Bayesian Distributional Non-Linear Multilevel Modeling with the R Package brms

Paul-Christian Bürkner
University of Münster

Abstract

The **brms** package allows R users to easily specify a wide range of Bayesian multilevel models, which are fitted with the probabilistic programming language Stan behind the scenes. A wide range of response distributions are supported in combination with an intuitive and powerful multilevel formula syntax. Non-linear relationships may be specified using non-linear predictor terms, splines or Gaussian processes. Additionally, all parameters of the response distribution can be predicted at the same time allowing for distributional regression. Model fit can be investigated and compared using leave-one-out cross-validation and graphical posterior-predictive checks. Many more post-processing and plotting methods are implemented.

Keywords: Bayesian inference, multilevel models, distributional models, **lme4**, Stan, R.

1. Introduction

Multilevel models (MLMs) offer great flexibility for researchers across sciences (Brown and Prescott 2015; Demidenko 2013; Gelman and Hill 2006; Pinheiro and Bates 2006). They allow modeling of data measured on different levels at the same time – for instance data of students nested within classes and schools – thus taking complex dependency structures into account. It is not surprising that many packages for R (R Core Team 2015) have been developed to fit MLMs. Usually, however, the functionality of these implementations is limited insofar as it is only possible to predict the mean of the response distribution. Other parameters of the response distribution, such as the residual standard deviation in linear models, are assumed constant across observations, which may be violated in many applications. Accordingly, it is desirable to allow for prediction of *all* response parameters at the same time. Models doing exactly that are often referred to as *distributional models* or more verbosely *models for location, scale and shape* (Rigby and Stasinopoulos 2005). Another limitation of basic MLMs is that they only allow for linear predictor terms. While linear predictor terms already offer good flexibility, they are of limited use when relationships are inherently non-linear. Such non-linearity can be handled in at least two ways: (1) by fully specifying a non-linear predictor term with corresponding parameters each of which can be predicted using MLMs (Lindstrom and Bates 1990), or (2) estimating the form of the non-linear relationship on the fly using splines (Wood 2004) or Gaussian processes (Rasmussen and Williams 2006). The former are often simply called *non-linear models*, while models applying splines are referred to as *generalized additive models* (GAMs; Hastie and Tibshirani, 1990).

Combining all of these modeling options into one framework is a complex task, both conceptually and with regard to model fitting. Maximum likelihood methods, which are typically applied in classical 'frequentist' statistics, can reach their limits at some point and fully Bayesian methods become the go-to solutions to fit such complex models (Gelman, Carlin, Stern, and Rubin 2014). In addition to being more flexible, the Bayesian framework comes with other advantages, for instance, the ability to derive probability statements for every quantity of interest or explicitly incorporating prior knowledge about parameters into the model.

Possibly the most powerful program for performing full Bayesian inference available to date is Stan (Stan Development Team 2017b; Carpenter, Gelman, Hoffman, Lee, Goodrich, Betancourt, Brubaker, Guo, Li, and Ridell 2017). It implements Hamiltonian Monte Carlo (Duane, Kennedy, Pendleton, and Roweth 1987; Neal 2011; Betancourt, Byrne, Livingstone, and Girolami 2014) and its extension, the No-U-Turn (NUTS) Sampler (Hoffman and Gelman 2014). These algorithms converge much more quickly than other Markov-Chain Monte-Carlo (MCMC) algorithms especially for high-dimensional models (Hoffman and Gelman 2014; Betancourt *et al.* 2014; Betancourt 2017). An excellent non-mathematical introduction to Hamiltonian Monte Carlo can be found in Betancourt (2017).

Stan comes with its own programming language, allowing for great modeling flexibility (cf., Stan Development Team 2017b; Carpenter *et al.* 2017). Many researchers may still be hesitent to use Stan directly, as every model has to be written, debugged and possibly also optimized. This may be a time-consuming and error-prone process even for researchers familiar with Bayesian inference. The **brms** package Bürkner (in press), presented in this paper, aims to remove these hurdles for a wide range of regression models by allowing the user to benefit from the merits of Stan by using extended **lme4**-like (Bates, Mächler, Bolker, and Walker 2015) formula syntax, with which many R users are familiar with. It offers much more than writing efficient and human-readable Stan code: **brms** comes with many post-processing and visualization functions, for instance to perform posterior predictive checks, leave-one-out cross-validation, visualization of estimated effects, and prediction of new data. The overarching aim is to have one general framework for regression modeling, which offers everything required to successfully apply regression models to complex data. To date, it already fully replaces and extends the functionality of dozens of other R packages, each of which is restricted to specific regression models.

The purpose of the present article is to provide an introduction of the advanced multilevel formula syntax implemented in **brms**, which allows you to fit a wide and growing range of non-linear distributional multilevel models. A general overview of the package is already given in Bürkner (in press). Accordingly, the present article focuses on more recent developments. We begin by explaining the underlying structure of distributional models. Next, the formula syntax of **lme4** and its extensions implemented in **brms** are explained. Three examples that demonstrate the use of the new syntax are discussed in detail. We end by describing future plans for extending the package. Users more interested in application and less in mathematical details and notation may want to skip Section 2 and 3.

2. Model description

The core of models implemented in **brms** is the prediction of the response y through predicting

all parameters θ_p of the response distribution D, which is also called the model `family` in many R packages. We write

$$y_i \sim D(\theta_{1i}, \theta_{2i}, ...)$$

to stress the dependency on the i^{th} observation. Every parameter θ_p may be regressed on its own predictor term η_p transformed by the inverse link function f_p that is $\theta_{pi} = f_p(\eta_{pi})$[1]. Such models are typically refered to as *distributional models*[2]. Details about the parameterization of each `family` are given in `vignette("brms_families")`.

Suppressing the index p in all quantities for simplicity, a predictor term η can generally be written as

$$\eta = \mathbf{X}\beta + \mathbf{Z}u + \sum_{k=1}^{K} s_k(x_k)$$

In this equation, β and u are the coefficients at population-level and group-level respectively and $\mathbf{X}, \mathbf{Z}$ are the corresponding design matrices. The terms $s_k(x_k)$ symbolize optional smoothing terms based on some covariates x_k fitted via splines (see Wood (2011) for the underlying implementation in the **mgcv** package). The response y as well as $\mathbf{X}$, $\mathbf{Z}$, and x_k make up the data, whereas β, u, and the smooth functions s_k are the model parameters being estimated. The coefficients β and u may be more commonly known as fixed and random effects, but I avoid theses terms following the recommendations of Gelman and Hill (2006). Details about prior distributions of β and u can be found in Bürkner (in press) and under `help("set_prior")`.

Each `family` in **brms** has one primary parameter θ_1, which is usually (but not necessarily) the mean of the distribution. The corresponding predictor term η_1 may have any form specifiable in Stan. We call it a *non-linear* predictor and write

$$\eta_1 = f(c_1, c_2, ..., \phi_1, \phi_2, ...)$$

The structure of the function f is given by the user, c_r are known or observed covariates, and ϕ_s are non-linear parameters each having its own linear predictor term η_{ϕ_s} of the form specified above. In fact, we should think of non-linear parameters as placeholders for linear predictor terms rather than as parameters themselves. A frequentist implementation of such models, which inspired the non-linear syntax in **brms**, can be found in the **nlme** package (Pinheiro, Bates, DebRoy, Sarkar, and R Core Team 2016). Theoretically, one could allow non-linear predictors for all parameters of D, but this would complicate the current implementation and might be unnecessary for the vast majority of research questions.

3. Extended multilevel formula syntax

The formula syntax applied in **brms** builds upon the syntax of the R package **lme4** (Bates *et al.* 2015). First, we will briefly explain the **lme4** syntax used to specify multilevel models and then introduce certain extensions that allow to specify much more complicated models in **brms**. An **lme4** formula has the general form

```
response ~ pterms + (gterms | group)
```

[1] A parameter can also be assumed constant across observations so that a linear predictor is not requireed.
[2] The models described in Bürkner (in press) are a sub-class of the here described models.

The `pterms` part contains the population-level effects that are assumed to be the same across obervations. The `gterms` part contains so called group-level effects that are assumed to vary accross grouping variables specified in `group`. Multiple grouping factors each with multiple group-level effects are possible. Usually, `group` contains only a single variable name pointing to a factor, but you may also use `g1:g2` or `g1/g2`, if both `g1` and `g2` are suitable grouping factors. The `:` operator creates a new grouping factor that consists of the combined levels of `g1` and `g2` (you could think of this as pasting the levels of both factors together). The `/` operator indicates nested grouping structures and expands one grouping factor into two or more when using multiple `/` within one term. If, for instance, you write `(1 | g1/g2)`, it will be expanded to `(1 | g1) + (1 | g1:g2)`. Instead of `|` you may use `||` in grouping terms to prevent group-level correlations from being modeled. This may be useful in particular when modeling so many group-level effects that convergence of the fitting algorithms becomes an issue due to model complexity. One limitation of the `||` operator in **lme4** is that it only splits up terms so that columns of the design matrix originating from the same term are still modeled as correlated (e.g., when coding a categorical predictor; see the `mixed` function of the **afex** package by Singmann, Bolker, and Westfall (2015) for a way to avoid this behavior).

While intuitive and visually appealing, the classic **lme4** syntax is not flexible enough to allow for specifying the more complex models supported by **brms**. In non-linear or distributional models, for instance, multiple parameters are predicted, each having their own population and group-level effects. Hence, multiple formulas are necessary to specify such models[3]. Then, however, specifying group-level effects of the same grouping factor to be correlated *across* formulas becomes complicated. The solution implemented in **brms** (and currently unique to it) is to expand the `|` operator into `|<ID>|`, where `<ID>` can be any value. Group-level terms with the same `ID` will then be modeled as correlated if they share same grouping factor(s)[4]. For instance, if the terms `(x1|ID|g1)` and `(x2|ID|g1)` appear somewhere in the formulas passed to **brms**, they will be modeled as correlated.

Further extensions of the classical **lme4** syntax refer to the `group` part. It is rather limited in its flexibility since only variable names combined by `:` or `/` are supported. We propose two extensions of this syntax: Firstly, `group` can generally be split up in its terms so that, say, `(1 | g1 + g2)` is expanded to `(1 | g1) + (1 | g2)`. This is fully consistent with the way `/` is handled so it provides a natural generalization to the existing syntax. Secondly, there are some special grouping structures that cannot be expressed by simply combining grouping variables. For instance, multi-membership models cannot be expressed this way. To overcome this limitation, we propose wrapping terms in `group` within special functions that allow specifying alternative grouping structures: `(gterms | fun(group))`. In **brms**, there are currently two such functions implemented, namely `gr` for the default behavior and `mm` for multi-membership terms. To be compatible with the original syntax and to keep formulas short, `gr` is automatically added internally if none of these functions is specified.

There are other syntax extensions implemented in **brms** that do not directly target grouping terms. Firstly, there are terms formally included in the `pterms` part that are handled

[3] Actually, it is possible to specify multiple model parts within one formula using interactions terms for instance as implemented in **MCMCglmm** (Hadfield 2010). However, this syntax is limited in flexibility and requires a rather deep understanding of the way R parses formulas, thus often being confusing to users.

[4] It might even be further extended to `|fun(<ID>)|`, where `fun` defines the type of correlation structure, defaulting to unstructured that is estimating the full correlation matrix. The `fun` argument is not yet supported by **brms** but could be supported in the future if other correlation structures, such as compound symmetry or Toeplitz, turn out to have reasonable practical applications and effective implementations in Stan.

separately. The most prominent examples are smooth terms specified through the `s` and `t2` functions of the **mgcv** package (Wood 2011). Other examples are category specific effects `cs`, monotonic effects `mo`, noise-free effects `me`, or Gaussian process terms `gp`. The former is explained in Bürkner (in press), while the latter thee are documented in `help(brmsformula)`. Internally, these terms are extracted from `pterms` and not included in the construction of the population-level design matrix $\mathbf{X}$. Secondly, making use of the fact that `|` is unused on the left-hand side of $\sim$ in formula, additional information on the response variable may be specified via `response | aterms ~ <predictor terms>`. The `aterms` part may contain multiple terms of the form `fun(<variable>)` seperated by `+` each providing special information on the response variable. `fun` can be replaced with either `se`, `weights`, `disp`, `trials`, `cat`, `cens`, `trunc`, or `dec`. As it is not the main topic of the present paper, we refer to `help("brmsformula")` and `help("addition-terms")` for more details.

4. Examples

In this section, we will discuss three examples in detail. The first is about the number of fish caught be different groups of people. It does not actually contain any multilevel structure, but helps in understanding how to set up distributional models. The second example is about cumulative insurance loss payments across several years, which is fitted using a rather complex non-linear multilevel model. The third example is about the performance of school children, who change school during the year, thus requiring a multi-membership model.

Despite not being covered in the three examples, there are a few more modeling options that we want to briefly describe. First, **brms** allows fitting so called phylogenetic models. These models are relevant in evolutionary biology when data of many species are analyzed at the same time. Species are not independent as they come from the same phylogenetic tree, implying that different levels of the same grouping-factor (i.e., species) are likely correlated. There is a whole vignette dedicated to this topic, which can be found via `vignette("brms_phylogenetics")`. Second, there is a canonical way to handle ordinal predictors, without falsely assuming they are either categorical or continuous. We call them monotonic effects and discuss them in `vignette("brms_monotonic")`. Last but not least, it is possible to account for measurement error in both response and predictor variables. This is often ignored in applied regression modeling (Westfall and Yarkoni 2016), although measurement error is very common in all scientific fields making use of observational data. There is no vignette yet covering this topic, but one will be added in the future. In the meantime, `help("brmsformula")` is the best place to start learning about such models as well as about other details of the **brms** formula syntax.

4.1. Example 1: Catching fish

An important application of the distributional regression framework of **brms** are so called zero-inflated and hurdle models. These models are helpful whenever there are more zeros in the response variable than one would naturally expect. For example, if you seek to predict the number of cigarettes people smoke per day and also includes non-smokers, there will be a huge amount of zeros which, when not modeled appropriately, can seriously distort parameter estimates. Here, we consider another example dealing with the number of fish caught by various groups of people. On the UCLA website (`http://www.ats.ucla.edu/stat/r/dae/zipoisson.htm`), the data are described as follows: "The state wildlife biologists want to

model how many fish are being caught by fishermen at a state park. Visitors are asked how long they stayed, how many people were in the group, were there children in the group and how many fish were caught. Some visitors do not fish, but there is no data on whether a person fished or not. Some visitors who did fish did not catch any fish so there are excess zeros in the data because of the people that did not fish."

```
R> zinb <- read.csv("http://stats.idre.ucla.edu/stat/data/fish.csv")
R> zinb$camper <- factor(zinb$camper, labels = c("no", "yes"))
R> head(zinb)
```

```
  nofish livebait camper persons child         xb         zg count
1      1        0     no       1     0 -0.8963146  3.0504048     0
2      0        1    yes       1     0 -0.5583450  1.7461489     0
3      0        1     no       1     0 -0.4017310  0.2799389     0
4      0        1    yes       2     1 -0.9562981 -0.6015257     0
5      0        1     no       1     0  0.4368910  0.5277091     1
6      0        1    yes       4     2  1.3944855 -0.7075348     0
```

As predictors we choose the number of people per group, the number of children, as well as whether the group consists of campers. Many groups may not catch any fish just because they do not try and so we fit a zero-inflated Poisson model. For now, we assume a constant zero-inflation probability across observations.

```
R> fit_zinb1 <- brm(count ~ persons + child + camper, data = zinb,
R>                  family = zero_inflated_poisson())
```

The model is readily summarized via

```
R> summary(fit_zinb1)
```

```
 Family: zero_inflated_poisson (log)
Formula: count ~ persons + child + camper
   Data: zinb (Number of observations: 250)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000
   WAIC: Not computed

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept    -1.01      0.17    -1.34    -0.67       2171    1
persons       0.87      0.04     0.79     0.96       2188    1
child        -1.36      0.09    -1.55    -1.18       1790    1
camper        0.80      0.09     0.62     0.98       2950    1

Family Specific Parameters:
   Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
```

```
zi      0.41       0.04      0.32      0.49        2409    1

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

A graphical summary is available through

```
R> marginal_effects(fit_zinb1)
```

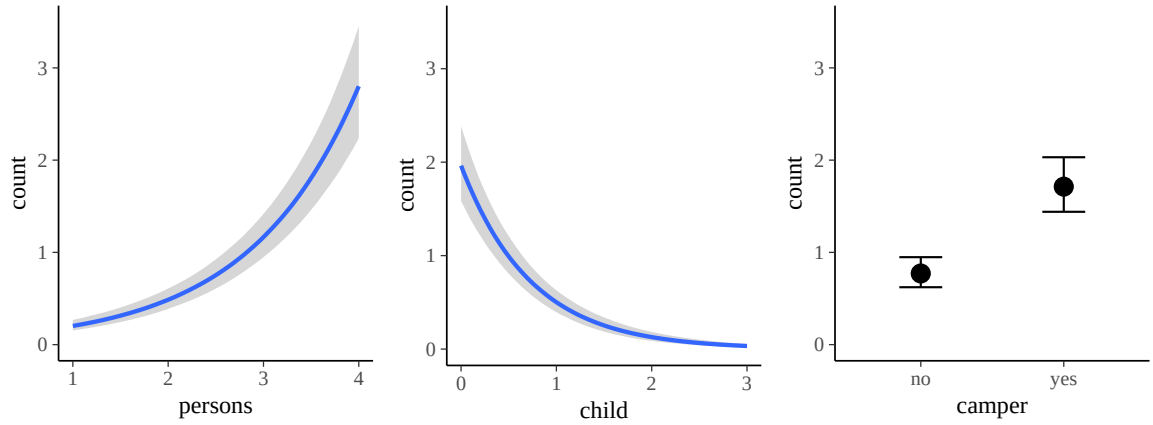

Figure 1: Marginal effects plots of the `fit_zinb1` model.

(see Figure 1). In fact, the `marginal_effects` method turned out to be so powerful in visualizing effects of predictors that I am using it almost as frequently as `summary`. According to the parameter estimates, larger groups catch more fish, campers catch more fish than non-campers, and groups with more children catch less fish. The zero-inflation probability `zi` is pretty large with a mean of 41%. Please note that the probability of catching no fish is actually higher than 41%, but parts of this probability are already modeled by the Poisson distribution itself (hence the name zero-*inflation*). If you want to treat all zeros as originating from a separate process, you can use hurdle models instead (not shown here).

Now, we try to additionally predict the zero-inflation probability by the number of children. The underlying reasoning is that we expect groups with more children to not even try catching fish, since children often lack the patience required for fishing. From a purely statistical perspective, zero-inflated (and hurdle) distributions are a mixture of two processes and predicting both parts of the model is natural and often very reasonable to make full use of the data.

```
R> fit_zinb2 <- brm(bf(count ~ persons + child + camper, zi ~ child),
R>                  data = zinb, family = zero_inflated_poisson())
```

To transform the linear predictor of `zi` into a probability, **brms** applies the logit-link, which takes values within $[0, 1]$ and returns values on the real line. Thus, it allows the transition between probabilities and linear predictors.

```
R> summary(fit_zinb2)

 Family: zero_inflated_poisson (log)
Formula: count ~ persons + child + camper
         zi ~ child
   Data: zinb (Number of observations: 250)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000
   WAIC: Not computed

Population-Level Effects:
             Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept       -1.07      0.18    -1.43    -0.73       2322    1
persons          0.89      0.05     0.80     0.98       2481    1
child           -1.17      0.10    -1.37    -1.00       2615    1
camper           0.78      0.10     0.60     0.96       3270    1
zi_Intercept    -0.95      0.27    -1.52    -0.48       2341    1
zi_child         1.21      0.28     0.69     1.79       2492    1

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

According to the model, trying to fish with children not only decreases the overall number fish caught (as implied by the Poisson part of the model) but also drastically increases your chance of catching no fish at all (as implied by the zero-inflation part), possibly because groups with more children spend less time or no time at all fishing. Comparing model fit via leave-one-out cross validation as implemented in the **loo** package (Vehtari, Gelman, and Gabry 2016a,b).

```
R> LOO(fit_zinb1, fit_zinb2)

                       LOOIC     SE
fit_zinb1            1639.52 363.30
fit_zinb2            1621.35 362.39
fit_zinb1 - fit_zinb2   18.16  15.71
```

reveals that the second model using the number of children to predict both model parts has better fit. However, when considering the standard error of the `LOOIC` difference, improvement in model fit is apparently modest and not substantial. More examples of distributional model can be found in `vignette("brms_distreg")`.

4.2. Example 2: Insurance loss payments

On his blog, Markus Gesmann predicts the growth of cumulative insurance loss payments over time, originated from different origin years (see http://www.magesblog.com/2015/11/loss-developments-via-growth-curves-and.html). We will use a slightly simplified version of his model for demonstration purposes here. It looks as follows:

$$cum_{AY,dev} \sim N(\mu_{AY,dev}, \sigma)$$
$$\mu_{AY,dev} = ult_{AY}\left(1 - \exp\left(-\left(\frac{dev}{\theta}\right)^{\omega}\right)\right)$$

The cumulative insurance payments *cum* will grow over time, and we model this dependency using the variable *dev*. Further, ult_{AY} is the (to be estimated) ultimate loss of accident each year. It constitutes a non-linear parameter in our framework along with the parameters θ and ω, which are responsible for the growth of the cumulative loss and are for now assumed to be the same across years. We load the data

```
R> url <- paste0("https://raw.githubusercontent.com/mages/",
R>               "diesunddas/master/Data/ClarkTriangle.csv")
R> loss <- read.csv(url)
R> head(loss)
```

```
    AY dev      cum
1 1991   6  357.848
2 1991  18 1124.788
3 1991  30 1735.330
4 1991  42 2182.708
5 1991  54 2745.596
6 1991  66 3319.994
```

and translate the proposed model into a non-linear **brms** model.

```
R> nlform <- bf(cum ~ ult * (1 - exp(-(dev / theta)^omega)),
R>              ult ~ 1 + (1|AY), omega ~ 1, theta ~ 1, nl = TRUE)

R> nlprior <- c(prior(normal(5000, 1000), nlpar = "ult"),
R>              prior(normal(1, 2), nlpar = "omega"),
R>              prior(normal(45, 10), nlpar = "theta"))

R> fit_loss1 <- brm(formula = nlform, data = loss, family = gaussian(),
R>                  prior = nlprior, control = list(adapt_delta = 0.9))
```

In the above functions calls, quite a few things are going on. The formulas are wrapped in `bf` to combine them into one object. The first formula specifies the non-linear model. We set argument `nl = TRUE` so that **brms** takes this formula literally and instead of using standard R formula parsing. We specify one additional formula per non-linear parameter (a) to clarify what variables are covariates and what are parameters and (b) to specify the predictor term for the parameters. We estimate a group-level effect of accident year (variable `AY`) for the ultimate loss `ult`. This also shows nicely how a non-linear parameter is actually a placeholder for a linear predictor, which in the case of `ult`, contains only a varying intercept for year. Both `omega` and `theta` are assumed to be constant across observations so we just fit a population-level intercept.

Priors on population-level effects are required and, for the present model, are actually mandatory to ensure identifiability. Otherwise, we may observe that different Markov chains converge to different parameter regions as multiple posterior distribution are equally plausible. In the `control` argument we increase `adapt_delta` to get rid of a few divergent transitions (cf. Stan Development Team, 2017a; Bürkner, in press). Again the model is summarized via

```
R> summary(fit_loss1)
```

```
 Family: gaussian (identity)
Formula: cum ~ ult * (1 - exp(-(dev / theta)^omega))
         ult ~ 1 + (1 | AY)
         omega ~ 1
         theta ~ 1
   Data: loss (Number of observations: 55)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000
   WAIC: Not computed

Group-Level Effects:
~AY (Number of levels: 10)
                  Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(ult_Intercept)   745.74    231.31   421.05  1306.04        916    1

Population-Level Effects:
                Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
ult_Intercept    5273.70    292.34  4707.11  5852.28        798    1
omega_Intercept     1.34      0.05     1.24     1.43       2167    1
theta_Intercept    46.07      2.09    42.38    50.57       1896    1

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma   139.93     15.52    113.6   175.33       2358    1

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

as well as

```
R> marginal_effects(fit_loss1)
```

(see Figure 2). We can also visualize the cumulative insurance loss over time separately for each year.

```
R> conditions <- data.frame(AY = unique(loss$AY))
R> rownames(conditions) <- unique(loss$AY)
```

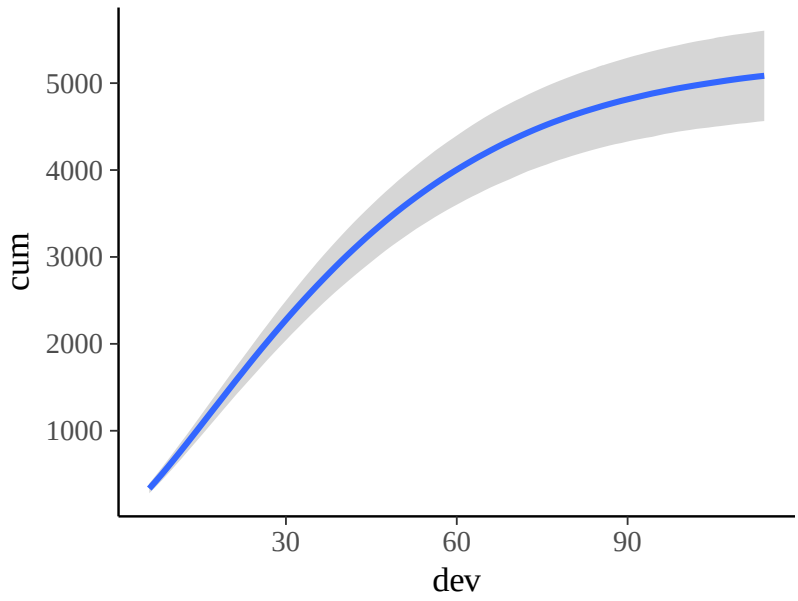

Figure 2: Marginal effects plots of the `fit_loss1` model.

```
R> me_year <- marginal_effects(fit_loss1, conditions = conditions,
R>                             re_formula = NULL, method = "predict")
R> plot(me_year, ncol = 5, points = TRUE)
```

(see Figure 3). It is evident that there is some variation in cumulative loss across accident years, for instance due to natural disasters happening only in certain years. Further, we see that the uncertainty in the predicted cumulative loss is larger for later years with fewer available data points.

In the above model, we considered `omega` and `delta` to be constant across years, which may not necessarily be true. We can easily investigate this by fitting varying intercepts for all three non-linear parameters also estimating group-level correlation using the above introduced `ID` syntax.

```
R> nlform2 <- bf(cum ~ ult * (1 - exp(-(dev / theta)^omega)),
R>               ult ~ 1 + (1|ID1|AY), omega ~ 1 + (1|ID1|AY),
R>               theta ~ 1 + (1|ID1|AY), nl = TRUE)

R> fit_loss2 <- update(fit_loss1, formula = nlform2,
R>                     control = list(adapt_delta = 0.90))
```

We could have also specified all predictor terms more conveniently within one formula as `ult + omega + theta ~ 1 + (1|ID1|AY)`, because the structure of the predictor terms is identical. To compare model fit, we perform leave-one-out cross-validation.

```
R> LOO(fit_loss1, fit_loss2)
```

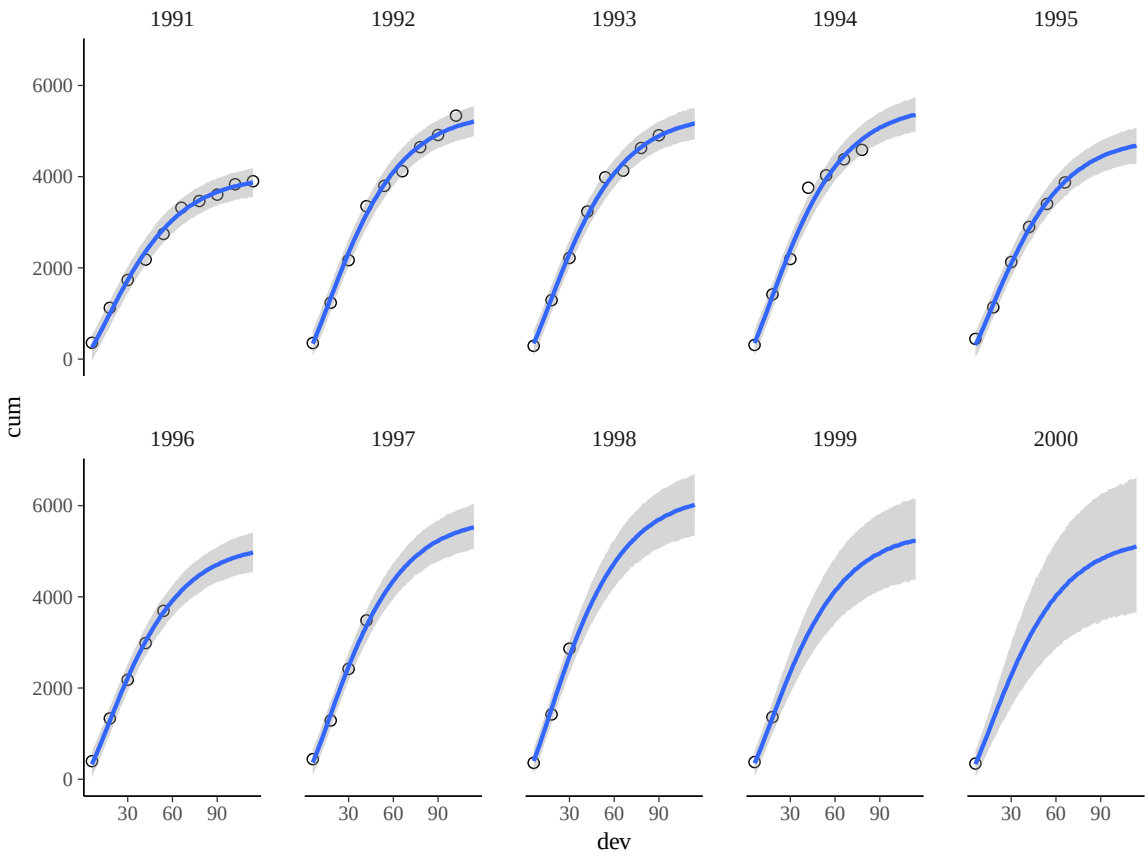

Figure 3: Marginal effects plots of the `fit_loss1` model separately for each accident year.

```
                      LOOIC    SE
fit_loss1              715.44 19.24
fit_loss2             720.60 19.85
fit_loss1 - fit_loss2  -5.15  5.34
```

Since smaller values indicate better expected out-of-sample predictions and thus better model fit, the simpler model that only has a varying intercept over parameter `ult` is preferred. This may not be overly surprising, given that three varying intercepts as well as three group-level correlations are probably overkill for data containing only 55 observations. Nevertheless, it nicely demonstrates how to apply the `ID` syntax in practice.

With the above models, we achieve good fit to the data thanks to a tailored non-linear function provided by subject matter experts. Now assume that such a knowledge is not available. We can try finding a reasonable function trough trial and error, or we can use a statistical procedure to find a good function for us. The latter approach naturally leads to fitting smooth terms via splines, which is readily applied to the insurance loss data using **brms**. All the complicated stuff related to the data for setting up smooth terms is handled by **mgcv** (Wood 2011) on the backend.

```
R> fit_loss3 <- brm(cum ~ s(dev) + (1|AY), data = loss)
```

As the smooth term itself cannot be modeled as varying by year in a multilevel manner, we add a basic varying intercept in an effort to account for variation between years. The usual methods such as `summary` or `marginal_effects` can be applied to obtain numerical or graphical summaries of the model. The predictions of each year are looking a follows.

```
R> me_year3 <- marginal_effects(fit_loss3, conditions = conditions,
R>                              re_formula = NULL, method = "predict")
R> plot(me_year3, ncol = 5, points = TRUE)
```

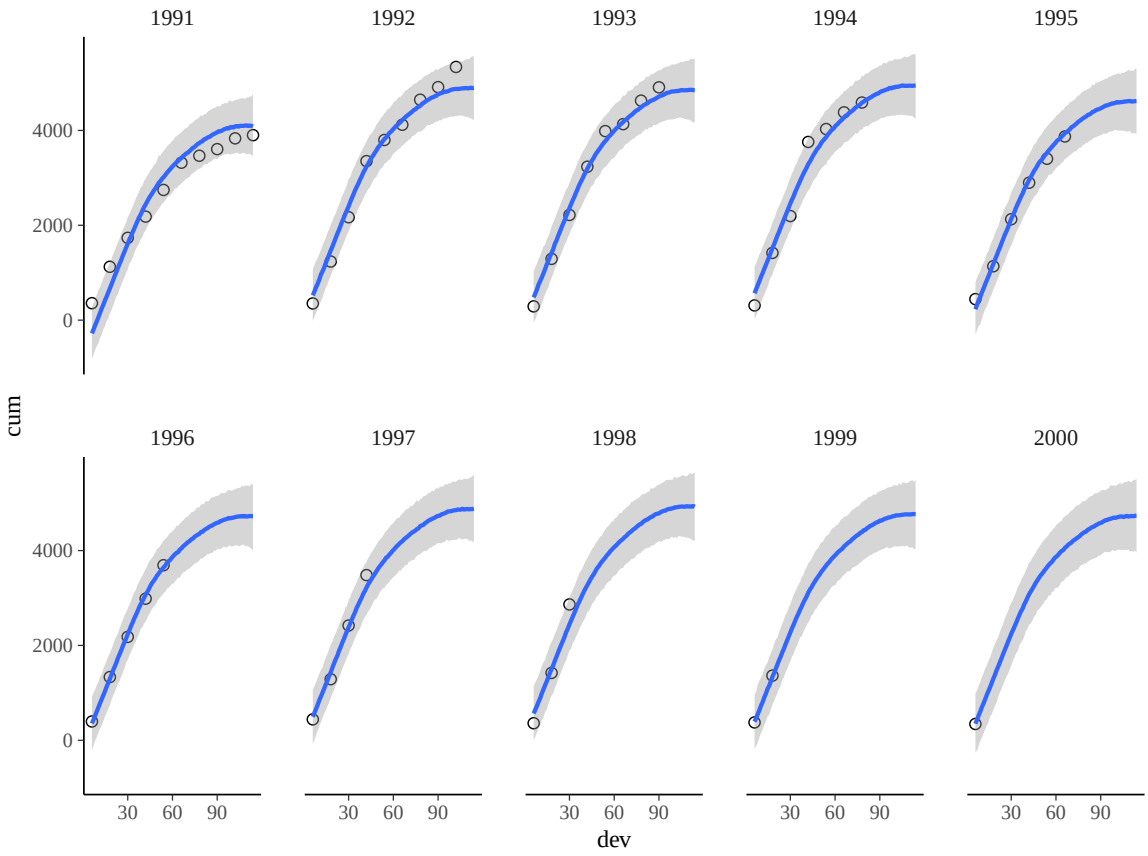

Figure 4: Marginal effects plots of the `fit_loss3` model separately for each accident year.

(see Figure 4). Again we can compare model fit via leave-one-out cross-validation.

```
R> LOO(fit_loss1, fit_loss3)
```

```
                       LOOIC    SE
fit_loss1             715.44 19.24
fit_loss3             780.03 12.91
fit_loss1 - fit_loss3 -64.58 18.59
```

We see that `fit_loss3` clearly fits worse than the specifically designed non-linear model `fit_loss1`, but the former still performs impressively given that we just fitted a non-linear

function to the data without knowing anything about its form a-priori. An alternative to splines for fitting a-priori unknown non-linear relationships are latent Gaussian processes (Rasmussen and Williams 2006). In **brms**, Gaussian process terms can be specified in the model formula via function `gp` in a similar manner as splines via `s` or `t2`. More examples of non-linear models can be found in `vignette("brms_nonlinear")`.

4.3. Example 3: Performance of school children

Suppose that we want to predict the performance of students in the final exams at the end of the year. There are many variables to consider, but one important factor will clearly be school membership. Schools might differ in the ratio of teachers and students, the general quality of teaching, in the cognitive ability of the students they draw, or other factors we are not aware of that induce dependency among students of the same school. Thus, it is advised to apply a multilevel modeling techniques including school membership as a group-level term. Of course, we should account for class membership and other levels of the educational hierarchy as well, but for the purpose of the present example, we will focus on schools only. Usually, accounting for school membership is pretty-straight forward by simply adding a varying intercept to the formula: `(1 | school)`. However, a non-negligible number of students might change schools during the year. This would result in a situation where one student is a member of multiple schools and so we need a multi-membership model. Setting up such a model not only requires information on the different schools students attend during the year, but also the amount of time spend at each school. The latter can be used to weight the influence each school has on its students, since more time attending a school will likely result in greater influence. For now, let us assume that students change schools maximally once a year and spend equal time at each school. We will later see how to relax these assumptions.

Real educational data are usually relatively large and complex so that we simulate our own data of 10 schools and 1000 students, with each school having the same expected number of 100 students. We model 10% of students as changing schools.

```
R> data_mm <- sim_multi_mem(nschools = 10, nstudents = 1000, change = 0.1)
R> head(data_mm)
```

```
  s1 s2  w1  w2        y
1  8  9 0.5 0.5 16.27422
2 10  9 0.5 0.5 18.71387
3  5  3 0.5 0.5 23.65319
4  3  5 0.5 0.5 22.35204
5  5  3 0.5 0.5 16.38019
6 10  6 0.5 0.5 17.63494
```

The code of function `sim_multi_mem` can be found in the online supplement of the present paper. For reasons of better illustration, students changing schools appear in the first rows of the data. Data of students being only at a single school looks as follows:

```
R> data_mm[101:106, ]
```

```
    s1 s2  w1  w2         y
101  2  2 0.5 0.5 27.247851
```

```
102  9  9 0.5 0.5 24.041427
103  4  4 0.5 0.5 12.575001
104  2  2 0.5 0.5 21.203644
105  4  4 0.5 0.5 12.856166
106  4  4 0.5 0.5  9.740174
```

Thus, school variables are identical, but we still have to specify both in order to pass the data appropriately. Incorporating no other predictors into the model for simplicity, a multimembership model is specified as

```
R> fit_mm <- brm(y ~ 1 + (1 | mm(s1, s2)), data = data_mm)
```

The only new syntax element is that multiple grouping factors (`s1` and `s2`) are wrapped in `mm`. Everything else remains exactly the same. Note that we did not specify the relative weights of schools for each student and thus, by default, equal weights are assumed.

```
R> summary(fit_mm)
```

```
 Family: gaussian (identity)
Formula: y ~ 1 + (1 | mm(s1, s2))
   Data: data_mm (Number of observations: 1000)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000
   WAIC: Not computed

Group-Level Effects:
~mms1s2 (Number of levels: 10)
              Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)     2.76      0.82     1.69     4.74        682 1.01

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept       19      0.93    17.06     20.8        610    1

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma     3.58      0.08     3.43     3.75       2117    1

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

One important way to evaluate model fit we did not cover so far in the present paper are graphical posterior-predictive checks Gelman *et al.* (2014). In **brms**, a wide variety of posterior-predictive checks are available using the **bayesplot** package (Gabry 2016) at the backend, which is an incredibly powerful tool for visualizing MCMC output[5].

[5] I hope that in the future, many Bayesian R packages (particularly those fitted via Stan) will adopt **bayesplot** to align graphical output and ease comparison between results obtained with different packages.

```
R> pp_check(fit_mm)
```

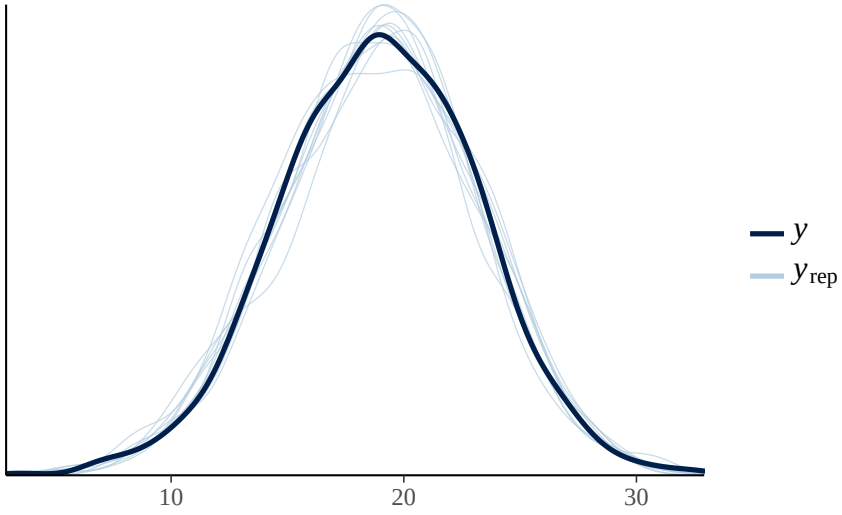

Figure 5: Density overlay plot of the `fit_mm1` model.

In Figure 5, the dark blue line represents the density of the observed response, while each light blue line represents draws from the posterior-predictive distribution, that is the distribution of the response as predicted by the model. Similar densities of observed and model-implied responses give a first indication of the appropriateness of the model, although two models with similar posterior-predictive distribution might perform differently when it comes to actual prediction of new data. In the present case, we see that densities are highly similar, which is, however, not overly surprising given that we fitted the exact same model to the data that was used to generate them. There is a continuously increasing number of posterior-predictive checks to apply via `pp_check`, and we recommend looking at `help("PPC-overview")` for an overview. To get a sense of bad model fit as detected via posterior-predictive checks, simply try modeling skewed data (e.g., response or survival times) using a normal model. We would hope that this display will jar researchers out of using normal distributions when they are not suitable.

With regard to the assumptions made in the above example, it is unlikely that all children who change schools stay in both schools equally long. To relax this assumption, we have to specify weights. First, we amend the simulated data to contain non-equal weights for students changing schools. For all other students, weighting does of course not matter as they stay in the same school anyway.

```
R> data_mm[1:100, "w1"] <- runif(100, 0, 1)
R> data_mm[1:100, "w2"] <- 1 - data_mm[1:100, "w1"]
R> head(data_mm)
```

```
  s1 s2        w1         w2        y
1  8  9 0.3403258 0.65967423 16.27422
2 10  9 0.1771435 0.82285652 18.71387
```

```
3  5  3 0.9059811 0.09401892 23.65319
4  3  5 0.4432007 0.55679930 22.35204
5  5  3 0.8052026 0.19479738 16.38019
6 10  6 0.5610243 0.43897567 17.63494
```

Incorporating these weights into the model is straight forward.

```
R> fit_mm2 <- brm(y ~ 1 + (1 | mm(s1, s2, weights = cbind(w1, w2))),
R>                data = data_mm)
```

The summary output is similar to the previous, so we do not show it here. The second assumption that students change schools only once a year, may also easily be relaxed by providing more than two grouping factors, say, `mm(s1, s2, s3)`.

5. Conclusion

The present paper is meant to introduce users to the flexibility of the distributional regression approach and corresponding formula syntax as implemented in **brms** and fitted with Stan behind the scenes. Only a subset of modeling options were discussed in detail, which ensured the paper was not too broad. For some of the more basic models that **brms** can fit, see Bürkner (in press). Many more examples can be found in the growing number of vignettes accompanying the package (see `vignette(package = "brms")` for an overview).

To date, **brms** is already one of the most flexible R packages when it comes to regression modeling. However, for the future, there are quite a few more features that I am planning to implement (see https://github.com/paul-buerkner/brms/issues for the current list of issues). In addition to smaller, incremental updates, I have three specific features in mind: extended multivariate models, extended autocorrelation structures, and missing value imputation (in order of current importance). I receive ideas and suggestions from users almost every day – for which I am always grateful – and so the list of features that will be implemented in the proceeding versions of **brms** will continue to grow.

Acknowledgments

First of all, I would like to thank the Stan Development Team for creating the probabilistic programming language Stan, which is an incredibly powerful and flexible tool for performing full Bayesian inference. Without it, **brms** could not fit a single model. Furthermore, I want to thank Donald Williams and Ruben Arslan for valuable comments on earlier versions of the paper. I also would like to thank the many users who reported bugs or had ideas for new features, thus helping to continuously improve **brms**.

References

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using **lme4**." *Journal of Statistical Software*, **67**(1), 1–48.

Betancourt M (2017). "A Conceptual Introduction to Hamiltonian Monte Carlo." *arXiv preprint*. URL https://arxiv.org/pdf/1701.02434.pdf.

Betancourt M, Byrne S, Livingstone S, Girolami M (2014). "The Geometric Foundations of Hamiltonian Monte Carlo." *arXiv preprint arXiv:1410.5110*.

Bürkner PC (in press). "**brms**: An R Package for Bayesian Multilevel Models using **Stan**." *Journal of Statistical Software*.

Brown H, Prescott R (2015). *Applied Mixed Models in Medicine*. John Wiley & Sons.

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P, Ridell A (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*.

Demidenko E (2013). *Mixed Models: Theory and Applications with R*. John Wiley & Sons.

Duane S, Kennedy AD, Pendleton BJ, Roweth D (1987). "Hybrid Monte Carlo." *Physics Letters B*, **195**(2), 216–222.

Gabry J (2016). ***bayesplot****: Plotting for Bayesian Models*. URL http://mc-stan.org/.

Gelman A, Carlin JB, Stern HS, Rubin DB (2014). *Bayesian Data Analysis*, volume 2. Taylor & Francis.

Gelman A, Hill J (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Hadfield JD (2010). "MCMC Methods for Multi-Response Generalized Linear Mixed Models: the **MCMCglmm** R Package." *Journal of Statistical Software*, **33**(2), 1–22.

Hastie TJ, Tibshirani RJ (1990). *Generalized Additive Models*, volume 43. CRC Press.

Hoffman MD, Gelman A (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *The Journal of Machine Learning Research*, **15**(1), 1593–1623.

Lindstrom MJ, Bates DM (1990). "Nonlinear Mixed Effects Models for Repeated Measures Data." *Biometrics*, pp. 673–687.

Neal RM (2011). *Handbook of Markov Chain Monte Carlo*, volume 2, chapter MCMC Using Hamiltonian Dynamics. CRC Press.

Pinheiro J, Bates D (2006). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlage Science & Business Media.

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2016). ***nlme****: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-124, URL http://CRAN.R-project.org/package=nlme.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rasmussen CE, Williams CKI (2006). "Gaussian processes for machine learning."

Rigby RA, Stasinopoulos DM (2005). "Generalized Additive Models for Location, Scale and Shape." *Journal of the Royal Statistical Society C (Applied Statistics)*, **54**(3), 507–554.

Singmann H, Bolker B, Westfall J (2015). ***afex**: Analysis of Factorial Experiments*. R package version 0.15-2, URL https://CRAN.R-project.org/package=afex.

Stan Development Team (2017a). *Stan: A C++ Library for Probability and Sampling, Version 2.14.0*. URL http://mc-stan.org/.

Stan Development Team (2017b). *Stan Modeling Language: User's Guide and Reference Manual*. URL http://mc-stan.org/manual.html.

Vehtari A, Gelman A, Gabry J (2016a). ***loo**: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*. R package version 1.0.0, URL https://github.com/stan-dev/loo.

Vehtari A, Gelman A, Gabry J (2016b). "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC." *Statistics and Computing*, pp. 1–20.

Westfall J, Yarkoni T (2016). "Statistically Controlling for Confounding Constructs is Harder than You Think." *PloS one*, **11**(3), e0152719.

Wood SN (2004). "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models." *Journal of the American Statistical Association*, **99**(467), 673–686.

Wood SN (2011). "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(1), 3–36.

Affiliation:

Paul-Christian Bürkner
Department of Statistics
Faculty of Psychology
University of Münster
48149, Münster
E-mail: paul.buerkner@wwu.de
URL: http://www.uni-muenster.de/PsyIFP/AEHolling/personen/buerkner.html