brms: An R Package for Bayesian Multilevel Models using Stan

Paul-Christian Bürkner
University of Münster

Abstract

The **brms** package implements Bayesian multilevel models in R using the probabilistic programming language **Stan**. A wide range of distributions and link functions are supported, allowing users to fit – among others – linear, robust linear, binomial, Poisson, survival, response times, ordinal, quantile, zero-inflated, hurdle, and even non-linear models all in a multilevel context. Further modeling options include autocorrelation of the response variable, user defined covariance structures, censored data, as well as meta-analytic standard errors. Prior specifications are flexible and explicitly encourage users to apply prior distributions that actually reflect their beliefs. In addition, model fit can easily be assessed and compared with the Watanabe-Akaike-Information Criterion and leave-one-out cross-validation. If you use the software, please cite this article as published in the Journal of Statistic Software Bürkner (in press).

Keywords: Bayesian inference, multilevel model, ordinal data, MCMC, **Stan**, R.

1. Introduction

Multilevel models (MLMs) offer a great flexibility for researchers across sciences (Brown and Prescott 2015; Demidenko 2013; Gelman and Hill 2006; Pinheiro and Bates 2006). They allow the 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 2015b) have been developed to fit MLMs. Possibly the most widely known package in this area is **lme4** (Bates, Mächler, Bolker, and Walker 2015), which uses maximum likelihood or restricted maximum likelihood methods for model fitting. Although alternative Bayesian methods have several advantages over frequentist approaches (e.g., the possibility of explicitly incorporating prior knowledge about parameters into the model), their practical use was limited for a long time because the posterior distributions of more complex models (such as MLMs) could not be found analytically. Markov chain Monte Carlo (MCMC) algorithms allowing to draw random samples from the posterior were not available or too time-consuming. In the last few decades, however, this has changed with the development of new algorithms and the rapid increase of general computing power. Today, several software packages implement these techniques, for instance **WinBugs** (Lunn, Thomas, Best, and Spiegelhalter 2000; Spiegelhalter, Thomas, Best, and Lunn 2003), **OpenBugs** (Spiegelhalter, Thomas, Best, and Lunn 2007), **JAGS** (Plummer 2013), **MCMCglmm** (Hadfield 2010) and **Stan** (Stan Development Team 2016a; Carpenter, Gelman, Hoffman, Lee, Goodrich, Betancour, Brubaker, Guo, Li, and Ridell 2015) to men-

tion only a few. With the exception of the latter, all of these programs are primarily using combinations of Metropolis-Hastings updates (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953; Hastings 1970) and Gibbs-sampling (Geman and Geman 1984; Gelfand and Smith 1990), sometimes also coupled with slice-sampling (Damien, Wakefield, and Walker 1999; Neal 2003). One of the main problems of these algorithms is their rather slow convergence for high-dimensional models with correlated parameters (Neal 2011; Hoffman and Gelman 2014; Gelman, Carlin, Stern, and Rubin 2014). Furthermore, Gibbs-sampling requires priors to be conjugate to the likelihood of parameters in order to work efficiently (Gelman *et al.* 2014), thus reducing the freedom of the researcher in choosing a prior that reflects his or her beliefs. In contrast, **Stan** implements Hamiltonian Monte Carlo (Duane, Kennedy, Pendleton, and Roweth 1987; Neal 2011) and its extension, the No-U-Turn Sampler (NUTS) (Hoffman and Gelman 2014). These algorithms converge much more quickly especially for high-dimensional models regardless of whether the priors are conjugate or not (Hoffman and Gelman 2014).

Similar to software packages like **WinBugs**, **Stan** comes with its own programming language, allowing for great modeling flexibility (cf., Stan Development Team 2016b; Carpenter *et al.* 2015). Many researchers may still hesitate 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 package **brms**, presented in this paper, aims at closing this gap (at least for MLMs) allowing the user to benefit from the merits of **Stan** only by using simple, **lme4**-like formula syntax. **brms** supports a wide range of distributions and link functions, allows for multiple grouping factors each with multiple group-level effects, autocorrelation of the response variable, user defined covariance structures, as well as flexible and explicit prior specifications.

The purpose of the present article is to provide a general overview of the **brms** package (version 0.10.0). We begin by explaining the underlying structure of MLMs. Next, the software is introduced in detail using recurrence times of infection in kidney patients (McGilchrist and Aisbett 1991) and ratings of inhaler instructions (Ezzet and Whitehead 1991) as examples. We end by comparing **brms** to other R packages implementing MLMs and describe future plans for extending the package.

2. Model description

The core of every MLM is the prediction of the response y through the linear combination η of predictors transformed by the inverse link function f assuming a certain distribution D for y. We write

$$y_i \sim D(f(\eta_i), \theta)$$

to stress the dependency on the i^{th} data point. In many R packages, D is also called the 'family' and we will use this term in the following. The parameter θ describes additional family specific parameters that typically do not vary across data points, such as the standard deviation σ in normal models or the shape α in Gamma or negative binomial models. The linear predictor can generally be written as

$$\eta = \mathbf{X}\beta + \mathbf{Z}u$$

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 response y as well as $\mathbf{X}$ and $\mathbf{Z}$ make

up the data, whereas β, u, and θ are the model parameters being estimated. The coefficients β and u may be more commonly known as fixed and random effects. However, we avoid these terms in the present paper following the recommendations of Gelman and Hill (2006), as they are not used unambiguously in the literature. Also, we want to make explicit that u is a model parameter in the same manner as β so that uncertainty in its estimates can be naturally evaluated. In fact, this is an important advantage of Bayesian MCMC methods as compared to maximum likelihood approaches, which do not treat u as a parameter, but assume that it is part of the error term instead (cf., Fox and Weisberg, 2011).

Except for linear models, we do not incorporate an additional error term for every observation by default. If desired, such an error term can always be modeled using a grouping factor with as many levels as observations in the data.

2.1. Prior distributions

Regression parameters at population-level

In **brms**, population-level parameters are not restricted to have normal priors. Instead, every parameter can have every one-dimensional prior implemented in **Stan**, for instance uniform, Cauchy or even Gamma priors. As a negative side effect of this flexibility, correlations between them cannot be modeled as parameters. If desired, point estimates of the correlations can be obtained after sampling has been done. By default, population level parameters have an improper flat prior over the reals.

Regression parameters at group-level

The group-level parameters u are assumed to come from a multivariate normal distribution with mean zero and unknown covariance matrix $\mathbf{\Sigma}$:

$$u \sim N(0, \mathbf{\Sigma})$$

As is generally the case, covariances between group-level parameters of different grouping factors are assumed to be zero. This implies that $\mathbf{Z}$ and u can be split up into several matrices $\mathbf{Z_k}$ and parameter vectors u_k, where k indexes grouping factors, so that the model can be simplified to

$$u_k \sim N(0, \mathbf{\Sigma_k})$$

Usually, but not always, we can also assume group-level parameters associated with different levels (indexed by j) of the same grouping factor to be independent leading to

$$u_{kj} \sim N(0, \mathbf{V_k})$$

The covariance matrices $\mathbf{V_k}$ are modeled as parameters. In most packages, an Inverse-Wishart distribution is used as a prior for $\mathbf{V_k}$. This is mostly because its conjugacy leads to good properties of Gibbs-Samplers (Gelman *et al.* 2014). However, there are good arguments against the Inverse-Wishart prior (Natarajan and Kass 2000; Kass and Natarajan 2006). The NUTS-Sampler implemented in **Stan** does not require priors to be conjugate. This advantage is utilized in **brms**: $\mathbf{V_k}$ is parameterized in terms of a correlation matrix $\mathbf{\Omega_k}$ and a vector of standard deviations σ_k through

$$\mathbf{V_k} = \mathbf{D}(\sigma_k)\mathbf{\Omega_k}\mathbf{D}(\sigma_k)$$

where $\mathbf{D}(\sigma_k)$ denotes the diagonal matrix with diagonal elements σ_k. Priors are then specified for the parameters on the right hand side of the equation. For $\mathbf{\Omega_k}$, we use the LKJ-Correlation prior with parameter $\zeta > 0$ by Lewandowski, Kurowicka, and Joe (2009)[1]:

$$\mathbf{\Omega_k} \sim \text{LKJ}(\zeta)$$

The expected value of the LKJ-prior is the identity matrix (implying correlations of zero) for any positive value of ζ, which can be interpreted like the shape parameter of a symmetric beta distribution (Stan Development Team 2016b). If $\zeta = 1$ (the default in **brms**) the density is uniform over correlation matrices of the respective dimension. If $\zeta > 1$, the identity matrix is the mode of the prior, with a sharper peak in the density for larger values of ζ. If $0 < \zeta < 1$ the prior is U-shaped having a trough at the identity matrix, which leads to higher probabilities for non-zero correlations. For every element of σ_k, any prior can be applied that is defined on the non-negative reals only. As default in **brms**, we use a half Student-t prior with 3 degrees of freedom. This prior often leads to better convergence of the models than a half Cauchy prior, while still being relatively weakly informative.

Sometimes – for instance when modeling pedigrees – different levels of the same grouping factor cannot be assumed to be independent. In this case, the covariance matrix of u_k becomes

$$\mathbf{\Sigma_k} = \mathbf{V_k} \otimes \mathbf{A_k}$$

where $\mathbf{A_k}$ is the known covariance matrix between levels and $\otimes$ is the Kronecker product.

Family specific parameters

For some families, additional parameters need to be estimated. In the current section, we only name the most important ones. Normal and Student's distributions need the parameter σ to account for residual error variance. By default, σ has a half Cauchy prior with a scale parameter that depends on the standard deviation of the response variable to remain only weakly informative regardless of response variable's scaling. Furthermore, Student's distributions needs the parameter ν representing the degrees of freedom. By default, ν has a wide gamma prior as proposed by Juárez and Steel (2010). Gamma, Weibull, and negative binomial distributions need the shape parameter α that also has a wide gamma prior by default.

3. Parameter estimation

The **brms** package does not fit models itself but uses **Stan** on the back-end. Accordingly, all samplers implemented in **Stan** can be used to fit **brms** models. Currently, these are the static Hamiltonian Monte-Carlo (HMC) Sampler sometimes also referred to as Hybrid Monte-Carlo (Neal 2011, 2003; Duane *et al.* 1987) and its extension the No-U-Turn Sampler (NUTS) by Hoffman and Gelman (2014). HMC-like algorithms produce samples that are much less autocorrelated than those of other samplers such as the random-walk Metropolis algorithm (Hoffman and Gelman 2014; Creutz 1988). The main drawback of this increased efficiency is the need to calculate the gradient of the log-posterior, which can be automated

[1]Internally, the Cholesky factor of the correlation matrix is used, as it is more efficient and numerically stable.

using algorithmic differentiation (Griewank and Walther 2008) but is still a time-consuming process for more complex models. Thus, using HMC leads to higher quality samples but takes more time per sample than other algorithms typically applied. Another drawback of HMC is the need to pre-specify at least two parameters, which are both critical for the performance of HMC. The NUTS Sampler allows setting these parameters automatically thus eliminating the need for any hand-tuning, while still being at least as efficient as a well tuned HMC (Hoffman and Gelman 2014). For more details on the sampling algorithms applied in **Stan**, see the **Stan** user's manual (Stan Development Team 2016b) as well as Hoffman and Gelman (2014).

In addition to the estimation of model parameters, **brms** allows drawing samples from the posterior predictive distribution as well as from the pointwise log-likelihood. Both can be used to assess model fit. The former allows a comparison between the actual response y and the response $\hat{y}$ predicted by the model. The pointwise log-likelihood can be used, among others, to calculate the Watanabe-Akaike information criterion (WAIC) proposed by Watanabe (2010) and leave-one-out cross-validation (LOO; Gelfand, Dey, and Chang 1992; Vehtari, Gelman, and Gabry 2015a; see also Ionides 2008) both allowing to compare different models applied to the same data (lower WAICs and LOOs indicate better model fit). The WAIC can be viewed as an improvement of the popular deviance information criterion (DIC), which has been criticized by several authors (Vehtari *et al.* 2015a; Plummer 2008; van der Linde 2005; see also the discussion at the end of the original DIC paper by Spiegelhalter, Best, Carlin, and Van Der Linde 2002) in part because of problems arising from fact that the DIC is only a point estimate. In **brms**, WAIC and LOO are implemented using the **loo** package (Vehtari, Gelman, and Gabry 2015b) also following the recommendations of Vehtari *et al.* (2015a).

4. Software

The **brms** package provides functions for fitting MLMs using **Stan** for full Bayesian inference. To install the latest release version of **brms** from CRAN, type `install.packages("brms")` within R. The current developmental version can be downloaded from GitHub via

```
devtools::install_github("paul-buerkner/brms")
```

Additionally, a C++ compiler is required. This is because **brms** internally creates **Stan** code, which is translated to C++ and compiled afterwards. The program **Rtools** (R Core Team 2015a) comes with a C++ compiler for Windows[2]. On OS X, one should use **Xcode** (Apple Inc. 2015) from the App Store. To check whether the compiler can be called within R, run `system("g++ -v")` when using **Rtools** or `system("clang++ -v")` when using **Xcode**. If no warning occurs and a few lines of difficult to read system code are printed out, the compiler should work correctly. For more detailed instructions on how to get the compilers running, see the prerequisites section on https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started.

Models are fitted in **brms** using the following procedure, which is also summarized in Figure 1. First, the user specifies the model using the `brm` function in a way typical for most model fitting R functions, that is by defining `formula`, `data`, and `family`, as well as some other optional arguments. Second, this information is processed and the `make_stancode` and `make_standata`

[2]During the installation process, there is an option to change the system `PATH`. Please make sure to check this options, because otherwise **Rtools** will not be available within R.

functions are called. The former generates the model code in **Stan** language and the latter prepares the data for use in **Stan**. These two are the mandatory parts of every **Stan** model and without **brms**, users would have to specify them themselves. Third, **Stan** code and data as well as additional arguments (such as the number of iterations and chains) are passed to functions of the **rstan** package (the R interface of **Stan**; Stan Development Team, 2016a). Fourth, the model is fitted by **Stan** after translating and compiling it in C++. Fifth, after the model has been fitted and returned by **rstan**, the fitted model object is post-processed in **brms** among others by renaming the model parameters to be understood by the user. Sixth, the results can be investigated in R using various methods such as `summary`, `plot`, or `predict` (for a complete list of methods type `methods(class = "brmsfit")`).

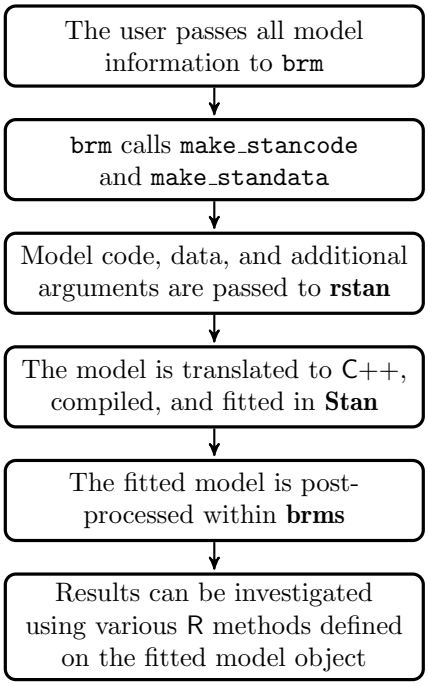

Figure 1: High level description of the model fitting procedure used in **brms**.

4.1. A worked example

In the following, we use an example about the recurrence time of an infection in kidney patients initially published by McGilchrist and Aisbett (1991). The data set consists of 76 entries of 7 variables:

```
R> library("brms")
R> data("kidney")
R> head(kidney, n = 3)
```

```
  time censored patient recur age    sex disease
1    8        0       1     1  28   male   other
2   23        0       2     1  48 female      GN
3   22        0       3     1  32   male   other
```

Variable `time` represents the recurrence time of the infection, `censored` indicates if `time` is right censored (`1`) or not censored (`0`), variable `patient` is the patient id, and `recur` indicates if it is the first or second recurrence in that patient. Finally, variables `age`, `sex`, and `disease` make up the predictors.

4.2. Fitting models with brms

The core of the **brms** package is the `brm` function and we will explain its argument structure using the example above. Suppose we want to predict the (possibly censored) recurrence time using a log-normal model, in which the intercept as well as the effect of `age` is nested within patients. Then, we may use the following code:

```
fit1 <- brm(formula = time | cens(censored) ~ age * sex + disease
            + (1 + age|patient),
            data = kidney, family = lognormal(),
            prior = c(set_prior("normal(0,5)", class = "b"),
                      set_prior("cauchy(0,2)", class = "sd"),
                      set_prior("lkj(2)", class = "cor")),
            warmup = 1000, iter = 2000, chains = 4,
            control = list(adapt_delta = 0.95))
```

4.3. `formula`: Information on the response and predictors

Without doubt, `formula` is the most complicated argument, as it contains information on the response variable as well as on predictors at different levels of the model. Everything before the $\sim$ sign relates to the response part of `formula`. In the usual and most simple case, this is just one variable name (e.g., `time`). However, to incorporate additional information about the response, one can add one or more terms of the form `| fun(variable)`. `fun` may be one of a few functions defined internally in **brms** and `variable` corresponds to a variable in the data set supplied by the user. In this example, `cens` makes up the internal function that handles censored data, and `censored` is the variable that contains information on the censoring. Other available functions in this context are `weights` and `disp` to allow different sorts of weighting, `se` to specify known standard errors primarily for meta-analysis, `trunc` to define truncation boundaries, `trials` for binomial models[3], and `cat` to specify the number of categories for ordinal models.

Everything on the right side of $\sim$ specifies predictors. Here, the syntax exactly matches that of **lme4**. For both, population-level and group-level terms, the `+` is used to separate different effects from each other. Group-level terms are of the form `(coefs | group)`, where `coefs` contains one or more variables whose effects are assumed to vary with the levels of the grouping factor given in `group`. Multiple grouping factors each with multiple group-level coefficients are possible. In the present example, only one group-level term is specified in which `1 + age` are the coefficients varying with the grouping factor `patient`. This implies that the intercept of the model as well as the effect of age is supposed to vary between patients. By default, group-level coefficients within a grouping factor are assumed to be correlated. Correlations

[3] In functions such as `glm` or `glmer`, the binomial response is typically passed as `cbind(success, failure)`. In **brms**, the equivalent syntax is `success | trials(success + failure)`.

can be set to zero by using the `(coefs || group)` syntax[4]. Everything on the right side of `formula` that is not recognized as part of a group-level term is treated as a population-level effect. In this example, the population-level effects are `age`, `sex`, and `disease`.

4.4. `family`: Distribution of the response variable

Argument `family` should usually be a family function, a call to a family function or a character string naming the family. If not otherwise specified, default link functions are applied. **brms** comes with a large variety of families. Linear and robust linear regression can be performed using the `gaussian` or `student` family combined with the `identity` link. For dichotomous and categorical data, families `bernoulli`, `binomial`, and `categorical` combined with the `logit` link, by default, are perfectly suited. Families `poisson`, `negbinomial`, and `geometric` allow for modeling count data. Families `lognormal`, `Gamma`, `exponential`, and `weibull` can be used (among others) for survival regression. Ordinal regression can be performed using the families `cumulative`, `cratio`, `sratio`, and `acat`. Finally, families `zero_inflated_poisson`, `zero_inflated_negbinomial`, `zero_inflated_binomial`, `zero_inflated_beta`, `hurdle_poisson`, `hurdle_negbinomial`, and `hurdle_gamma` can be used to adequately model excess zeros in the response. In our example, we use `family = lognormal()` implying a log-normal "survival" model for the response variable `time`.

4.5. `prior`: Prior distributions of model parameters

Every population-level effect has its corresponding regression parameter. These parameters are named as `b_<coef>`, where `<coef>` represents the name of the corresponding population-level effect. The default prior is an improper flat prior over the reals. Suppose, for instance, that we want to set a normal prior with mean `0` and standard deviation `10` on the effect of `age` and a Cauchy prior with location `1` and scale `2` on `sexfemale`[5]. Then, we may write

```
prior <- c(set_prior("normal(0,10)", class = "b", coef = "age"),
           set_prior("cauchy(1,2)", class = "b", coef = "sexfemale"))
```

To put the same prior (e.g., a normal prior) on all population-level effects at once, we may write as a shortcut `set_prior("normal(0,10)", class = "b")`. This also leads to faster sampling, because priors can be vectorized in this case. Note that we could also omit the `class` argument for population-level effects, as it is the default class in `set_prior`.

A special shrinkage prior to be applied on population-level effects is the horseshoe prior (Carvalho, Polson, and Scott 2009, 2010). It is symmetric around zero with fat tails and an infinitely large spike at zero. This makes it ideal for sparse models that have many regression coefficients, although only a minority of them is non-zero. The horseshoe prior

[4] In contrast to **lme4**, the `||` operator in **brms** splits up the design matrix computed from `coefs` instead of decomposing `coefs` in its terms. This implies that columns of the design matrix originating from the same factor are also assumed to be uncorrelated, whereas **lme4** estimates the correlations in this case. For a way to achieve **brms**-like behavior with **lme4**, see the `mixed` function of the **afex** package by Singmann, Bolker, and Westfall (2015).

[5] When factors are used as predictors, parameter names will depend on the factor levels. To get an overview of all parameters and parameter classes for which priors can be specified, use function `get_prior`. For the present example, `get_prior(time | cens(censored) ~ age * sex + disease + (1 + age|patient), data = kidney, family = lognormal())` does the desired.

can be applied on all population-level effects at once (excluding the intercept) by using `set_prior("horseshoe(1)")`. The 1 implies that the Student-t prior of the local shrinkage parameters has 1 degrees of freedom. In **brms** it is possible to increase the degrees of freedom (which will often improve convergence), although the prior no longer resembles a horseshoe in this case[6]. For more details see Carvalho *et al.* (2009, 2010).

Each group-level effect of each grouping factor has a standard deviation parameter, which is restricted to be non-negative and, by default, has a half Student-t prior with 3 degrees of freedom and a scale parameter that is minimally 10. For non-ordinal models, **brms** tries to evaluate if the scale is large enough to be considered only weakly informative for the model at hand by comparing it with the standard deviation of the response after applying the link function. If this is not the case, it will increase the scale based on the aforementioned standard deviation[7]. **Stan** implicitly defines a half Student-t prior by using a Student-t prior on a restricted parameter (Stan Development Team 2016b). For other reasonable priors on standard deviations see Gelman (2006). In **brms**, standard deviation parameters are named as `sd_<group>_<coef>` so that `sd_patient_Intercept` and `sd_patient_age` are the parameter names in the example. If desired, it is possible to set a different prior on each parameter, but statements such as `set_prior("student_t(3,0,5)", class = "sd", group = "patient")` or even `set_prior("student_t(3,0,5)", class = "sd")` may also be used and are again faster because of vectorization.

If there is more than one group-level effect per grouping factor, correlations between group-level effects are estimated. As mentioned in Section 2, the LKJ-Correlation prior with parameter $\zeta > 0$ (Lewandowski *et al.* 2009) is used for this purpose. In **brms**, this prior is abbreviated as `"lkj(zeta)"` and correlation matrix parameters are named as `cor_<group>`, (e.g., `cor_patient`), so that `set_prior("lkj(2)", class = "cor", group = "patient")` is a valid statement. To set the same prior on every correlation matrix in the model, `set_prior("lkj(2)", class = "cor")` is also allowed, but does not come with any efficiency increases.

Other model parameters such as the residual standard deviation `sigma` in normal models or the `shape` in Gamma models have their priors defined in the same way, where each of them is treated as having its own parameter class. A complete overview on possible prior distributions is given in the **Stan** user's manual (Stan Development Team 2016b). Note that **brms** does not thoroughly check if the priors are written in correct **Stan** language. Instead, **Stan** will check their syntactical correctness when the model is parsed to C++ and return an error if they are not. This, however, does not imply that priors are always meaningful if they are accepted by **Stan**. Although **brms** tries to find common problems (e.g., setting bounded priors on unbounded parameters), there is no guarantee that the defined priors are reasonable for the model.

4.6. `control`: Adjusting the sampling behavior of Stan

In addition to choosing the number of iterations, warmup samples, and chains, users can con-

[6] This class of priors is often referred to as hierarchical shrinkage family, which contains the original horseshoe prior as a special case.

[7] Changing priors based on the data is not truly Bayesian and might rightly be criticized. However, it helps avoiding the problem of too informative default priors without always forcing users to define their own priors. The latter would also be problematic as not all users can be expected to be well educated Bayesians and reasonable default priors will help them a lot in using Bayesian methods.

trol the behavior of the NUTS sampler by using the `control` argument. The most important reason to use `control` is to decrease (or eliminate at best) the number of divergent transitions that cause a bias in the obtained posterior samples. Whenever you see the warning `"There were x divergent transitions after warmup."`, you should really think about increasing `adapt_delta`. To do this, write `control = list(adapt_delta = <x>)`, where `<x>` should usually be a value between `0.8` (current default) and `1`. Increasing `adapt_delta` will slow down the sampler but will decrease the number of divergent transitions threatening the validity of your posterior samples.

Another problem arises when the depth of the tree being evaluated in each iteration is exceeded. This is less common than having divergent transitions, but may also bias the posterior samples. When it happens, **Stan** will throw out a warning suggesting to increase `max_treedepth`, which can be accomplished by writing `control = list(max_treedepth = <x>)` with a positive integer `<x>` that should usually be larger than the current default of `10`.

4.7. Analyzing the results

The example model `fit1` is fitted using 4 chains, each with 2000 iterations of which the first 1000 are warmup to calibrate the sampler, leading to a total of 4000 posterior samples[8]. For researchers familiar with Gibbs or Metropolis-Hastings sampling, this number may seem far too small to achieve good convergence and reasonable results, especially for multilevel models. However, as **brms** utilizes the NUTS sampler (Hoffman and Gelman 2014) implemented in **Stan**, even complex models can often be fitted with not more than a few thousand samples. Of course, every iteration is more computationally intensive and time-consuming than the iterations of other algorithms, but the quality of the samples (i.e., the effective sample size per iteration) is usually higher.

After the posterior samples have been computed, the `brm` function returns an R object, containing (among others) the fully commented model code in **Stan** language, the data to fit the model, and the posterior samples themselves. The model code and data for the present example can be extracted through `stancode(fit1)` and `standata(fit1)` respectively[9]. A model summary is readily available using

```
R> summary(fit1, waic = TRUE)

 Family: lognormal (identity)
Formula: time | cens(censored) ~ age * sex + disease + (1 + age | patient)
   Data: kidney (Number of observations: 76)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000
   WAIC: 673.51

Group-Level Effects:
~patient (Number of levels: 38)
                  Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
```

[8] To save time, chains may also run in parallel when using argument `cluster`.

[9] Both model code and data may be amended and used to fit new models. That way, **brms** can also serve as a good starting point in building more complicated models in **Stan**, directly.

```
sd(Intercept)          0.40      0.28      0.01      1.01        1731    1
sd(age)                0.01      0.01      0.00      0.02        1137    1
cor(Intercept,age)    -0.13      0.46     -0.88      0.76        3159    1

Population-Level Effects:
             Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept        2.73      0.96     0.82     4.68        2139    1
age              0.01      0.02    -0.03     0.06        1614    1
sexfemale        2.42      1.13     0.15     4.64        2065    1
diseaseGN       -0.40      0.53    -1.45     0.64        2664    1
diseaseAN       -0.52      0.50    -1.48     0.48        2713    1
diseasePKD       0.60      0.74    -0.86     2.02        2968    1
age:sexfemale   -0.02      0.03    -0.07     0.03        1956    1

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma     1.15      0.13     0.91     1.44        4000    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).
```

On the top of the output, some general information on the model is given, such as family, formula, number of iterations and chains, as well as the WAIC. Next, group-level effects are displayed separately for each grouping factor in terms of standard deviations and correlations between group-level effects. On the bottom of the output, population-level effects are displayed. If incorporated, autocorrelation and family specific parameters (e.g., the residual standard deviation `sigma`) are also given.

In general, every parameter is summarized using the mean (`Estimate`) and the standard deviation (`Est.Error`) of the posterior distribution as well as two-sided 95% Credible intervals (`l-95% CI` and `u-95% CI`) based on quantiles. The `Eff.Sample` value is an estimation of the effective sample size; that is the number of independent samples from the posterior distribution that would be expected to yield the same standard error of the posterior mean as is obtained from the dependent samples returned by the MCMC algorithm. The `Rhat` value provides information on the convergence of the algorithm (cf., Gelman and Rubin, 1992). If `Rhat` is considerably greater than 1 (i.e., > 1.1), the chains have not yet converged and it is necessary to run more iterations and/or set stronger priors.

To visually investigate the chains as well as the posterior distribution, the `plot` method can be used (see Figure 2). An even more detailed investigation can be achieved by applying the **shinystan** package (Gabry 2015) through method `launch_shiny`. With respect to the above summary, `sexfemale` seems to be the only population-level effect with considerable influence on the response. Because the mean of `sexfemale` is positive, the model predicts longer periods without an infection for females than for males. Effects of population-level predictors can also be visualized with the `marginal_effects` method (see Figure 3).

Looking at the group-level effects, the standard deviation parameter of `age` is suspiciously small. To test whether it is smaller than the standard deviation parameter of `Intercept`, we

apply the `hypothesis` method:

```
R> hypothesis(fit1, "Intercept - age > 0", class = "sd", group = "patient")

Hypothesis Tests for class sd_patient:
                 Estimate Est.Error l-95% CI u-95% CI Evid.Ratio
Intercept-age > 0     0.39      0.27     0.03      Inf      67.97 *
---
'*': The expected value under the hypothesis lies outside the 95% CI.
```

The one-sided 95% credibility interval does not contain zero, thus indicating that the standard deviations differ from each other in the expected direction. In accordance with this finding, the `Evid.Ratio` shows that the hypothesis being tested (i.e., `Intercept - age > 0`) is about 68 times more likely than the alternative hypothesis `Intercept - age < 0`. It is important to note that this kind of comparison is not easily possible when applying frequentist methods, because in this case only point estimates are available for group-level standard deviations and correlations.

When looking at the correlation between both group-level effects, its distribution displayed in Figure 2 and the 95% credibility interval in the summary output appear to be rather wide. This indicates that there is not enough evidence in the data to reasonably estimate the correlation. Together, the small standard deviation of `age` and the uncertainty in the correlation raise the question if `age` should be modeled as a group specific term at all. To answer this question, we fit another model without this term:

```
R> fit2 <- update(fit1, formula. = ~ . - (1 + age|patient) + (1|patient))
```

A good way to compare both models is leave-one-out cross-validation (LOO)[10], which can be called in **brms** using

```
R> LOO(fit1, fit2)

              LOOIC    SE
fit1         675.45 45.18
fit2         674.17 45.06
fit1 - fit2   1.28  0.99
```

In the output, the LOO information criterion for each model as well as the difference of the LOOs each with its corresponding standard error is shown. Both LOO and WAIC are approximately normal if the number of observations is large so that the standard errors can be very helpful in evaluating differences in the information criteria. However, for small sample sizes, standard errors should be interpreted with care (Vehtari *et al.* 2015a). For the present example, it is immediately evident that both models have very similar fit, indicating that there is little benefit in adding group specific coefficients for `age`.

[10]The WAIC is an approximation of LOO that is faster and easier to compute. However, according to Vehtari *et al.* (2015a), LOO may be the preferred method to perform model comparisons.

4.8. Modeling ordinal data

In the following, we want to briefly discuss a second example to demonstrate the capabilities of **brms** in handling ordinal data. Ezzet and Whitehead (1991) analyze data from a two-treatment, two-period crossover trial to compare 2 inhalation devices for delivering the drug salbutamol in 286 asthma patients. Patients were asked to rate the clarity of leaflet instructions accompanying each device, using a four-point ordinal scale. Ratings are predicted by `treat` to indicate which of the two inhaler devices was used, `period` to indicate the time of administration, and `carry` to model possible carry over effects.

```
R> data("inhaler")
R> head(inhaler, n = 1)

  subject rating treat period carry
1       1      1   0.5    0.5     0
```

Typically, the ordinal response is assumed to originate from the categorization of a latent continuous variable. That is there are K latent thresholds (model intercepts), which partition the continuous scale into the $K + 1$ observable, ordered categories. Following this approach leads to the cumulative or graded-response model (Samejima 1969) for ordinal data implemented in many R packages. In **brms**, it is available via family `cumulative`. Fitting the cumulative model to the inhaler data, also incorporating an intercept varying by subjects, may look this:

```
fit3 <- brm(formula = rating ~ treat + period + carry + (1|subject),
            data = inhaler, family = cumulative)
```

While the support for ordinal data in most R packages ends here[11], **brms** allows changes to this basic model in at least three ways. First of all, three additional ordinal families are implemented. Families `sratio` (stopping ratio) and `cratio` (continuation ratio) are so called sequential models (Tutz 1990). Both are equivalent to each other for symmetric link functions such as `logit` but will differ for asymmetric ones such as `cloglog`. The fourth ordinal family is `acat` (adjacent category) also known as partial credits model (Masters 1982; Andrich 1978b). Second, restrictions to the thresholds can be applied. By default, thresholds are ordered for family `cumulative` or are completely free to vary for the other families. This is indicated by argument `threshold = "flexible"` (default) in `brm`. Using `threshold = "equidistant"` forces the distance between two adjacent thresholds to be the same, that is

$$\tau_k = \tau_1 + (k - 1)\delta$$

for thresholds τ_k and distance δ (see also Andrich 1978a; Andrich 1978b; Andersen 1977). Third, the assumption that predictors have constant effects across categories may be relaxed for non-cumulative ordinal models (Van Der Ark 2001; Tutz 2000) leading to category specific effects. For instance, variable `treat` may only have an impact on the decision between category 3 and 4, but not on the lower categories. Without using category specific effects, such a pattern would remain invisible.

[11] Exceptions known to us are the packages **ordinal** (Christensen 2015) and **VGAM** (Yee 2010). The former supports only cumulative models but with different modeling option for the thresholds. The latter supports all four ordinal families also implemented in **brms** as well as category specific effects but no group-specific effects.

To illustrate all three modeling options at once, we fit a (hardly theoretically justified) stopping ratio model with equidistant thresholds and category specific effects for variable `treat` on which we apply an informative prior.

```
fit4 <- brm(formula = rating ~ period + carry + cs(treat) + (1|subject),
            data = inhaler, family = sratio, threshold = "equidistant",
            prior = set_prior("normal(-1,2)", coef = "treat"))
```

Note that priors are defined on category specific effects in the same way as for other population-level effects. A model summary can be obtained in the same way as before:

```
R> summary(fit4, waic = TRUE)

 Family: sratio (logit)
Formula: rating ~ period + carry + cs(treat) + (1 | subject)
   Data: inhaler (Number of observations: 572)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000
   WAIC: 911.9

Group-Level Effects:
~subject (Number of levels: 286)
              Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)     1.05      0.23     0.56      1.5        648    1

Population-Level Effects:
             Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept[1]     0.72      0.13     0.48     0.99       2048    1
Intercept[2]     2.67      0.35     2.00     3.39        969    1
Intercept[3]     4.62      0.66     3.36     5.95       1037    1
period           0.25      0.18    -0.09     0.61       4000    1
carry           -0.26      0.22    -0.70     0.17       1874    1
treat[1]        -0.96      0.30    -1.56    -0.40       1385    1
treat[2]        -0.65      0.49    -1.60     0.27       4000    1
treat[3]        -2.65      1.21    -5.00    -0.29       4000    1

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
delta     1.95      0.32     1.33      2.6       1181    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).
```

Trace and density plots of the model parameters as produced by `plot(fit4)` can be found in Figure 4. We see that three intercepts (thresholds) and three effects of `treat` have been estimated, because a four-point scale was used for the ratings. The treatment effect seems to

be strongest between category 3 and 4. At the same time, however, the credible interval is also much larger. In fact, the intervals of all three effects of `treat` are highly overlapping, which indicates that there is not enough evidence in the data to support category specific effects. On the bottom of the output, parameter `delta` specifies the distance between two adjacent thresholds and indeed the intercepts differ from each other by the magnitude of `delta`.

5. Comparison between packages

Over the years, many R packages have been developed that implement MLMs, each being more or less general in their supported models. Comparing all of them to **brms** would be too extensive and barely helpful for the purpose of the present paper. Accordingly, we concentrate on a comparison with four packages. These are **lme4** (Bates *et al.* 2015) and **MCMCglmm** (Hadfield 2010), which are possibly the most general and widely applied R packages for MLMs, as well as **rstanarm** (Gabry and Goodrich 2016) and **rethinking** (McElreath 2016), which are both based on **Stan**. As opposed to the other packages, **rethinking** was primarily written for teaching purposes and requires the user to specify the full model explicitly using its own simplified **BUGS**-like syntax thus helping users to better understand the models that are fitted to their data.

Regarding model families, all five packages support the most common types such as linear and binomial models as well as Poisson models for count data. Currently, **brms** and **MCMCglmm** provide more flexibility when modeling categorical and ordinal data. In addition, **brms** supports robust linear regression using Student's distribution, which is also implemented on a GitHub branch of **rstanarm**. **MCMCglmm** allows fitting multinomial models that are currently not available in the other packages.

Generalizing classical MLMs, **brms** and **MCMCglmm** allow fiting zero-inflated and hurdle models dealing with excess zeros in the response. Furthermore, **brms** supports non-linear models similar to the **nlme** package (Pinheiro, Bates, DebRoy, Sarkar, and R Core Team 2016) providing great flexibility but also requiring more care to produce reasonable results. Another flexible model class are generalized additive mixed models (Hastie and Tibshirani 1990; Wood 2012; Zuur 2014), which can be fitted with **brms** and **rstanarm**.

In all five packages, there are quite a few additional modeling options. Variable link functions can be specified in all packages except for **MCMCglmm**, in which only one link is available per family. **MCMCglmm** generally supports multivariate responses using data in wide format, whereas **brms** currently only offers this option for families `gaussian` and `student`. It should be noted that it is always possible to transform data from wide to long format for compatibility with the other packages. Autocorrelation of the response can only be fitted in **brms**, which supports auto-regressive as well as moving-average effects. For ordinal models in **brms**, effects of predictors may vary across different levels of the response as explained in the inhaler example. A feature currently exclusive to **rethinking** is the possibility to impute missing values in the predictor variables.

Information criteria are available in all three packages. The advantage of WAIC and LOO implemented in **brms**, **rstanarm**, and **rethinking** is that their standard errors can be easily estimated to get a better sense of the uncertainty in the criteria. Comparing the prior options of the Bayesian packages, **brms** and **rethinking** offer a little more flexibility than **MCMCglmm** and **rstanarm**, as virtually any prior distribution can be applied on population-level effects

as well as on the standard deviations of group-level effects. In addition, we believe that the way priors are specified in **brms** and **rethinking** is more intuitive as it is directly evident what prior is actually applied. A more detailed comparison of the packages can be found in Table 1 and Table 2. To facilitate the understanding of the model formulation in **brms**, Table 3 shows **lme4** function calls to fit sample models along with the equivalent **brms** syntax.

So far the focus was only on capabilities. Another important topic is speed, especially for more complex models. Of course, **lme4** is usually much faster than the other packages as it uses maximum likelihood methods instead of MCMC algorithms, which are slower by design. To compare the efficiency of the four Bayesian packages, we fitted multilevel models on real data sets using the minimum effective sample size divided by sampling time as a measure of sampling efficiency. One should always aim at running multiple chains as one cannot be sure that a single chain really explores the whole posterior distribution. However, as **MCMCglmm** does not come with a built-in option to run multiple chains, we used only a single chain to fit the models after making sure that it leads to the same results as multiple chains. The R code allowing to replicate the results is available as supplemental material.

The first thing that becomes obvious when fitting the models is that **brms** and **rethinking** need to compile the C++ model before actually fitting it, because the **Stan** code being parsed to C++ is generated on the fly based on the user's input. Compilation takes about a half to one minute depending on the model complexity and computing power of the machine. This is not required by **rstanarm** and **MCMCglmm**, although the former is also based on **Stan**, as compilation takes place only once at installation time. While the latter approach saves the compilation time, the former is more flexible when it comes to model specification. For small and simple models, compilation time dominates the overall computation time, but for larger and more complex models, sampling will take several minutes or hours so that one minute more or less will not really matter, anymore. Accordingly, the following comparisons do not include the compilation time.

In models containing only group-specific intercepts, **MCMCglmm** is usually more efficient than the **Stan** packages. However, when also estimating group-specific slopes, **MCMCglmm** falls behind the other packages and quite often refuses to sample at all unless one carefully specifies informative priors. Note that these results are obtained by running only a single chain. For all three **Stan** packages, sampling efficiency can easily be increased by running multiple chains in parallel. Comparing the **Stan** packages to each other, **brms** is usually most efficient for models with group-specific terms, whereas **rstanarm** tends to be roughly 50% to 75% as efficient at least for the analyzed data sets. The efficiency of **rethinking** is more variable depending on the model formulation and data, sometimes being slightly ahead of the other two packages, but usually being considerably less efficient. Generally, **rethinking** loses efficiency for models with many population-level effects presumably because one cannot use design matrices and vectorized prior specifications for population-level parameters. Note that it was not possible to specify the exact same priors across packages due to varying parameterizations. Of course, efficiency depends heavily on the model, chosen priors, and data at hand so that the present results should not be over-interpreted.

6. Conclusion

The present paper is meant to provide a general overview on the R package **brms** implement-

	brms	**lme4**	**MCMCglmm**
Supported model types:			
Linear models	yes	yes	yes
Robust linear models	yes	no	no
Binomial models	yes	yes	yes
Categorical models	yes	no	yes
Multinomial models	no	no	yes
Count data models	yes	yes	yes
Survival models	yes[1]	yes	yes
Ordinal models	various	no	cumulative
Zero-inflated and hurdle models	yes	no	yes
Generalized additive models	yes	no	no
Non-linear models	yes	no	no
Additional modeling options:			
Variable link functions	various	various	no
Weights	yes	yes	no
Offset	yes	yes	using priors
Multivariate responses	limited	no	yes
Autocorrelation effects	yes	no	no
Category specific effects	yes	no	no
Standard errors for meta-analysis	yes	no	yes
Censored data	yes	no	yes
Truncated data	yes	no	no
Customized covariances	yes	no	yes
Missing value imputation	no	no	no
Bayesian specifics:			
parallelization	yes	–	no
population-level priors	flexible	–[3]	normal
group-level priors	normal	–[3]	normal
covariance priors	flexible	–[3]	restricted[4]
Other:			
Estimator	HMC, NUTS	ML, REML	MH, Gibbs[2]
Information criterion	WAIC, LOO	AIC, BIC	DIC
C++ compiler required	yes	no	no
Modularized	no	yes	no

Table 1: Comparison of the capabilities of the **brms**, **lme4** and **MCMCglmm** package. Notes: (1) Weibull family only available in **brms**. (2) Estimator consists of a combination of both algorithms. (3) Priors may be imposed using the **blme** package (Chung *et al.* 2013). (4) For details see Hadfield (2010).

	brms	**rstanarm**	**rethinking**
Supported model types:			
Linear models	yes	yes	yes
Robust linear models	yes	yes[1]	no
Binomial models	yes	yes	yes
Categorical models	yes	no	no
Multinomial models	no	no	no
Count data models	yes	yes	yes
Survival models	yes[2]	yes	yes
Ordinal models	various	cumulative[3]	no
Zero-inflated and hurdle models	yes	no	no
Generalized additive models	yes	yes	no
Non-linear models	yes	no	limited[4]
Additional modeling options:			
Variable link functions	various	various	various
Weights	yes	yes	no
Offset	yes	yes	yes
Multivariate responses	limited	no	no
Autocorrelation effects	yes	no	no
Category specific effects	yes	no	no
Standard errors for meta-analysis	yes	no	no
Censored data	yes	no	no
Truncated data	yes	no	yes
Customized covariances	yes	no	no
Missing value imputation	no	no	yes
Bayesian specifics:			
parallelization	yes	yes	yes
population-level priors	flexible	normal, Student-t	flexible
group-level priors	normal	normal	normal
covariance priors	flexible	restricted[5]	flexible
Other:			
Estimator	HMC, NUTS	HMC, NUTS	HMC, NUTS
Information criterion	WAIC, LOO	AIC, LOO	AIC, LOO
C++ compiler required	yes	no	yes
Modularized	no	no	no

Table 2: Comparison of the capabilities of the **brms**, **rstanarm** and **rethinking** package. Notes: (1) Currently only implemented on a branch on GitHub. (2) Weibull family only available in **brms**. (3) No group-level terms allowed. (4) The parser is mainly written for linear models but also accepts some non-linear model specifications. (5) For details see https://github.com/stan-dev/rstanarm/wiki/Prior-distributions.

Dataset	Function call
cake	
lme4	`lmer(angle ~ recipe * temperature + (1|recipe:replicate), data = cake)`
brms	`brm(angle ~ recipe * temperature + (1|recipe:replicate), data = cake)`
sleepstudy	
lme4	`lmer(Reaction ~ Days + (Days|Subject), data = sleepstudy)`
brms	`brm(Reaction ~ Days + (Days|Subject), data = sleepstudy)`
cbpp[1]	
lme4	`glmer(cbind(incidence, size - incidence) ~ period + (1 | herd), family = binomial("logit"), data = cbpp)`
brms	`brm(incidence | trials(size) ~ period + (1 | herd), family = binomial("logit"), data = cbpp)`
grouseticks[1]	
lme4	`glmer(TICKS ~ YEAR + HEIGHT + (1|BROOD) + (1|LOCATION), family = poisson("log"), data = grouseticks)`
brms	`brm(TICKS ~ YEAR + HEIGHT + (1|BROOD) + (1|LOCATION), family = poisson("log"), data = grouseticks)`
VerbAgg[2]	
lme4	`glmer(r2 ~ (Anger + Gender + btype + situ)^2 + (1|id) + (1|item), family = binomial, data = VerbAgg)`
brms	`brm(r2 ~ (Anger + Gender + btype + situ)^2 + (1|id) + (1|item), family = bernoulli, data = VerbAgg)`

Table 3: Comparison of the model syntax of **lme4** and **brms** using data sets included in **lme4**. Notes: (1) Default links are used to that the link argument may be omitted. (2) Fitting this model takes some time. A proper prior on the population-level effects (e.g., `prior = set_prior("normal(0,5)")`) may help in increasing sampling speed.

ing MLMs using the probabilistic programming language **Stan** for full Bayesian inference. Although only a small selection of the modeling options available in **brms** are discussed in detail, I hope that this article can serve as a good starting point to further explore the capabilities of the package.

For the future, I have several plans on how to improve the functionality of **brms**. I want to include multivariate models that can handle multiple response variables coming from different distributions as well as new correlation structures for instance for spatial data. Similarily, distributional regression models as well as mixture response distributions appear to be valuable extensions of the package. I am always grateful for any suggestions and ideas regarding new features.

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. Two anonymous reviewers provided very detailed and thoughtful suggestions to substantially improve both the package and the paper. Furthermore, Prof. Philipp Doebler and Prof. Heinz Holling have given valuable feedback on earlier versions of the paper. Lastly, I want to thank the many users who reported bugs or had ideas for new features, thus helping to continuously improve **brms**.

References

Andersen EB (1977). "Sufficient Statistics and Latent Trait Models." *Psychometrika*, **42**(1), 69–81.

Andrich D (1978a). "Application of a Psychometric Rating Model to Ordered Categories which are Scored with Successive Integers." *Applied Psychological Measurement*, **2**(4), 581–594.

Andrich D (1978b). "A Rating Formulation for Ordered Response Categories." *Psychometrika*, **43**(4), 561–573.

Apple Inc (2015). ***Xcode*** *Software, Version 7*. Cupertino, USA. URL https://developer.apple.com/xcode/.

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

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

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

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancour M, Brubaker MA, Guo J, Li P, Ridell A (2015). "**Stan**: A Probabilistic Programming Language." *Journal of Statistical Software*. URL http://www.stat.columbia.edu/~gelman/research/published/Stan-paper-aug-2015.pdf.

Carvalho CM, Polson NG, Scott JG (2009). "Handling Sparsity via the Horseshoe." In *International Conference on Artificial Intelligence and Statistics*, pp. 73–80.

Carvalho CM, Polson NG, Scott JG (2010). "The Horseshoe Estimator for Sparse Signals." *Biometrika*, pp. 1–16.

Christensen RHB (2015). "**ordinal** – Regression Models for Ordinal Data." R package version 2015.6-28. http://www.cran.r-project.org/package=ordinal/.

Chung Y, Rabe-Hesketh S, Dorie V, Gelman A, Liu J (2013). "A nondegenerate penalized likelihood estimator for variance parameters in multilevel models." *Psychometrika*, **78**(4), 685–709. URL http://gllamm.org/.

Creutz M (1988). "Global Monte Carlo Algorithms for Many-Fermion Systems." *Physical Review D*, **38**(4), 1228.

Damien P, Wakefield J, Walker S (1999). "Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical Models by Using Auxiliary Variables." *Journal of the Royal Statistical Society B, Statistical Methodology*, pp. 331–344.

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.

Ezzet F, Whitehead J (1991). "A Random Effects Model for Ordinal Responses from a Crossover Trial." *Statistics in Medicine*, **10**(6), 901–907.

Fox J, Weisberg S (2011). *An R companion to Applied Regression, Second Edition*. Sage.

Gabry J (2015). ***shinystan**: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models*. R Package Version 2.0.0, URL http://CRAN.R-project.org/package=shinystan.

Gabry J, Goodrich B (2016). *rstanarm: Bayesian Applied Regression Modeling via **Stan***. R package version 2.9.0-3, URL https://CRAN.R-project.org/package=rstanarm.

Gelfand AE, Dey DK, Chang H (1992). "Model Determination Using Predictive Distributions with Implementation via Sampling-Based Methods." *Technical report*, DTIC Document.

Gelfand AE, Smith AF (1990). "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, **85**(410), 398–409.

Gelman A (2006). "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis*, **1**(3), 515–534.

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.

Gelman A, Rubin DB (1992). "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science*, pp. 457–472.

Geman S, Geman D (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6), 721–741.

Griewank A, Walther A (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Siam.

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.

Hastings WK (1970). "Monte Carlo Sampling Methods Using Markov Chains and their Applications." *Biometrika*, **57**(1), 97–109.

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.

Ionides EL (2008). "Truncated Importance Sampling." *Journal of Computational and Graphical Statistics*, **17**(2), 295–311.

Juárez MA, Steel MF (2010). "Model-Based Clustering of Non-Gaussian Panel Data Based on Skew-t Distributions." *Journal of Business & Economic Statistics*, **28**(1), 52–66.

Kass RE, Natarajan R (2006). "A Default Conjugate Prior for Variance Components in Generalized Linear Mixed Models (Comment on Article by Browne and Draper)." *Bayesian Analysis*, **1**(3), 535–542.

Lewandowski D, Kurowicka D, Joe H (2009). "Generating Random Correlation Matrices Based on Vines and Extended Onion Method." *Journal of Multivariate Analysis*, **100**(9), 1989–2001.

Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000). "**WinBUGS** a Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and Computing*, **10**(4), 325–337.

Masters GN (1982). "A Rasch Model for Partial Credit Scoring." *Psychometrika*, **47**(2), 149–174.

McElreath R (2016). *rethinking: Statistical Rethinking Course and Book Package*. R package version 1.58, URL https://github.com/rmcelreath/rethinking.

McGilchrist C, Aisbett C (1991). "Regression with Frailty in Survival Analysis." *Biometrics*, pp. 461–466.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953). "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics*, **21**(6), 1087–1092.

Natarajan R, Kass RE (2000). "Reference Bayesian Methods for Generalized Linear Mixed Models." *Journal of the American Statistical Association*, **95**(449), 227–237.

Neal RM (2003). "Slice Sampling." *The Annals of Statistics*, pp. 705–741.

Neal RM (2011). "MCMC Using Hamiltonian Dynamics." *Handbook of Markov Chain Monte Carlo*, **2**.

Pinheiro J, Bates D (2006). *Mixed-Effects Models in S and S-PLUS*. Springer 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.

Plummer M (2008). "Penalized Loss Functions for Bayesian Model Comparison." *Biostatistics*.

Plummer M (2013). ***JAGS**: Just Another Gibs Sampler*. URL http://mcmc-jags.sourceforge.net/.

R Core Team (2015a). ***Rtools** Software, Version 3.3*. R Foundation for Statistical Computing, Vienna, Austria. URL https://cran.r-project.org/bin/windows/Rtools/.

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

Samejima F (1969). "Estimation of Latent Ability Using a Response Pattern of Graded Scores." *Psychometrika Monograph Supplement*.

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.

Spiegelhalter D, Thomas A, Best N, Lunn D (2003). ***WinBUGS** Version - 1.4 User Manual*. URL http://www.mrc-bsu.cam.ac.uk/bugs.

Spiegelhalter D, Thomas A, Best N, Lunn D (2007). ***OpenBUGS** User Manual, Version 3.0.2*.

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002). "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society B, Statistical Methodology*, **64**(4), 583–639.

Stan Development Team (2016a). ***Stan**: A C++ Library for Probability and Sampling, Version 2.11.1*. URL http://mc-stan.org/.

Stan Development Team (2016b). ***Stan** Modeling Language: User's Guide and Reference Manual*. URL http://mc-stan.org/manual.html.

Tutz G (1990). "Sequential Item Response Models with an Ordered Response." *British Journal of Mathematical and Statistical Psychology*, **43**(1), 39–55.

Tutz G (2000). *Die Analyse Kategorialer Daten: Anwendungsorientierte Einführung in Logit-Modellierung und Kategoriale Regression*. Oldenbourg Verlag.

Van Der Ark LA (2001). "Relationships and Properties of Polytomous Item Response Theory Models." *Applied Psychological Measurement*, **25**(3), 273–282.

van der Linde A (2005). "DIC in Variable Selection." *Statistica Neerlandica*, **59**(1), 45–56.

Vehtari A, Gelman A, Gabry J (2015a). "Efficient Implementation of Leave-One-Out Cross-Validation and WAIC for Evaluating Fitted Bayesian Models." *Unpublished manuscript*, pp. 1–22. URL http://www.stat.columbia.edu/~gelman/research/unpublished/loo_stan.pdf.

Vehtari A, Gelman A, Gabry J (2015b). ***loo**: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*. R Package Version 0.1.3, URL https://github.com/jgabry/loo.

Watanabe S (2010). "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory." *The Journal of Machine Learning Research*, **11**, 3571–3594.

Wood SN (2012). "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models." *Journal of the American Statistical Association*.

Yee TW (2010). "The **VGAM** Package for Categorical Data Analysis." *Journal of Statistical Software*, **32**(10), 1–34.

Zuur AF (2014). *A beginner's Guide to Generalized Additive Models with R*. Highland Statistics Limited.

Affiliation:

Paul-Christian Bürkner
Department of Statistics
Faculty of Psychology
University of Münster
48149, Münster
E-mail: paul.buerkner@wwu.de
URL: http://wwwpsy.uni-muenster.de/Psychologie.inst4/AEHolling/personen/buerkner.html

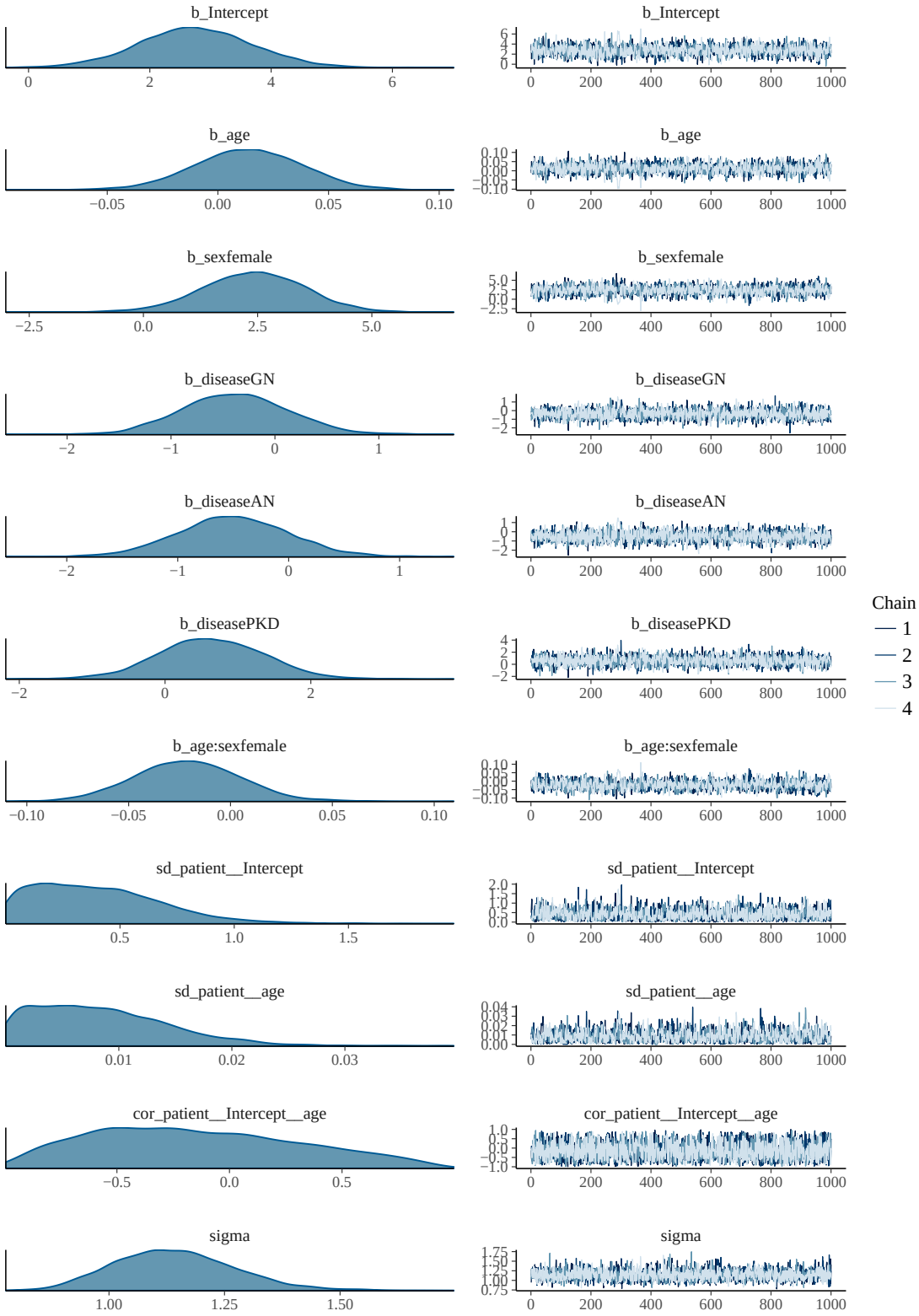

Figure 2: Trace and Density plots of all relevant parameters of the kidney model discussed in Section 4.

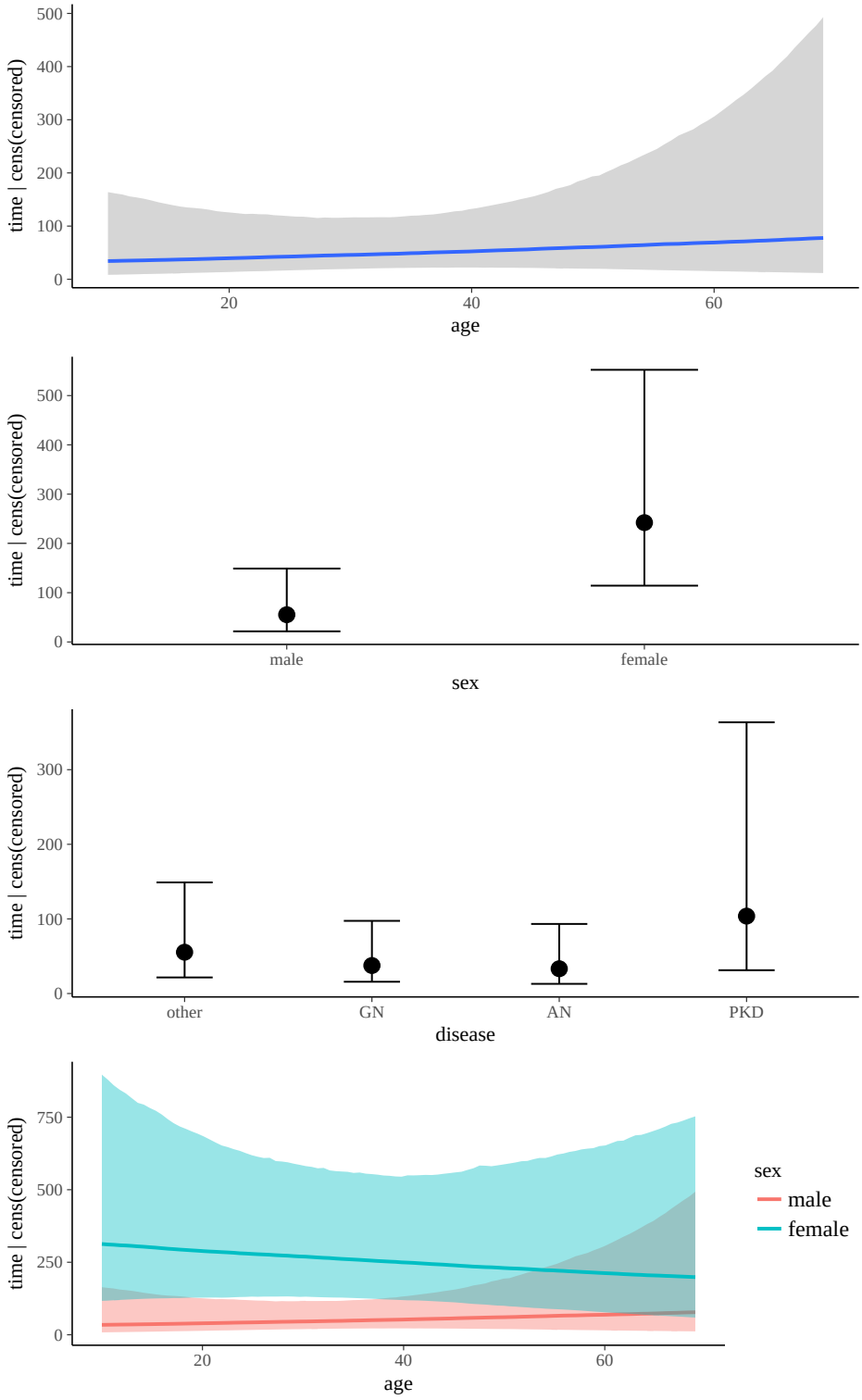

Figure 3: Marginal effects plots of all population-level predictors of the kidney model discussed in Section 4.

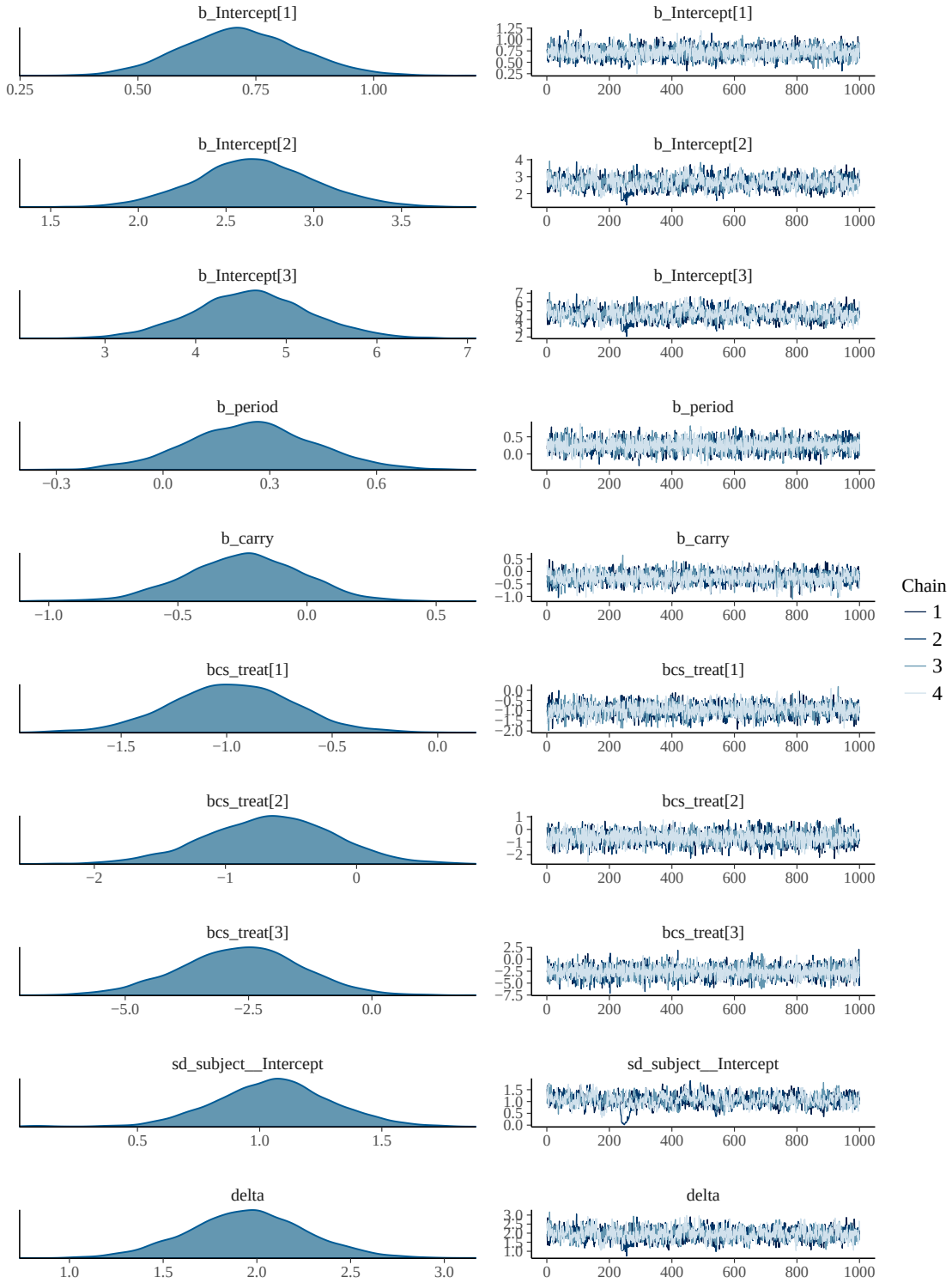

Figure 4: Trace and Density plots of all relevant parameters of the inhaler model discussed in Section 4.