Review of “A Lego Toolbox for Flexible Bayesian Regression (and Beyond)”

Main comments

The paper describes the bamlss package, which provides a modular and extensible framework for building and fitting general GAM models, in a Bayesian context. The package does a good job in providing a common parsimonious interface to different samplers and optimizers, and in breaking down the model building and fitting phases. Hence, the package is definitely useful. The paper is well written, but the text could be clearer in several places (see the minor comments below).

I have two main comments. The first is that it would be very useful to have a table of all the samplers and optimizers provided by the package. It would also be useful to specify which classes of models are compatible with which optimizers/samplers (as I understand some samplers are specifically aimed at certain models, e.g. cox_mcmc.). Maybe the best way to do this is to include a function in the package which, given a bamlss model specification as input, returns a list of optimizers/samplers that could be used to fit it. A related suggestion is that it would be useful to adopt prefixes in the package to indicate sampler and optimizers (e.g., all the sampler could start with sam_ and all optimizers with opt_), because the package provides quite a lot of functions but no consistent naming convention for functions that do similar things. The second comment is that the paper needs to clarify the fact that some features of bamlss implicitly depend on packages and stand alone software such as BayesX, which must be installed separately. Are you planning to make bamlss depend on R2BayesX explicitly (i.e. via the Depends or Imports fields) in future versions of the package?

Minor comments

- pg 1 says “(3) Enhanced inference infrastructure, typically Bayesian, beyond classical frequentist significance tests.” maybe rephrase this, as it seems to say that frequentist inference consists only of significance tests.
- pg 3, talking about gradient boosting “However, obtaining MCMC samples from the posterior distributions corresponding to such models is not easily available in these packages” in my understanding, the problem is not that it is difficult to do MCMC sampling for boosting, the problem is that that posterior distribution is undefined (or at least not explicily defined) for such models. Some rephrasing would be helpful here.
- pg 4 “Now, a standard Bayesian binomial logit model using the default MCMC algorithm can be fitted.” what MCMC sampler has been used here (slice sampling?) and where does the sampling occur (in BayesX or in bamlss)?
- pg 5 “In addition, the acceptance probabilities alpha are reported and indicate proper behavior of the MCMC algorithm.” I am not sure I know what the acceptance probabilities are. That is, I guess that they are not just the acceptance ratio of the MCMC chain. Please clarify in the text.
- pg 6 “maximum auto-correlation for all parameters” please clarify what you mean by maximum auto-correlation, as it might be, for example, the maximum auto-correlation across the parameters at each lag or averaged across the lags.
- pg 10 “Moreover, note that all smooth terms, i.e., te(), ti(), etc., are supported by bamlss. This way, it is also possible to incorporate user defined model terms.” Please add more explanations, as the fact that bamlss support te(), ti(), etc. does not necessarily imply that user defined terms can be used.
- pg 11 “According the histogram and the quantile-quantile plot of the resulting randomized quantile residuals in Figure 5, the model seems to fit relatively well. Only for very low and very high values of accel the fitted distributions seem to be less appropriate.” I guess that the last part of the sentence refers to the QQ-plot, but the plot does not provide confidence intervals, hence it might be that there is no significant departure from normality in the tails.
- pg 13 “For fully Bayesian inference the log-posterior is either used for posterior mode estimation, or for solving high-dimensional integrals. e.g., for posterior mean estimation MCMC samples need to be computed.” I would remove “fully” as for some people mode estimation if not full Bayesian inference.

- pg 13 bottom, the paragraph starting with “Using a basis function approach...” needs some rephrasing. In fact “a basis function approach” is a bit vague and later in the paragraph the $\mathbf{G}_{jk}$’s are referred to as both derivative matrices and penalty matrix, which can be understood only by readers already familiar with these models.

- pg 14 bottom “Estimated functions $f^{jk}(\cdot)$ are usually centered around their mean...” please clarify what this means in the text, as I guess that the effect are centered to have mean equal to 0 across the data (i.e., $\sum_i f^{jk}(x_i) = 0$), but this is not what the text says.

- pg 17, top “In summary, the architecture is very flexible such that users interested in implementing new models only need to focus on the estimation step, i.e., write optimizer() or sampler() functions and get all post-processing and extractor functionalities “for free”.” This sounds a bit confusing, in fact if I want to implement a new model, why should I implement a new sampler rather than use one of those already provided by bamlss? Are you referring to non-standard models that cannot use the standard samplers? Please clarify.

- pg 21, table 3. I see that the link functions have to be specified by name, is it possible to pass the link functions directly? I guess that that would allow more flexibility, in particular the use of user-defined link functions.

- pg 21 table 3 and text under the table. I am a bit surprised that the mixed elements of the Fisher information are not needed. Is this because the model are fitted by backfitting? Please state explicitly that you are using the diagonal elements of the Fisher information matrix. Also, I guess that the observed Fisher information cannot be used in place of its expected version?

- pg 23 to make the example more interesting, it would make sense providing a brief description of the variables contained in the FlashAustria data set.

- pg 24 “With a statistical model on hand one could predict lightning counts for the time before 2010 and thus analyze lightning events in the past for which no observations are available.” To better motivate the example, please add a sentence to explain what is the purpose of reconstructing lightning events before 2010 (is this an useful thing to do in practice? I guess so, but the text does not say anything about it).

- pg 24 “The second element specifies the formula for parameter θ. Hence well known for their sampling properties, we are applying P-splines (Eilers and Marx 1996) for all terms.” Please rephrase the second sentence, which is quite unclear (e.g., “We adopt P-splines (Eilers and Marx 1996), which are well known for their sampling properties.”).

- pg 24 the model seems to take 5 + 27 = 32 minutes to compute. How does this compare with, for instance, a brms or jagam implementation using the same (or similar) model? Is the sampler implemented in R or C++?

- pg 25 “After 1000 iterations the term s(q_prof_PC1).mu has the highest contribution to the log- likelihood with 282 followed by s(sqrt_cape).mu with 344. The term of the parameter θ s(sqrt_lsp).theta has a relatively small contribution with 38.” but from figure 7 it seems that s(sqrt_cape).mu has the highest contribution, while s(sqrt_lsp).theta does not appear in the plot.

Typos

- pg 3 bottom “lighting” -> “lightning”

- pg 3 “show cases” -> “showcases”

- pg 4 “optimizes the posterior mode” -> “finds the posterior mode”

- pg 5 bottom “Before proceding the” -> “Before proceding with the”

- pg 15 “staring value” -> “starting value”

- pg 17 bottom “in the above” -> “above”

- pg 21 “Merely all” I guess you meant “Almost all”?