\ell(\cdot)$ and priors $p(\cdot)$, e.g., given by (3), where $\boldsymbol{\vartheta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\tau}_1^2, \ldots, \boldsymbol{\tau}_K^2)^\top$. From a frequentist perspective (4) can be viewed as a penalized log-likelihood.
Moreover, gradient based algorithms require the evaluation of the first derivative or score vector as well as the second derivatives, e.g., when applying a Newton-Raphson type algorithm, or MCMC sampling using IWLS proposals (see, e.g., Fahrmeir et al. 2013). Because these quantities can be nicely decomposed using the chain rule and model terms are represented by an unified approach, algorithms for distributional regression models can be build by combining the following "Lego-bricks":

- The log-likelihood function $\ell(\boldsymbol{\vartheta}|\mathbf{y})$.
- The first order derivatives $\partial\ell(\boldsymbol{\vartheta}|\mathbf{y})/\partial\boldsymbol{\theta}_k$, $\partial\boldsymbol{\theta}_k/\partial\boldsymbol{\eta}_k$ and $\partial\boldsymbol{\eta}_k/\partial\boldsymbol{\vartheta}_k$.
- Second order derivatives $\partial^2\ell(\boldsymbol{\vartheta}|\mathbf{y})/\partial\boldsymbol{\eta}_k\partial\boldsymbol{\eta}_k^\top$ (and expectations).
- Derivatives for priors, e.g., $ln\, p(\boldsymbol{\beta}_{jk}|\tau_{jk}^2)$ and $ln\, p(\tau_{jk}^2)$.

Hence, a modular system can in principle be used to implement various estimation algorithms (also using existing software). A simple generic algorithm for distributional regression models is outlined by the following pseudo code:

```
while(eps > ε & i < maxit) {
    for(k in 1:K) {
        for(j in 1:p) {
            Compute η_-j^[k] = η^[k] − f_j^[k].
            Obtain new (β_j^[k], τ^2_j^[k])^⊤ = u_j^[k](y, η_-j^[k], X_j^[k], β_j^[k], τ^2_j^[k], family, k).
            Update η^[k].
        }
    }
    Compute new eps
}
```

The algorithm does not distinguish between optimization or sampling, because the functions $\mathtt{u}_\mathtt{j}^{[\mathtt{k}]}(\cdot)$ could either return proposals from a MCMC sampler or updates from an optimization algorithm. Moreover, it is possible to use different update functions for model terms within predictors, e.g., IWLS proposals combined with slice sampling or Hamiltonian Monte Carlo. An implementation of the modular infrastructure is provided in the R package **bamlss** (available at `https://R-forge.R-project.org` at the time of writing).

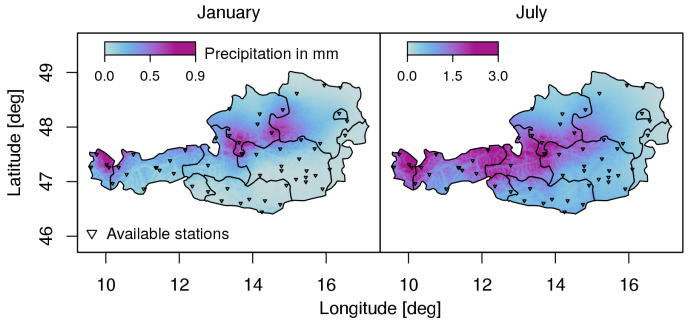

FIGURE 1. Predicted average precipitation for 10th of January and 10th of July. Animation available at `http://eeecon.uibk.ac.at/~umlauf/data/austria.gif`

4 Example

As an illustration, we analyze precipitation data taken from the HOMSTART project conducted at the Zentralanstalt für Meteorologie und Geodynamik (ZAMG, see also Umlauf et. al 2012). The aim is to estimate a good climatology which can be used for subsequent meteorolical models. Since precipitation data is skewed and exhibits high density at zero observations, we estimate a censored normal additive regression model with latent Gaussian variable $\mathbf{y}^\star$ and observed response $\mathbf{y}$, the square root of daily precipitation observations. The model is given by

$$\mathbf{y}^\star \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2), \quad \boldsymbol{\mu} = \boldsymbol{\eta}_\mu, \quad \log(\boldsymbol{\sigma}) = \boldsymbol{\eta}_\sigma, \quad \mathbf{y} = max(\mathbf{0}, \mathbf{y}^\star).$$

For both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, we use the following additive predictor:

$$\boldsymbol{\eta} = \beta_0 + f_1(\texttt{day}, \texttt{lon}, \texttt{lat}) + f_2(\texttt{lon}, \texttt{lat}) + f_3(\texttt{day}) + f_4(\texttt{alt}),$$

where function f_1 is a spatially varying seasonal effect, f_2 a spatially correlated effect, f_3 the seasonal and f_4 the altitude effect. The resulting climatology for two particular days of the year are shown in Figure 1.

References

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013). *Regression Models, Methods and Applications*. Springer-Verlag, Berlin. ISBN 978-3-642-34332-2.

Rigby, R.A. and Stasinopoulos, D.M. (2005) Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society C*, 54(3), 507 – 554.

Umlauf, N., Mayr, G., Messner, J., and Zeileis, A. (2012). Why does it always rain on me? A spatio-temporal analysis of precipitation in Austria. *Austrian Journal of Statistics*, 41(1), 81 – 92.