Generalized additive models to analyze non-linear trends in biomedical longitudinal data using R:

Beyond repeated measures ANOVA and Linear Mixed Models

Response to Reviewer's Comments

Journal of Submission: Statistics in Medicine

Manuscript ID: SIM-21-0640.R1

Corresponding author: Timothy J. Muldoon*

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

tmuldoon@uark.edu

Contents

General Comments to the Reviewers

We would like to thank the reviewers 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 reviewers 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 January 4th, 2022, improve the manuscript so it is deemed worthy of publication in *Statistics in Medicine*. Following are our detailed responses to reviewer comments.

Reviewer #1

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

Comments from Reviewer

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:

- L206-L209: The term "missing data" has been removed to avoid confusion. We now refer to different number of observations as "unbalanced data".
- L213: We have made clear that LMEMs can also work with missing observations.
- L221-L224: 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.

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 approach because we follow the statistical analysis logic followed by most biomedical researchers to analyze longitudinal data: Fit a LMM or an rm-ANOVA, and if a *post-hoc* analysis gives *p-values* below a certain threshold (usually 0.05) that's good enough. The purpose of the figure is to convey that changing models (from rm-ANOVA to LMEM) without visualizing the model fit is in this case is something that will propagate bias in the results and does not provide any substantial improvement to the analysis. However, we acknowledge that this point was not explicitly stated in the original manuscript and the following changes have been made:

- L95-L97: We mention that LMEMs assume a linear relationship "by default", but we indicate that polynomial effects can be used as well, though they have their own shortcomings.
- L120-L121 now reads "the limitations of LMEMs and rm-ANOVA regarding an expected trend in the response", instead of "linearity of response" (L119 in the original manuscript).
- L137 now reads "assumed relationship" instead of "linear relationship" (L135 in the original manuscript).
- Section 3.2 now is titled "Assumed relationship" instead of "Linear relationship"
- L180-L183 now indicate that polynomial effects can be used in a LMEM, but that they have little predictive power and lack biological or mechanistic interpretation.
- L247-L250: We provide additional context on the purpose of using linear models to fit quadratic data in which now read "We are aware that the simulated data used in this section present an extreme case that might not occur frequently in biomedical research, but they are used to 1) present the consequences of modeling non-linear trends in data with a linear model such as rm-ANOVA or a LMEM with"default" (linear) effects and, 2) demonstrate that, a visual assessment of model fit is an important tool that helps determine the validity of any statistical assumptions."

We believe that the aforementioned change makes a clear statement to the reader about our intentions in Fig.1. We want to emphasize that strictly speaking the term "default" (used for linear effects in a LMEM) might not be correct, but we are stretching the language to make the point that doing a simple switch from rm-ANOVA to a LMEMs (in R) without any additional considerations on the data results in the assumption of linear effects that still provide biased estimates.

- L270-L273 now include some of the reasons that complicate the use of polynomial effects in a LMEM for the data in Figure 1. Specifically, that in reality the true function is not known and that this complicates choosing the degree of the polynomial. We also reference Section 4.3.1, which provides more detail on the limitations of polynomial effects.
- L550-L552 In Section 6 (Discussion) emphasize again the limitations of polynomial effects when used on biological data, such as in the case of the simulated data in Fig. 1.

GLMMs

Reviewer's Comments

You don't even mention generalized linear mixed models; why? In the context of the tutorial they represent the logical bridge between GAM and LMM, especially when you consider that you are really fitting GAMMs here.

Reply to Reviewer's Comments

We appreciate Dr. Simpson's comment here and the equations provided, which we have now incorporated in the manuscript. We acknowledge that a brief overview of the different statistical model classes was missing in the original manuscript, and we have now extensively edited Section 4 to provide context on linear models (LMs), generalized linear models (GLMs) and generalized linear mixed models (GLMMs) before introducing GAMs.

- L282- L327: Section 4 has been renamed to "Linear Models and Beyond" and the different Subsections cover LMs, GLMs, GLMMs and GAMs. Because the different model classes are not commonly known in the biomedical field, we contextualize the previously presented rm-ANOVA and LMEM into their respective classes to provide the reader with a clear understanding of where each model fits (Section 4.1). We also provide equations for the linear predictors in each case (Equations (6) and (8)) following the notation of Equations (1) through (4). We want to emphasize that because the target audience of this paper are biomedical researchers that are not familiar with Mathematical Statistics, we follow notation that is easy to read (e.g., using $time_t$ instead of x_t), although it might not be notation followed in other Statistical publications.

We believe that because Section 4 now provides context and follows a logical progression of LMs, GLMs, GLMMs and GAMs in a way that is easily understandable, these changes properly addressed the reviewer comments.

- L325: We also acknowledge that the GAM model (Equation 5 in the original manuscript and Equation 10 in the revised document) corresponds to the **linear case**. This change will avoid the propagation of confusion that Dr. Simpson correctly indicated in his comment.
- L326-L327: We indicate that the term ε_{ijt} is the deviation of each observation from the mean.

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:

- L350-L358: The information that originally appeared in L289-294 pertaining splines has been changed, as we 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. Here, we no longer refer to TPRS as an "optimized version" but we indicate that they are a truncated version of thin plate splines (TPS).

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

- L363-L366: These lines correspond to L297-L300 on the original submission 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. We believe that this will make sense to the reader due to the information provided in L350-353.

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.

Reply to Reviewer's Comments

- We appreciate this comment regarding wiggliness. For this particular topic, we have implemented changes that address the reviewer's concerns. We have used a minimalist approach here in order to provide some facts about penalization to give a general idea of the logic behind it but without adding additional notation, which we believe will obscure the topic to the reader.

L367-375 has been added, and is a paragraph which briefly discusses wiggliness, the smoothness parameter λ and the penalization process using the integrated square of the second derivative of the spline. Our goal here is for the reader to understand the effect that λ in order to prevent overfitting.

- L374-L375: We have substituted "weak" and "strong" for "low" and "high".

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

- L394-L395 now clarifies this by indicating 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 5.2 L457-460, 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:

- L398-L405 now contain a more detailed and accurate explanation of"across-the-function" CIs. We have used the reviewer's comments to indicate that in GAM fitted using *mgcv*, the close to nominal across-the-function coverage is true when averaged over the entire function.
- L406-L413 now provide a brief explanation of simultaneous CIs and how they correct for the across the function CI. With this explanation we want to provide the reader with an intuitive understanding on the advantages of simultaneous CIs over across the function CIs when computing comparisons between different groups (which we discuss in L485-L503).
- We have modified Figure 3 to now include across the function and simultaneous CIs. Figure 3B, D now show the fitted smooths and both intervals (across the function and simultaneous), and we explain the rationale for their inclusion in L478-L483. In this way, the reader has a visual understanding of the difference between across the function and simultaneous CIs.

Differences in smooths

Reviewer's Comments

The essence of this is not "if the difference between [credible] intervals of the fitted smooths… is non-zero". For a variety of reasons this is not how it works. Instead we compute the difference of the smooths and then compute a credible interval around this difference. Also, to interpret this correctly we really need to do a simultaneous interval as you are looking at all points over the range of the covariate. If we just use the usual interval we have to consider the under and over coverage of the Nychka across-the-function intervals. You should talk about this issue in this section if you don't compute simultaneous intervals. This section

could also do with a little more explanation; {gratia} (and hence I) tens to consider differences of smooths to exclude the groups means. Here, if I read the paper and the code correctly, you included the group means. This works OK for simple models like this, but what if you had two or more parametric factor terms in the model? Then you would not (at least by default) have any way to isolate the group means for the by factor independently of the levels of the other factor. You would have to create differences conditional upon a level of the other factor. Users, however, tend to expect differences to include the group means, as you have chosen to present things here. Hence a note explaining the implications of this choice would be useful to forewarn the user.

Reply to Reviewer's Comments

We have made major changes to this section based on the reviewer's comments.

- We have removed the sentence "if the difference between the credible intervals of the fitted smooths... is non-zero", and now L487-L489 now read: "is that the difference between them can be computed, followed by the estimation of an empirical Bayesian simultaneous CI around this difference."
- L497-L498 The sentence now reads "unlikely to be distinguishable from zero".
- L507-513 now indicate that we are computing a simultaneous CI around the difference, and that we *included* the group means in our pairwise comparison. We also indicate to the reader the fact that if multiple parametric terms exist in the model the inclusion of the means can become problematic.

However, we want to point that our belief is that including the means makes the comparison more intuitive, as in that way the researcher can directly compare both the fitted smooths and the pairwise comparisons to infer any significant changes and their associated magnitude. In Appendix A we also have made the corrections to the calculation of the pairwise comparisons in Section A.5 (the GAM workflow section) so the reader is aware again that in our code we keep the group means.

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

L430 (in the revised manuscript) indicates 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, the more likely it is that there is missed pattern left in the residuals").

Reviewer's Comments

A.4.3 You say that gam_00 doesn't account for the nesting of the data by Group. It does but only partially; you get a separate smooth per group and those smooth will be trying to also account for the different means of the response in the two groups. I think it is much clearer to say just that; it is better to include parametric terms for the group means so the smooths capture the time-course differences about these overall group means of Y .

Reply to Reviewer's Comments

We believe Dr. Simpson refers here to `gam_01` instead of `gam_00`, as the latter does not include `Group` as part of the model specification and because of this, we don't get a separate smooth per group, which can be seen in the resulting single smooth fitted by the model. However, we believe that the clarification he provided here is important, and we have added it to the first paragraph of Section A.4 (Third model) so it is more clear to the reader.

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.
- Additionally, we have used some of the source code of the function `gam.check` in order to use the graphical output of `gratia` and the numerical diagnostic information of `gam.check` without repeating the diagnostic plots, the code for this is now in the Appendix B.

Suggested Changes

Reviewer's Comments

Essentially now that we have knot-free splines, we can focus on the concept of the user's prior for the upper limit on the wiggliness of the each smooth; this is set using k. You also should explain that we need to increase k a little above this prior expectation because the basis of dimension k + K has a richer set of functions of complexity k than a basis of dimension k. From a practical point of view, the user really only needs to think about some anticipated amount of wiggliness and set k to be a little larger than this, then fit the model. Next the user should do the check.k() test on the size of the basis to see if it was large enough. If it isn't, consider increasing k a bit and then refit and recheck k. Rinse and repeat.

Beyond the usual model diagnostics and choosing the conditional distribution of the response, that's all the user really needs to do to get fitting GAMs. And this basic concept is largely missing from the manuscript and the worked tutorial.

Reply to Reviewer's Comments

Our goal in the main manuscript is to familiarize biomedical researchers with the theory of GAMs and their advantages to analyze non-linear trends in longitudinal data. We have incorporated Dr. Simpson's comments about the concept of GAM fitting in the Section "Final Considerations" in Appendix A. Our rationale for adding the information Dr. Simpson mentions here is that we want the reader to understand that fitting GAMs needs has to be more than a mechanic exercise in order to be effective. Too often, researchers fit a statistical model in a mechanic fashion, and our goal in this work is to convince the read to move away from such approach.

After presenting the theory in the main manuscript and the workflow for model selection in Appendix A, we consider that the reader will better appreciate that using `k.check` is the last step of the chain of thought required to fit a GAM.

Reviewer's Comments

For the Appendix, I would suggest that you break up the code a bit to include a little narrative code explaining what each sub-chunk is doing. At the moment you have these monolithic blocks of code that produce a figure and/or some printed console output and the user is largely left to figure out what the code is doing. I

appreciate that there are comments, but for a tutorial, it would be better to have something more like a vignette rather than monolithic code blocks with comments.

Reply to Reviewer's Comments

We have taken the following approach to address this comment:

- We now have two Appendices: Appendix A, which is a tutorial for GAM selection, and Appendix B, which has the code used through the manuscript.

Additionally, the R functions used through the manuscript are now stored in a directory called *scripts* so they can be easily accessed if desired.

Because we consider that the majority of the readers will be interested in Appendix A, we have used a vignette-style approach by breaking up the code chunks to provide narrative in order to make it more understandable.

Readers interested in the code itself can then revise Appendix B, and in this way we avoid to show long sections of code upfront to the reader by using a single Appendix.

Figures

Reviewer's Comments

The figures are too tall — the aspect ratio used for the individual panels is wrong. You should be emphasizing the time axis so you should have individual panels that are wider than they are tall.

Reply to Reviewer's Comments

We appreciate this comment regarding figure aesthetics, we have now made the figures wider to emphasize the time axis.

Reviewer's Comments

Figure 3: on panel A, are the solid lines the truth in the main panel? If they are, you don't need the inset plot, and if they aren't, are they the mean of the simulated observations? If they are this latter, I think this is superfluous and it would be better to make the solid line the truth, about which you simulated noisy data.

Reply to Reviewer's Comments

Although data simulation is a very common practice in Statistics/Ecology, it not so much in biomedical research. The main goal of Figure 3 A is to show that simulation is indeed a plausible tool to explore data behavior and to test statistical models in biomedical research. It might seem too basic or repetitive, but we are trying to convince the reader that simulation can indeed produce synthetic data that can be useful. However, we do acknowledge that the inset was repetitive.

We have updated Figure 3 so that panel A shows both the simulated and real data and therefore it is easier for the reader to see the simulation of the output.

Figures

Reviewer's Comments

The figures are too tall — the aspect ratio used for the individual panels is wrong. You should be emphasising the time axis so you should have individual panels that are wider than they are tall.

Figure 3: on panel A, are the solid lines the truth in the main panel? If they are, you don't need the inset plot, and if they aren't, are they the mean of the simulated observations? If they are this latter, I think this is superfluous and it would be better to make the solid line the truth, about which you simulated noisy data.

Reply to Reviewer's Comments

We have changed the aspect of the figures to emphasize the time axis, making them wider.

On Figure 3, we have added the original data to panel A, which now displays both the simulated data (thin lines) and the original data (truth) in order to simplify the figure. We have updated also places in the manuscript that reference this Figure as well.

Minor Comments

Reviewer's Comments

L57 p values cannot be significant or otherwise, they just are. They might indicate statistically significant effects, where the weight of evidence against the null hypothesis is sufficient to meet some threshold of "significance". I note the use of scare quotes so perhaps this sentence was intended in jest but even so it is confusing and unnecessary.

Reply to Reviewer's Comments

We appreciate the reviewer's comments here. However, the indicated sentence is not intended in jest. We are using in this sentence (and in the paper in general) a statistical approach that is understandable by biomedical researchers in general. To this day, the field values "significance" (defined by the arbitrarily set threshold of $p<0.05$). We are trying to convince the reader that low *p-values* are possible with ill-fitted models, and we use language that is understandable by the reader.

The updated sentence in L59 now reads "significant effects (p-*value*<0.05)".

Reviewer's Comments

Also here we have "non-linear data" which almost surely should be "non-linear effects" — data aren't non-linear, but relationships between variables, "effects", might be. See also L316

Reply to Reviewer's Comments

We have changed the expression to "data that shows non-linear trends". The change has also been applied to L316 which is now L391 in the revised manuscript.

Reviewer's Comments

L93 "distribution of the errors of the random effects" makes no sense. Also LMMs do not restrict random effects to be i.i.d. Gaussian — they could be i.i.d. gamma or t distributed for example. That typical R software (and mgcv from the GAM side) makes an assumption that the distribution of individual random effects terms is Gaussian, doesn't mean the model class is so restricted.

Reply to Reviewer's Comments

- We have changed the sentence to "the distribution of the random effects, which need to be independent", which also indicates that errors need not be normally-distributed (L95 in the revised manuscript).

Reviewer's Comments

L185 here and throughout, check for situations where you have combined numerical citations to literature within narrative text; Here it should be "...see refs 40 & 42.)" for example, not superscript numerals.

Reply to Reviewer's Comments

We have updated all the occurrences of citations within the narrative text by using the last name of the author followed by the citation, which fixes the issue.

Reviewer's Comments

L244 "global" is perhaps more conventionally described as "population" in the mixed model literature.

Reply to Reviewer's Comments

We have changed the word from "global" to "population".

Reviewer's Comments

L283 I think it is better to say "...the mathematical space within which the true but unknown $f(x_t)$ is thought to exist."

Reply to Reviewer's Comments

We have updated this line to "the mathematical space within which the true but unknown $f(x_t \mid \beta_j)$ is thought to exist", in L339 in the revised manuscript.

Reviewer's Comments

L293 You never really describe what "data-drive flexibility" is. Why are GAMs data driven and LMMs and GLMs not? This is a critical concept to convey I feel.

Reply to Reviewer's Comments

This line has been removed from the revised manuscript. However, we acknowledge that we didn't clearly conveyed this point to the reader. In Section 6 (Discussion), L549-550 we now clearly state that a major advantage of GAMs is that they can learn non-linear trends from the data, thus eliminating the requirement of other models such as LMEMs where the relationship needs to be provided *a priori* by the user.

Reviewer's Comments

L351 Better to say "separate smooths" not "independent". These factor-by smooths share a common basis, they just get an entirely separate set of model coefficients and smoothness parameters for each level of the factor.

Reply to Reviewer's Comments

The sentence now reads "separate smooths for..." (L442 in the revised manuscript)

Reviewer's Comments

L354 "difference between group means" because the smooth encodes differences between the groups too, just not difference in the means. Be a little more specific here.

Reply to Reviewer's Comments

The sentence now reads "any differences between the group means" (L445).

Reviewer's Comments

L356 use straight quotes in code, not curly ones. And speficy the `family` here?

Reply to Reviewer's Comments

We use the code notation provided by the *LaTeX* package `listings`, and have updated the quotes to be straight in all code chunks. Regarding `family`, we believe that it is not necessary to include it here, but we indicate in L459-460 what the argument exists and that other distributions can be used.

Reviewer's Comments

L359 "contain the fitted" instead of "store the"

Reply to Reviewer's Comments

The sentence now reads "contains the fitted model" (L450-L451).

Reviewer's Comments

L360 Not "knots" - there are k-1 basis functions, but as described above, because this is a TPRS there are knots at the unique data values and then we decompose the full basis and take the required eigenvectors as the basis to use.

Reply to Reviewer's Comments

The updated sentence now reads "using four basis functions (plus the intercept)" (L452).

Reviewer's Comments

L362 why highlight Gaussian process smooths? They're not as useful as they seem as you need to optimise the length scale parameter yourself as otherwise, especially for longitudinal data they aren't great - the length scale is the largest separate in time of any pair of observations.

Reply to Reviewer's Comments

For the length of the covariates explored in this paper TPRS are sufficient. However, we want to make clear to the reader that other types of splines that might be appropriate in certain situations are available. To acknowledge this, L454 in the updated manuscript now reads "(a description of all the available smooths can be found by typing `?mgcv::smooth.terms` in the Console)".

Reviewer's Comments

L366 See also Matteo Fasiolo's {mgcViz} package

Reply to Reviewer's Comments

We acknowledge the existence of Dr. Fasiolo's *mgcvViz* package in Appendix A in the Section "Additional Resources".

Reviewer's Comments

L373 Do rm-ANOVA and LMMs require equally-spaced or complete observations?

Reply to Reviewer's Comments

Only rm-ANOVA requires equally-spaced and complete observations. To avoid confusion, the line now reads "Because GAMs do not require equally-spaced or complete observations for all subjects (as rm-ANOVA does)" (L471).

Reviewer's Comments

L377 delete "total"

Reply to Reviewer's Comments

The word "total" has been removed. The sentence now reads "consider the random deletion of 40% of the..." (L475)

Reviewer's Comments

L381 "little" -> "few"

Reply to Reviewer's Comments

The sentence now reads "...with as few as 4 observations per group..." (L483)

Reviewer's Comments

L393 "observed difference is significant" (not "change")

Reply to Reviewer's Comments

The sentence has been rewritten. It now reads "...the observed difference is significant..." (L496-L497).

Reviewer's Comments

Also, it is (exceedingly) unlikely to yield 0. Better to say that the estimated difference is unlikely to be distinguishable from 0.

Reply to Reviewer's Comments

The sentence in L497-L498 now reads "is unlikely to be distinguishable from zero".

Reviewer's Comments

L410 This needs to be reworded as what is written can be read as any difference between the control and treatment (and there is a difference as the controls are higher than the treated group before day 3) where you mean that the treatment group is only higher than the control group post day 3.

Reply to Reviewer's Comments

This was an unintentional mistake based on some previous simulation results. We have updated the paragraph, and it now reads "However, because the model is still able to pick the overall trend in StO_2, the pairwise comparison is able to estimate the change on day 3 where the mean difference between groups becomes statistically significant as in the full dataset smooth pairwise comparison" (L523-L525). The change acknowledges that Control was higher before day 3, and that the Treatment Group becomes significant after day 3 in relation to L515-L519.

Reviewer's Comments

L441 Delete "significant" here - we estimate the difference of the effects (what you call "change") and make inference using the uncertainty estimates of that difference. Perhaps you mean "identify" or similar instead of "estimate" here, instead.

Reply to Reviewer's Comments

We have made the appropriate changes. The sentence now reads "The GAM is therefore able to identify changes between the groups at time points..." (L566)

Reviewer # 2

Reviewer's Introduction

This is a well-written paper that gives a useful overview of methods that are underused in the field. It is written with great clarity all is well-targeted towards this journal's audience. I recommend accepting with

only minor revisions.

Reply to Introduction

The authors thank the reviewer for deeming our within the scope of the Journal.

Comments from Reviewer

There is one are where I feel expanded discussion would improve the paper: interpretability. A common challenge in presenting GAMs is the lack of interpretability of coefficients. While, in general, linear model coefficients simultaneously can be used to present effect sizes and significance, GAM models do not present estimands with such clear meaning. This is an area that practitioners will run into in taking up this new method, and it would serve them best to address it directly. A common approach to this is thinking carefully about the quantity of interest (such as the difference in outcomes at two specific values of a predictor, rather than a slope), and calculating this value, with uncertainties, using posterior samples. This is a natural extension of section 6.3., where one can discuss how calculations from the posterior can be used to determine the value of these quantities of interest in addition to significance. It would be useful to discuss posterior probabilities and the notion of a q-value as a Bayesian measure of significance that can be quantitative in addition to visual. The {gratia} package provides some tools for sampling from GAM posteriors that could be demonstrated.

Reply to Reviewer's Comments

Although we acknowledge that these concepts are important, in our manuscript we want to focus on the visualization part as we believe is more intuitive and informative in a biomedical context. We have addressed the comment from the reviewer by providing context about interpretability and posterior simulation without doing an in-depth discussion which would be beyond the scope of our work. To this end, the following changes have been made:

L409-L413: We describe the use of simulation from the empirical Bayesian posterior distribution to correct for the coverage of the across the function CI, in a similar fashion of the correction that *q-values* provide.

L485-493: We indicate here that *p-values* for GAMs do not have the same interpretation of effect size as in unlike linear models, but that we can use the model to estimate the instantaneous effect size by doing pairwise comparisons and constructing a simultaneous CI around the difference. Additionally, in Appendix B, Section B.5 we indicate that we use {gratia} to estimate the simulatenous and pointwise CIs.

Reviewer's Small Notes

L14: GAMs aren't really what permit missing data or other correlative structures, all of which can be incorporated into LMEMs. I see this is addressed down in section 3.4, but it is somewhat misleading here.

Reply to Reviewer's Comments

We have corrected the line above and now L11-L12 make clear that LMEMs can work with unbalanced data and non-constant correlation as well. Later, in L16-L17 we indicate that GAMs can also permit incomplete observations and different correlation structures. We believe this will clear any confusion about the capabilities of LMEMs.

Reviewer's Comments

L99: LMEMs do not need normally distributed or independent data, though non-normal hierarchical effects tend to require more advanced tooling.

Reply to Reviewer's Comments

L95: we have removed the part about normality, and we now only indicate that random effects need to be independent.

Reviewer's Comments

L179: I believe equation 4 need a term of random effect times time

Reply to Reviewer's Comments

Our random effect α_{ij} is just a random intercept for each individual, and therefore it does not vary over time, which is the general case of a LMEM we cover in the paper.

Reviewer's Comments

L303: I believe it is important to note here that GAMs do not extrapolate well either. While they don't suffer from steep slopes right at the end of the data, beyond the range of the data they only reflect the assumptions built into the basis functions, be that flat values or linear extrapolation of the slope.

Reply to Reviewer's Comments

We have incorporated the reviewer suggestion in Section 6 (Discussion), as we believe it is the appropriate place to briefly present the limitations of GAMs regarding extrapolation. In L552-L557 we have followed the basic structure of the reviewer's comment in order to indicate the extrapolation limitations from GAMs. The section of the paragraph now reads:

"This does not mean, however, that as any other statistical model GAMs do not have certain limitations. In particular, beyond the range of the data GAMs only reflect the assumptions built into the basis functions, be that flat values or linear extrapolation of the slope. Therefore, researchers need to be careful when using GAMs for extrapolating purposes. In addition, both polynomial and GAMs show higher variance in estimates near the boundary of the data, but additive models generally have less variance than polynomials."

We believe this change acknowledges the limitations of GAMs regarding extrapolation.

Reviewer's Comments

L342, Figure 2: I believe the Y axis of (C), "Penalized Basis Functions", should be ""Weighted Basis Functions".

Reply to Reviewer's Comments

The label on the y axis now reads 'Weighted Basis Functions'.

Reviewer's Comments

L360: The word "framework" is used twice here

Reply to Reviewer's Comments

We believe the reviewer refers to L331 in the original manuscript. In the revised manuscript the sentence now reads:

"The previous sections provided the basic understanding of the GAM framework and how these models..." (L421).

Reviewer's Comments

Finally, I was pleased to find I could reproduce all your results and re-render the results from the GitHub repository. I commend the authors on this good work.

Reply to Reviewer's Comments

We appreciate the dedication of the reviewer to ensure the reproducibility of our results.