Generalized additive models to analyze biomedical non-linear longitudinal data in R:
Beyond repeated measures ANOVA and Linear Mixed Models

Journal of Submission: Statistics in Medicine
Manuscript ID: SIM-21-0640

Corresponding author: Timothy J. Muldoon*

Department of Biomedical Engineering, University of Arkansas, Fayetteville, AR, USA

tmuldoon@uark.edu

General Comments to the Reviewers

We would like to thank the reviewer for the careful and thorough analysis of this manuscript to *Statistics in Medicine* and for the thoughtful comments and constructive suggestions, which helped improve the quality of this manuscript. We carefully considered the reviewer's comments and in this document, explain how we revised the manuscript based on those comments and suggestions.

General Comments to the Editor

Dr. Platt,

The authors thank you for your determination that our manuscript may be suitable for resubmission in *Statistics in Medicine* after addressing the reviewer's comments. To this end, we have addressed all critiques. We hope these revisions, submitted on December 5, 2021, improve the manuscript so it is deemed worthy of publication in *Statistics in Medicine*. Following are our detailed responses to reviewer comments.

Reviewer's Introduction

Mundo and colleagues present a tutorial on the use of generalized additive models to analyze longitudinal data. A comparison with repeated measures ANOVA and linear mixed models is provided and a recurring example using simulated data is used to illustrate the differences among the methods and the advantages of GAM. While I like the general approach and aim of the manuscript, there are a number of inaccuracies and omissions, especially in the discussion of GAMs that make it difficult for me to support publication at this stage. I believe — given evidence elsewhere in the manuscript that the authors really do know the subject — that these inaccuracies and omissions stem from preparing a tutorial in a scientific paper format where word-length considerations come into play. Below I outline the main areas where I feel the GAM methodology is inaccurately presented or important topics omitted, and make suggestions to improve the manuscript.

Reply to Introduction

We appreciate Dr. Simpson's comments and critiques to our manuscript.

Comments from reviewers

Missing data

Reviewer's Comments

I found this use of "missing" data to be a little confusing. I understand what the authors are getting at, but it suggests that GAMs can handle this whereas rm-ANOVA and LMMs can't. While I appreciate that rm-ANOVA might require balanced observations for the group errors, LMMs are just as able to handle "missingness" (in the sense implied by the authors) as GAMs are. Additionally, the "missingness" isn't meant solely as missing in the statistical sense, but really it is due to irregular or incomplete sampling, more generally. It would be helpful to make the discussion of missingness about balanced data and to try to avoid terms like "missing" as that may be inferred to suggest GAMs and LMMs are immune to missing data problems (they aren't), they just don't require balance.

Reply to Reviewer's Comments

We appreciate the comments about clarity in the use of the term "missingness". Please note that the title for Section 3.4 has been changed from "Missing observations" to "Unbalanced data". We have restructured this section in the following manner:

- The term "missing data" has been removed to avoid confusion. We now refer to different number of observations as "unbalanced data" (L201-204).
- We have made clear that LMEMs can also work with missing observations (L209)
- Emphasis has been provided to the fact that GAMs are not immune to missing data problems, and that researchers need to minimize missing observation rates (L217-220).

Unnecesary restriction of LMMs

Reviewer's Comments

Why not use quadratic effects of time in the LMM? I appreciate this would make the model more complex, but trying to fit a quadratic effect with a linear model strikes me as futile and somewhat of a straw person argument to make.

Reply to Reviewer's Comments

We appreciate the reviewer's perspective regarding the somewhat contradictory argument of fitting a linear effects LMM to quadratic data. However, this is an intentional mistake because we follow the Statistical analysis logic followed by most Biomedical researchers: Fit a LMM, and if you get significant p-values that's good enough. We are trying to convey the point that although "significant" p-values can be obtained, visualizing the model fit is something that should be required, but that is seldomly done. **Incorporate in the section that a strenght of GAMs is that they can learn the function from the data, while LMMs can't.**

GLMMs

GLMs, GAMs, and conditional distributions

Thin plate regression splines

Reviewer's Comments

The authors define the default basis in {mgcv}, the software being used here to illustrate fitting GAMs, as "...thin plate regression splines are an optimized version that work well with noisy data." This is not a good description of what a thin plate regression spline (hereafter TPRS) is, and is not a good description of what the low-rank versions in {mgcv} are.

Reply to Reviewer's Comments

The following changes have been made:

- The information on L289-294 pertaining splines has been changed. L301-309 now describe cubic splines (CS), thin plate splines (TPS) and thin plate regression splines (TPRS) in a manner that conveys their general properties and advantages (or disadvantages) to a non-Statistical reader.

Reviewer's Comments

Much better, I believe, is to talk about this in terms of basis functions, of which there are five in this example. You can use the same terminology to refer to the CRS basis, and others. That in the CRS if you want k basis functions you need k knots (IIRC). CRS require you to specify the knot locations (or let the software spread them evenly through the data); TPRS don't. Hence the paragraph starting on L195 is confusing and misleading. The basis functions are not piece-wise polynomial and there are not "region[s] where a different set of basis functions will be used". Indeed, the functions you show operate throughout the range of the covariate.

Reply to Reviewer's Comments

- L297-300 no longer refer to "knots" when referring to the construction of the smooth, now indicating that the number of *basis functions* is what is specified when constructing the smoother. In the updated manuscript, this change appears in L310-315, and is more clear to the reader due to the changes introduced in L301-L309.

Penalised splines

Reviewer's Comments

You talk a little about "wiggliness" and penalising the weights for each basis function towards 0 but you don't really explain this most crucial concept of the model fitting problem and why modern GAMs are so much better than the GAMs developed by Hastie and Tibshirani when they first introduced GAMs to the world. You really need to define wiggliness beyond something that is used to avoid overfitting. Typically it is squared second derivative of the estimated function; so we limit the curvature of the fitted function by default. Technically the penalty is the penalty matrix - it is a matrix and it is fixed once we define the basis functions. What isn't fixed is (are) the smoothness parameter(s) of the smooth; it is those that control how much penalty we subtract from the log-likelihood of the data given the model estimates. To speak of a "weak" or "strong" penalty was a little confusing for me. Hence fitting a GAM requires one to estimate parameters and one or more smoothness parameters for each smooth in the GAM, plus any parameters required for parametric terms. By glossing over these important concepts, the reader is left wondering what wiggliness is, how we measure overfit? etc.

Bayesian

Reviewer's Comments

GAMs in {mgcv} are considered to be empirical Bayesian models, but in general GAMs can be fully Bayesian. You can fit fully Bayesian GAMs using INLA and JAGS using functions from {mgcv}, and the {brms} package allows full Bayesian GAMs to be estimated simply using Stan for example. This needs to be clarified. I'm not really sure what you mean by the sentence beginning "Moreover, the use of the restricted maximum…". I think I get what you mean; by casting the wiggly parts of smooths as random effects we can estimate the fit using REML and standard linear mixed model software. However, you can fit the model using the full fat version of maximum likelihood and these models would still be empirical Bayes if fitted by {mgcv}.

Reply to Reviewer's Comments

- We have clarified the concept "empirical Bayesian", which appeared on L318-L319 in the original manuscript. In the revised manuscript, L335-338 now indicate that Stan, JAGS or other probabilistic programming language can be used to estimate GAMs using a full Bayesian approach.
- The sentence "Moreover, the use of the restricted maximum likelihood (REML) to estimate the smoothing parameters gives an empirical estimate of the smooth model", which appears in L319 in the original manuscript has been removed from the text. Instead, the concept of REML has been moved to Section 6.2 L392-395, where we state some of the reasons indicated by Wood when choosing restricted maximum likelihood (REML) over the default general cross validation (GCV) method for smooth parameter estimation in *mgcv*.

Coverage of confidence intervals

Reviewer's Comments

This entire section from L320 onward through to the end of Section 5 needs some work. When viewed from the Bayesian perspective, the intervals are Bayesian credible intervals. When viewed from a frequentist perspective, the same intervals are confidence intervals but instead of having the typical point-wise interpretation they have an across the function interpretation. The description of what this means is wrong — you are almost giving the incorrect definition of a confidence interval here. The interval either does or does not contain the true function. That is given. I'm sure this is just a slip then on L324-325. Also, your description implies a simultaneous interval, although I don't think you intended this. One could interpret "95% of the time" as meaning 95% of the functions are contained in their entirety.

What across-the-function means is simply that if we average the coverage of the interval over the entire function we get approximately the nominal coverage, 95% say. For this to occur then, some areas of the function must have more than nominal coverage and some areas less than the nominal coverage.

Reply to Reviewer's Comments

The content in L320-329 regarding confidence intervals (CIs) in the original submission has been changed in the following manner:

- L341-L355 now contain a more detailed and accurate explanation on the differences and interpretation of "poin-twise" CIs and "across-the-function" CIs. We removed the "95% of the time" phrase to avoid confusion, and instead we provide an explanation that focuses on random sampling: "if 100 random samples are obtained and a GAM and CI is calculated for each one of them, it would be expected that 95 out of the 100 fitted CIs entirely contain the true function" (L352-354). We also reference the work of Marra and Wood if the reader desires a more in-depth exploration of across the function CIs.

Differences in smooths

Reviewer's Comments

Reply to Reviewer's Comments

Appendix

Reviewer's Comments

In the appendix I think you are needlessly restricting the size of the basis dimension to be k = 5, hence 4 basis functions per smooth when identifiability constraints are applied. Is there a reason I'm seeing here why you could leave this at the default k = 10 and really see the effect of the shrinkage as the EDF of the resulting smooths should be similar to the EDF you have with k = 5? This would also likely help with the k-index being low because there's not a lot of shrinkage you can do when the maximum EDF possible is 4 (per smooth).

Reply to Reviewer's Comments

In L367 it is indicated that the simulated data used to fit the GAM that the reviewer alludes to has only 5 unique covariates (days 0, 2, 5, 7 and 10). The computational requisite in the smooth estimation of having the maximum number of basis equal to the number of unique values in the covariate makes it impossible to fit a GAM with a k>5, as the error "A term has fewer unique covariate combinations than specified maximum degrees of freedom" is thrown by *mgcv*. We consider that because the k-index for the model is 1.04 the basis dimension for the smooth is adequate (Wood says that "the further below 1 this is, th emore likely it is that there is missed pattern left in the residuals").

Reviewer's Comments

Why change the model object name notation here? You had gam_00 previously and now you call the model m1? As you are already showing appraise() output for the diagnostic plots, you could use check.k() from {mgcv} to just get the basis dimension check.

Reply to Reviewer's Comments

The name of the model has been changed to `gam_02`, and it appears with this notation in Section 6.2, L383 so it matches the workflow of model selection presented in the Appendix.