Using GAMMs to model trial-by-trial fluctuations in experimental data: More risks but hardly any benefit

Rüdiger Thul
University of Nottingham

Kathy Conklin
University of Nottingham

Dale J. Barr
University of Glasgow

Abstract

Data from each subject in a repeated-measures experiment forms a time series, which may include trial-by-trial fluctuations arising from human factors such as practice or fatigue. Concerns about the statistical implications of such effects have increased the popularity of Generalized Additive Mixed Models (GAMMs), a powerful technique for modeling wiggly patterns. We question these statistical concerns and investigate the costs and benefits of using GAMMs relative to linear mixed-effects models (LMEMs). In two sets of Monte Carlo simulations, LMEMs that ignored time-varying effects were no more prone to false positives than GAMMs. Although GAMMs generally boosted power for within-subject effects, they reduced power for between-subject effects, sometimes to a severe degree. Our results signal the importance of proper subject-level randomization as the main defense against statistical artifacts due to by-trial fluctuations.

Studies including repeated measurements on individual subjects are extremely common in the social sciences. Because all the data for a single subject cannot be collected simultaneously, the set of observations for that subject will form a time series, and the full dataset a collection of such. By itself, this observation may seem trivial, but its statistical implication—non-independence over time—is not. Human subjects often fluctuate in their performance over the course of an experimental session, reflecting changing environmental, physiological, or psychological factors as a subject completes a task. The psychological

---

Corresponding author: Dale J. Barr, Institute of Neuroscience and Psychology, 62 Hillhead Street, University of Glasgow, United Kingdom, G12 8QB. Data, code, and instructions for reproducing the simulations are available through the project archive at https://osf.io/cp9z8. Thanks to Jonathon Holland for assisting with pilot simulations and a literature review, and to Mante Nieuwland and Christoph Scheepers for feedback on a draft of this manuscript.

literature has identified a broad range of factors that can give rise to time-varying effects, such as repetition of stimuli (Forbach, Stanners, & Hochhaus, 1974), task switching (Monsell, 2003), mental fatigue (Ackerman & Kanfer, 2009), mind wandering (McVay & Kane, 2009), and statistical learning (Jones, Curran, Mozer, & Wilder, 2013). Occasionally, these time-varying effects are phenomena of interest in their own right, but more often they are just treated as irrelevant and ignored.

By-trial fluctuations are seen as a problem because it is believed that datasets with such effects violate the assumption of independently and identically distributed errors underlying parametric statistical analyses. When such effects are ignored in the analysis, the residuals for each subject may show *temporal autocorrelation*: pairs of observations within a series are correlated (usually positively), with the correlation strength depending on the time lag (Baayen, Vasishth, Kliegl, & Bates, 2017). In traditional analyses using t-test and ANOVA, trials for each subject in each condition are usually aggregated into a set of means, and the analysis is performed on these means rather than on the trial-level observations. Although aggregation reduces the degrees of freedom, it is not immediately clear that it eliminates all concerns about non-independence, since the means themselves or their variances may be biased. In more modern analysis approaches, the need to simultaneously model crossed random factors such as subjects and stimuli (Baayen, Davidson, & Bates, 2008) precludes aggregation, thus exposing the analyst to the potential consequences of this non-independence. Later we will question the relevance of temporal autocorrelation for meeting statistical assumptions in most experimental contexts; but to fully understand the nature of these concerns, let us provisionally accept the premise.

To illustrate the potential problem, consider the contrived example in Figure 1, which shows simple response-time data from a single participant, fluctuating around a mean of 600 milliseconds (left panel). A sinusoidal effect such as this might arise through changing psychological factors during the experimental session. As the participant becomes familiar with the task, reaction time speeds up (trials 1 through 12), but then boredom and fatigue set in, gradually slowing responses (trials 12 through 36). As the end of the session comes into view, the subject speeds up in order to finish earlier (trials 36 through 48). A traditional linear mixed-model analysis fit to a collection of such data would be likely to include by-subject random intercepts, which would account for the mean height of the curve for each subject (the dashed line in the left panel), but would remain static over time. We could also envision an alternative analysis that captures the sinusoidal pattern in the data by incorporating a kind of time-varying random intercept (solid line). The latter model provides a better fit to the data, and also would remove the temporal autocorrelation. Temporal autocorrelation is usually diagnosed through an *autocorrelelogram* of the residuals (right panels), which plots the correlation coefficient between any two arbitrary time points in the series as a function of the lag between them.[1] (A lag of zero always has a perfect correla-

[1]The validity of autocorrelelograms for psychological data is questionable. Calculating correlations as a function of the time lag between observations is only valid when the underlying process has the mathematical property of *stationarity*; roughly, when the process is not itself changing over time. There is every reason to think that psychological data would *not* exhibit stationarity, since human subjects in an experiment are highly reactive to changing conditions (e.g., the learning, practice, task switching, and mind wandering effects cited above). Moreover, such plots are not fully diagnostic: it is possible to devise a non-stationary autocorrelated process that would yield a 'false negative' autocorrelation plot, i.e., one that suggests no

tion of one, because it is the observation's correlation with itself.) As can be seen in the autocorrelelogram for the static intercept model, failure to model the time-varying pattern has induced initial positive autocorrelation that decreases as a function of lag. In contrast, the model with the time-varying intercept has removed all temporal autocorrelation.

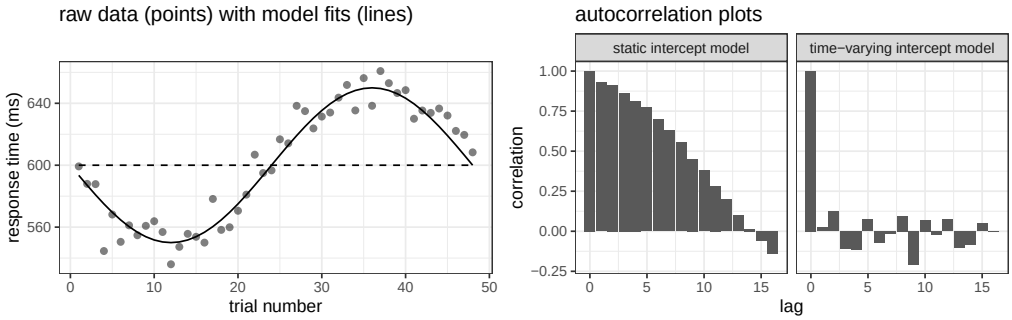

Figure 1. Hypothetical observations showing a sinusoidal pattern along with model fits (lines) and autocorrelation plots.

Models that ignore by-trial fluctuations may violate statistical assumptions, but with what consequences? There is a large statistical literature on temporal autocorrelation and potential remedies in time series analysis, with the typical message that failing to account for autocorrelation results in underestimation of standard errors, thereby inflating false positive rates for hypothesis tests, or equivalently, producing confidence intervals that are too narrow (e.g., Bence, 1995; Cochrane & Orcutt, 1949; Griffiths & Beesley, 1984). However, the statistical literature largely involves the analysis of just one or perhaps several time series—a common situation in economic forecasting or political polling—but very much unlike the situation in experimental studies with human subjects. In contrast, the literature on temporal autocorrelation in functional magnetic resonance imaging (fMRI) supplies examples with study designs and statistical considerations that are much closer to studies with human subjects. In experiments using fMRI, time-varying effects in the BOLD signal are generally seen as nuisance variation that will tend to inflate false positive rates if not taken into account (Purdon & Weisskoff, 1998). Proposed remedies include estimating and removing the autocorrelation parameter ('pre-whitening,' Bullmore et al., 1996; Woolrich, Ripley, Brady, & Smith, 2001) or 'coloring' the signal with known autocorrelation using temporal filtering (Worsley & Friston, 1995).

An alternative to these spectral approaches that is attractive for univariate situations is to simply estimate fluctuations as part of the model by including time-based predictors. For instance, one could include the polynomial effects of time (Mirman, 2016). To the extent the fluctuations are accurately modeled, this would remove the temporal autocorrelation from the residuals. However, determining the appropriate degree of the polynomial can be challenging, and fitting such models with the appropriate random effects structure can lead to convergence problems (Winter & Wieling, 2016). Also, polynomials perform poorly on patterns with discontinuities or asymptotes.

---

autocorrelation.

Examining data from three psycholinguistic datasets, Baayen and colleagues have presented compelling evidence for fluctuating performance over time and have shown how to use Generalized Additive Mixed Models (GAMMs) to model these effects. GAMMs are different from polynomial models in that they can represent arbitrary wiggly patterns in data as the sum of a set of mathematical *basis functions*, with the complexity of these functions determined by cross-validation on the data itself or by Bayesian techniques (Wood, 2017). As with a linear mixed-effects model, the analyst can specify random intercepts and random slopes for crossed random factors.

Baayen and colleagues depart from traditional approaches to analysis by construing these fluctuating effects not as irrelevant variation but as *nuisance variation*—that is, variation that is not necessarily of interest in itself, but that must be taken into account in order to obtain precise and unbiased estimates of parameters of interest. They take it for granted that ignoring such effects is harmful: "if the errors indeed show autocorrelational structure, evaluation of the significance of predictors in the model becomes problematic due to potential anti-conservatism of p-values" (Baayen, van Rij, de Cat, & Wood, 2018, p. 49). Anti-conservatism refers to the situation where p-values are smaller than they should be, corresponding to an increase in the rate of false positives. We see nothing in the work of Baayen and colleagues to support this assertion.

In contrast, the statistical and fMRI literatures mentioned above might be taken as support for such concerns about anti-conservativity. However, upon closer consideration, the warnings in these literatures may not be fully relevant. As noted above, statistical discussions of time-series analyses often focus on datasets including one or only a few time series, rather than many, and in scenarios where change over time almost always represents variation of interest. Also, the time series are usually observational (e.g., economic or political time series), where the timing and presentation order of interventions is not under control of the researcher, unlike the experimental context where randomization and counterbalancing can remove any confounding of effects of independent variables with fluctuations in performance, likely neutralizing the kinds of problems found in other types of studies. Finally, although addressing autocorrelation is seen as imporant in fMRI analysis, traditional fMRI studies often used 'boxcar' designs that lack true randomization over time: trials from a given experimental condition are often grouped together in 'epochs' because of the slowly evolving hemodynamic response underlying the BOLD signal (Amaro & Barker, 2006). By their very nature, boxcar designs confound treatment effects with fluctuations in the dependent variable. For fMRI studies using event-related designs with true randomization, false positives may be less of a concern than false negatives (Olszowy, Aston, Rua, & Williams, 2019).

Modeling fluctuations in performance may provide additional insight into data, but the nuisance variation perspective suggests that such modeling is not optional. To the extent that modeling this variation is important for inference, then re-analyses of datasets with GAMMs should lead to substantively different conclusions from original analyses that treat such variation as irrelevant. In one re-analysis by Baayen et al. (2017), a few effects that were significant in the original analysis no longer were in the GAMM analysis. This might seem to support the nuisance perspective; however, because the ground truth is unknown,

such differences in outcome could either reflect false positives from LMEMs or false negatives from GAMMs.

In this paper, we contend that the use of GAMMs to model by-trial fluctuations in experiments with randomized presentation is generally unnecessary and potentially harmful. Traditional LMEMs show reasonable Type I error rates in the face of temporal autocorrelation, so long as the presentation order has been appropriately randomized. And GAMMs are not only harder to implement and interpret compared to LMEMs, but they may also impair power for between-subject effects.

We begin by explaining the use of GAMMs in multi-level data, focusing on the use of *factor smooths* to account for time-varying patterns in individual subjects' data. Then, a brief thought experiment questions the necessity of accounting for these patterns to satisfy statistical assumptions. Next, Monte Carlo simulations examine the performance of GAMMs and LMEMs across a range of theoretical patterns, including sinusoidal and random-walk Gaussian patterns. The simulations support the thought experiment: outcomes for models ignoring temporal autocorrelation are statistically indistinguishable from outcomes of models fit to data with comparable residual variation but no temporal autocorrelation, with false positive rates close to nominal, so long as the presentation of within-subject levels is truly randomized over time. It is only in designs with blocked presentation order of within-subject levels that LMEM models perform poorly relative to GAMMs, due to their inability to deconfound treatment variation from time-varying effects. In designs with fully randomized presentation order, GAMMs sometimes offer a modest boost to power for within-subject effects, but usually at the cost of increased Type II errors for between-subject effects.

Next, due to concerns about the external validity of our theoretical simulations, we undertook a second set of simulations modeled upon real data, namely the Stroop task from the Many Labs 3 mega-study (Ebersole et al., 2016). Here, we simulated data based on estimates of fixed and random effects from the main study, 'grafting' on real residuals from randomly chosen subjects. Results were similar to the theoretical simulations: LMEMs show no evidence of inflated false positives, GAMMs modestly boost power for within-subject effects, but impair power for between-subject effects under some circumstances.

The bottom line from these simulations is that concerns about inflated Type I error rates for LMEMs applied to experimental data that include trial-by-trial fluctuations are completely unwarranted, so long as the presentation order of stimuli and experimental conditions have been properly randomized. If model residuals show signs of autocorrelation, this should not be cause for alarm, since proper randomization and counterbalancing is already a sufficient remedy. GAMMs may modestly increase power for within-subject effects in these situations, but given their complexity, and the potential costs to power for between-subject effects, their use should be seen as optional rather than mandatory. It is only where true randomization and counterbalancing of presentation order is absent or imperfect that trial-by-trial fluctations become a dangerous nuisance. In these circumstances, GAMMs (or other time-series modeling) are likely to be useful to deconfound variation of interest from time-varying noise.

While we question the need to model by-trial fluctuations in fully randomized experiments,

we do not doubt the utility of GAMMs in data exploration, or in situations when the variable of time is of theoretical interest. Thus, longitudinal studies, such as in cognitive development or language evolution, are exempt from our recommendations. We focus on studies where each subject's data is a series of trials with just one observation per trial. Our recommendations therefore may not generalize to studies where multiple observations are sampled during each trial or stimulus, as is the case in mouse-tracking, visual-world eyetracking, MEG, EEG, and pupillometry studies, except when the data from each trial has been reduced to a single summary statistic. We do not question the applicability of GAMMs for investigating the time-course of effects at the trial level, such as demonstrated by van Rij, Hendriks, van Rijn, Baayen, and Wood (2019). In short, there is still a wide range of situations in which the use GAMMs may be advantageous. Even so, users should be cautious about the potential dangers we document here.

How to model time-varying nuisance effects with GAMMs

In this section, we illustrate and explain those features of Generalized Additive Mixed Models that are most relevant to modeling fluctuations over time. For a more in-depth tutorial for psychological or linguistic data, see Baayen et al. (2017), Baayen et al. (2018), Sóskuthy (2017), and Winter and Wieling (2016). The textbook by Wood (2017) provides a more comprehensive, technical treatment. Our investigation centers on four main considerations in dealing with data containing by-trial fluctuations: (1) the form the pattern takes over time, and whether individual patterns share common structure or are completely idiosyncratic; (2) the consequences of the temporal structure with respect to model assumptions; namely, whether the pattern only violates independence assumptions or whether it additionally violates the assumption of normally distributed residuals; (3) whether the presentation of the levels of any within-subject factors are randomized over time or blocked; and (4) whether the independent variables under investigation are administered as between-subject or within-subject factors.

Let us start by considering a contrived dataset in which we assume that subjects' responses exhibit a sinusoidal pattern such as that shown in Figure 1. Using the `{autocorr}` package for R that accompanies this manuscript (https://github.com/dalejbarr/autocorr), we can simulate a dataset with by-trial fluctuations using the `sim_2x2()` function. This function simulates data from a 2x2 mixed design with one within-subject factor ('A') and one between-subject factor ('B'), and allows the user to explore different time-varying patterns. The resulting dataset has variables representing either a randomized or blocked design. We will start with the randomized design.

```
## devtools::install_github("dalejbarr/autocorr")
library("autocorr")
library("mgcv")
library("tidyverse")

set.seed(62) # for reproducibility
dat <- sim_2x2(int = 3,  # set the intercept to an arbitrary non-zero value
         version = 2L) # form of the time varying effect; see ?errsim
```

This gives us a dataset with simulated data from 48 subjects, with 48 trials per subject, and where individual subjects show a sinusoidal pattern, much like in the original example above. The `{mgcv}` package (Wood, 2017) provides functions for fitting GAMMs. Although the functions are designed for fitting data with 'smooth' terms to capture wiggly patterns in the data, they can also be used to fit a generic linear mixed-effects model (LMEM) if standard random effects and no smooth terms are specified. For comparability with GAMM models, we will fit standard LMEMs using `{mgcv}` functions instead of using `{lme4}` . We can use either the `gam()` or `bam()` functions, which are similar, except the latter has been optimized for large datasets.

Given the design includes one within-subject factor with multiple observations per level per subject, an appropriate LMEM model for these data would include by-subject random intercepts and slopes for the within ('A') factor.

```
mod_lmem <- gam(Y_r ~ A_c * B_c +
      s(subj_id, bs = "re") +     # by-subject random intercept
      s(subj_id, A_c, bs = "re"), # by-subject random slope for A_c
    data = dat)
```

The above syntax models the dependent variable `Y_r` in terms of an (implicit) intercept plus main effects of the within-subject factor (`A_c`) and the between-subject factor (`B_c`). The by-subject random intercepts and slopes are included using the `s()` function with the option `bs="re"`.[2]

How can an analyst detect the presence of by-trial fluctuations in data? Although it is useful and convenient to simply look at each subject's raw data, for the purpose of checking assumptions it is essential to look at model residuals. Autocorrelations that appear in the raw data may be absent from the residuals of an appropriately specified model, and may lead the analyst to pursue unnecessary remedies that may cloud interpretation or otherwise harm inference (Huitema & McKean, 1998). Let us plot the residuals as a function of trial number (`tnum_r`) for the first four subjects.

It is apparent from Figure 2 that we have time-varying effects that have not been accounted for. Also, the residuals do not appear to be normally distributed; in fact, we prove in the Appendix that in a hypothetical scenario where all subjects show variations on a basic sinusoidal pattern, the residuals from a LMEM will always depart from normality.

The GAMM approach for resolving these problems of non-independence and non-normality would be to add certain 'smooth' terms to the model. We can start by changing the static intercept to a time-varying intercept, by estimating a smooth function for time (in this case, indexed by trial number). However, if we stopped there, this would assume that the sinusoidal pattern is identical for all subjects. So we can add *factor smooths* to the model, which captures any leftover wigglyness for each subject after accounting for the time-varying fixed intercept. The wigglyness of these smooth terms is determined by a cross-validation procedure or by Bayesian techniques to prevent overfitting (Wood, 2017). Although the

[2] Unlike a standard LMEM fit using `{lme4}`, random intercepts and slopes in a GAMM are specified individually and treated as independent.

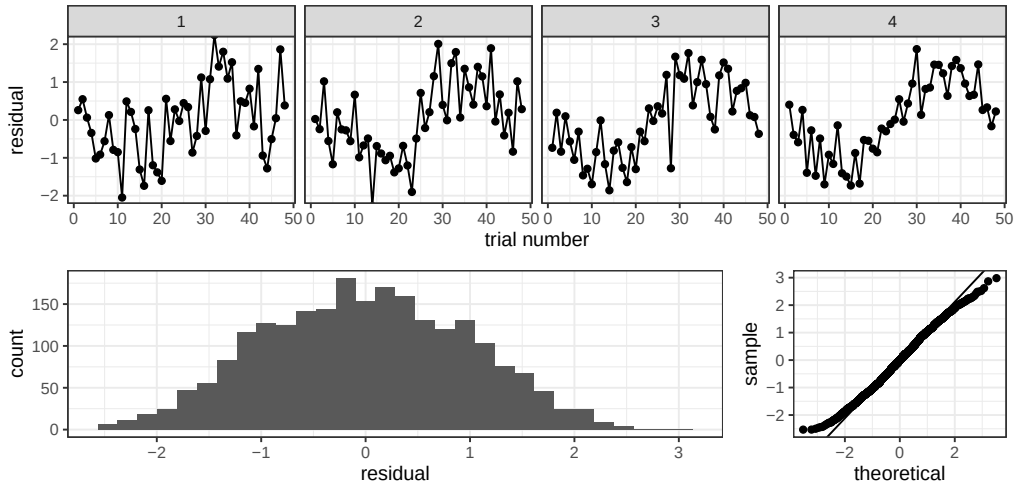

Figure 2. Residuals from a linear-mixed effects model. Top row: residuals plotted by time. Bottom row: histogram and Q-Q plot.

user can control many aspects of the estimation process (see the `{mgcv}` vignettes and documentation), we will just use the defaults.

```
mod_gamm <- gam(Y_r ~ A_c * B_c +
      s(tnum_r) + # time-varying fixed intercept
      s(subj_id, tnum_r, m = 1, bs = "fs") + # time-varying random intercepts
      s(subj_id, A_c, bs = "re"), # by-subject random slope for A_c
    data = dat)
```

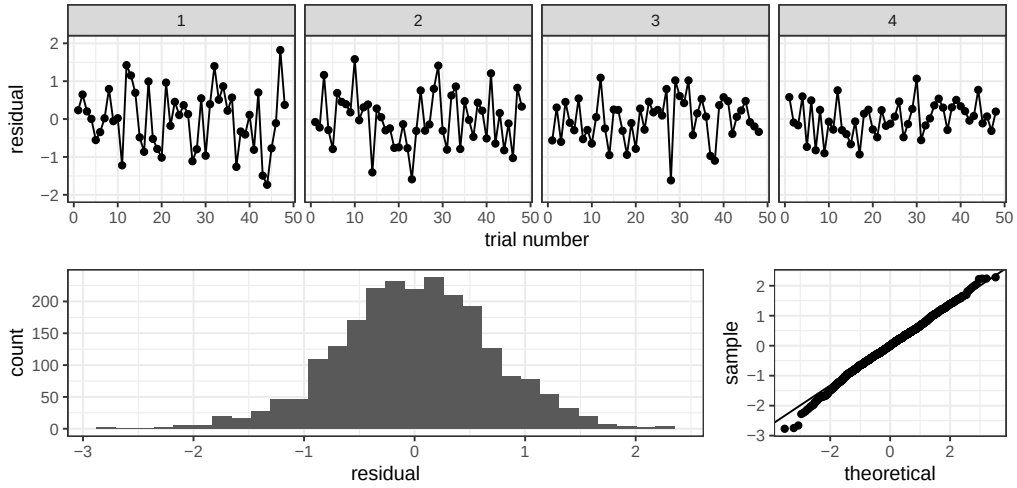

Figure 3. Residuals from a generalized additive mixed model. Top row: residuals plotted by time. Bottom row: histogram and Q-Q plot.

The `s(tnum_r)` term adds a fixed time-varying intercept for the time predictor `tnum_r`, estimated with the default "thin plate" basis functions. The `s(subj_id, tnum_r, m = 1,`

`bs = "fs")` term specifies factor smooths, which allows additional subject-specific wiggliness over time. The `m = 1` argument specifies a penalty to the first basis function, which is completely smooth. This ensures the factor smooths will behave like proper *random smooths*; that is, like a typical random effect rather than a fixed effect (Baayen et al., 2017). We kept random slopes for the within-subject factor (`A_c`) in the model, because the factor smooths function as a time-varying intercept only, and do not capture random slope variation. Checking the residuals from this model, (Figure 3), we see that the smooth terms have removed the temporal autocorrelation, and made the residuals look more normally distributed.

A limitation of the above model is that it treats the within-subject effect as static over time. It is possible that the effect varies over time, and that different subjects show distinct time-varying patterns. The model could be further enriched to capture such effects. However, the simpler model turns out to be adequate for our simulated data, and may also generally be so for real data; indeed, the models fit by Baayen et al. (2017) only included time-varying intercepts with conventional random slopes, and nonetheless proved effective in removing temporal autocorrelation.[3]

In the current scenario, the fluctuations for each subject come from a simple variation on a common sinusoidal pattern. When a model ignores these patterns, the residuals end up autocorrelated and non-normally distributed. In our simulations, we considered additional scenarios where the time-varying patterns were unique to each subject, as well as patterns that yielded autocorrelated residuals that were normally distributed.

Design considerations

The impact of autocorrelation on model performance is likely to depend not only on the pattern of autocorrelation (e.g., degree of common versus idiosyncratic structure) but also on how it relates to the experimental design. Ideally, in experimental studies, the presentation order of within-subject factors is randomized independently for each subject. This subject-level randomization ensures the temporal fluctuations will not be systematically confounded with any variation of interest, possibly making it safe to ignore them. It seems likely that modeling these effects with GAMMs could boost power, but it is unclear whether there are any hidden costs to this approach.

Occasionally, researchers use a counterbalanced blocked presentation instead of randomized presentation; for instance, half of the participants may receive the first half of trials in condition A1 and the second half in condition A2, while the other half gets them in the contrary order. In blocked designs, it may be problematic to ignore temporal fluctuations. Although counterbalancing across subjects should keep the false positive rate in check, power could be reduced if the variation of interest gets swamped by the nuisance variation. GAMMs or other types of time-series models might prove especially valuable for separating signal from noise.

As already discussed, GAMMs may increase power for within subject effects, but it is

[3] For further discussion of specifying random effects with GAMMs, see van Rij et al. (2019).

unclear how they will perform relative to LMEMs for between-subject effects. In a linear mixed-effects model, variation associated with between-subject treatment effects is likely to be masked by random intercept variation; for GAMMs, it is masked by variation captured by the factor smooths. It is important to verify whether GAMMs with factor smooths perform as well as LMEMs at distinguishing signal from noise.

Irrelevant or nuisance variation?

We have seen how temporal fluctuations can be accounted for by GAMMs but up to now we have deferred answering a more basic question: Is it even necessary to account for these fluctuations? In other words, should such effects be considered as nuisance variation, as Baayen and colleagues evidently view them when emphasizing the importance of 'cleaning up' autocorrelation, or are they irrelevant? We contend that the latter view is appropriate for randomized experimental data. Our Monte Carlo simulations confirm this contention, but let us first argue the case analytically.

Imagine you are simulating data from an experiment with a single two-level within-subjects factor. For each of 48 subjects, you simulate data from 48 trials, 24 from each of two conditions ("main" and "control"), appearing in a random order.[4] You generate observations Y_{ij} for subject i on trial j according to the linear model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + S_i + e_{ij}$$

where X_{ij} is a categorical indicator variable for condition, S_i is the random intercept for subject i with $S_i \sim N(0, {\sigma_s}^2)$, and e_{ij} is the error for a particular trial, $e_{ij} \sim N(0, {\sigma_e}^2)$. Let us assume a further variable t_{ij} which indexes trial number; note that this variable is not represented anywhere in the generative model. We can obtain residuals by fitting a random-intercepts model to the simulated data.

Estimated residuals for three subjects are plotted against time (Figure 4, panel A) next to autocorrelation plots for these same subjects (panel B). Because the errors come from a normal distribution and have been generated without reference to the trial index t_{ij}, there is no temporal pattern in the estimated residuals, nor is there any autocorrelation.

Simulating the data puts you in the privileged position of knowing the error values ($e_{ij}s$) behind each observation, and you can exploit this knowledge to induce temporal autocorrelation simply by re-ordering the t_{ij}s, as if you had collected the observations in a different order. For example, you could re-define t_{ij} to follow the size of the errors in descending order, such that trial 1 is assigned to the observation with the largest e_{ij} value, trial 2 with the next largest, and so on. As would be expected, the residuals from a random intercept linear mixed-effects model for these data form a descending pattern, and we now see strong autocorrelation in the data (Figure 4, panels C and D). This plot now suggests that the data violate independence assumptions, but only the t_{ij}s have changed, and these play no role in the model. Each subject has the exact same set of observations as in the original scenario;

[4] R code for this demonstration is included in the project repository.

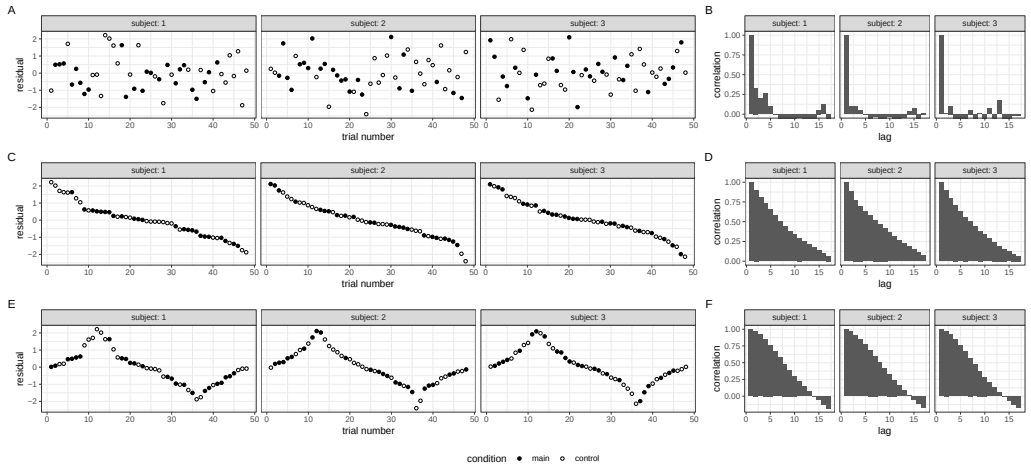

Figure 4. Estimated model residuals by time from original data and autocorrelation plots (A and B); model residuals by time from data where trial number is assigned based on the value of the error values in descending order (C and D) or arranged to form a sawtooth pattern (E and F). The open an closed symbols represent the two levels of a within-subject factor.

so, from the point of view of estimation, *you are fitting exactly the same model to exactly the same data.* Indeed, the model would give identical results under *any possible reordering of the trials*; for instance, for panels E and F you have re-defined the t_{ij}s to form a sawtooth pattern, yet the same parameter estimates are obtained (Table 1). Put differently, what the model sees for each subject is just a collection of unordered observations. You can change the values of the t_{ij}s variable to anything you please—you could even go so far as to delete the variable from the data—because they are wholly irrelevant from a modeling standpoint.

model	$\hat{\beta}_0$	SE($\hat{\beta}_0$)	$\hat{\beta}_1$	SE($\hat{\beta}_1$)
original order	-0.01712	0.10303	0.03347	0.04167
trials ordered by residual	-0.01712	0.10303	0.03347	0.04167
trials ordered into sawtooth	-0.01712	0.10303	0.03347	0.04167

Table 1
Parameter estimates and standard errors from the three models.

Keeping in mind that we are focused on experiments with univariate data and random presentation order, observing temporal autocorrelation in residuals from models where time does not appear as a variable should not be cause for concern about violating independence assumptions. If temporal autocorrelation appears in the residuals and the analyst chooses to ignore it, it is wrong to view the analyst's model as statistically unsound on that basis alone. They are non-independent only with respect to a variable that is invisible to the model. It is more accurate to view the analyst as having foregone the opportunity to obtain more precise estimates by incorporating the temporal structure in the model. Opting out from doing so can be a rational choice: GAMMs are technically challenging, they yield complex output that can be difficult to interpret, and their potential pitfalls are not well

known. Consequently, using them does not guarantee better insight into data.

Lest it still seem 'wrong' or 'inappropriate' for an analyst to ignore residual autocorrelation when analyzing data from a randomized experiment, let us attempt to further dislodge this notion by considering yet another way in which experimental data form a time series. Just as the individual measurements taken from a given subject are influenced by the characteristics of the particular moment at which the measurement is taken, so is the overall performance of a subject influenced by the time of day at which testing occurs, with subjects often showing more efficient performance on tasks during afternoons compared to mornings (Blake, 1967; Kleitman, 1963). Time of day fluctuations are very likely to induce residual autocorrelation, at least for tasks that are minimally cognitively demanding. In a study where you attempt to predict participants' mean performance by experimental group, this autocorrelation could be seen by plotting the residual for each subject against the time of day at which the session took place. Yet despite this variable being a likely source of non-independence, and despite it being a variable that is recorded in every experiment (if only by computer timestamps) it is almost always ignored during analysis. Why does no one question this?

We do not worry about time-of-day effects for the same reason we need not worry about time-of-trial effects: because we randomized. Randomization prospectively guards against the contamination of variance we care about by variance that we don't. Even beyond time of trial or time of day effects, in any study there is a potentially infinite number of possible variables for which residuals, when plotted against the variable values, would show non-independence—day of the week, season of the year, temperature of the room, degree of ambient visual or acoustic noise, subject conscienciousness, visual angle subtended by stimuli relative to each subject's visual acuity, and so on. So long as we have randomized, we don't need to worry about non-independence relative to these variables. Trial-by-trial fluctuations in performance are no different.

We have argued that modeling trial-by-trial fluctuations is not necessary, but we haven't answered the question of whether doing so is worthwhile. The rest of this article describes two sets of Monte Carlo simulations aimed at illuminating the potential costs and benefits of GAMM modeling so that analysts can be more informed in their decisions. The first set of simulations is meant to illuminate properties of GAMMs and LMEMs on data with time-varying effects by challenging them with a variety of patterns, including sinusoidal and Gaussian random walk patterns. The second set of simulations examines more realistic patterns, where the residuals from the simulated data contain real practice effects from a model fit to Stroop task data from the Many Labs 3 project (Ebersole et al., 2016).

Simulation Set 1: Sinusoidal and Random Walk Patterns

Method

Different approaches for dealing with autocorrelation might impact within-subject factors in a different way from between-subject factors. Thus, for our hypothetical experiment of interest we chose a mixed 2x2 design, including one within-subject factor, one between-subject factor, and their interaction. For simplicity, the design included no stimulus effects.

Each hypothetical experiment comprised 48 time-series, each representing data from a single hypothetical participant. The `sim_2x2()` function in the accompanying `{autocorr}` generated the datasets used in the Monte Carlo study, according to the following procedure.

The Y_{ij} observations for participant i on trial j were generated according to the multi-level linear model

$$Y_{ij} = \beta_0 + S_{0i} + (\beta_1 + S_{1i})A_j + \beta_2 B_i + \beta_3 A_j B_i + e_{ij}$$

where A_j is a deviation-coded (-.5, .5) predictor representing the level of the within-subject factor for trial j, B_i is a deviation-coded predictor representing the level of the between-subject factor for subject i, and the e_{ij}s comprise a 48-length vector of residuals. The by-subject random intercept and random slope for participant i, $\langle S_{0i}, S_{1i} \rangle$, were drawn from the bivariate normal distribution

$$\langle S_{0i}, S_{1i} \rangle \sim N(\, \langle 0, 0 \rangle \,, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} {\tau_0}^2 & \rho\tau_0\tau_1 \\ \rho\tau_0\tau_1 & {\tau_1}^2 \end{pmatrix}.$$

The `sim_2x2()` function generated sample datasets based on the above equations, with specific fixed-effects parameter values $\beta_0, \beta_1, \beta_2, \beta_3$, random-effects parameter values τ_0, τ_1, ρ, and residuals (as defined in the next section). The intercept β_0 remained fixed at zero for simplicity. We considered six unique values for the main-effect of the within-subject factor, $\beta_1 \in \{0, .05, .10, .15, .20, .25\}$: . We also considered six unique values for the effect of the between-subject factor, $\beta_2 \in \{0, .10, .20, .30, .40, .50\}$, as well as for the interaction, $\beta_3 \in \{0, .10, .20, .30, .40, .50\}$: . These values were determined through trial and error to yield a good range of power for effects in the "no autocorrelation" (baseline) scenario, given the ranges for the parameters defining the variance components. The values for β_1, β_2, and β_3 were not varied independently: whenever β_1 was set to the nth value in the series (e.g., if $n = 4, \beta_1 = .15$), then β_2 and β_3 were also set to their nth values (e.g., $\beta_2 = \beta_3 = .30$). This seemed like a reasonable way to reduce the number of simulations required, since all effects are independent and the design was always balanced.

Parameters for the variance components that define the random effects should ideally be chosen to approximate the ratio of subject-level to trial-level noise found in real data. Since data on this ratio are lacking, we derived values from the convenience sample of data from 13 psycholinguistic experiments provided by Barr, Levy, Scheepers, and Tily (2013) in their online appendix. The estimated variance components from these studies, expressed as a proportion of residual variance, are shown in Table 2 and are also available as the object `blst_studies` in `autocorr`. Each simulated dataset had parameters τ_0 and τ_1 drawn from a uniform distribution spanning from the 20th to the 80th percentiles of the corresponding empirical distribution, specifically $\tau_0 \sim U\,(0.105, 0.420)$ and $\tau_1 \sim U\,(0.001, 0.261)$. The

ID	σ	τ_0/σ	τ_1/σ
1	3572	0.216	0.000
2	8439	0.140	0.001
3	24388	0.105	0.000
4	29934	0.321	0.000
5	0.494	1.877	0.385
6	275363	0.449	0.002
7	230191	0.106	0.059
8	231824	0.172	0.003
9	51372	0.412	0.021
10	7536625	0.036	0.066
11	406043	0.229	0.222
12	0.128	0.104	0.366
13	242830	0.426	0.287

Table 2
Estimated variance components from the sample of 13 psycholinguistic studies in Barr et al. (2013).

random correlation ρ was also from a uniform distribution, reflecting a range of realistic values: $\rho \sim U(-.8, .8)$.

For each sample dataset generated by `sim_2x2()`, the residuals for all 48 participants either had no autocorrelation (baseline scenario), or had the same autocorrelation structure representing one of the eight scenarios described below. All samples were homogeneous with respect to the autocorrelation structure. That is, the residuals for all subjects within a sample were generated according to the same selected scenario; we did not mix scenarios within a sample.

The scenarios of autocorrelated residuals were chosen to represent a variety of theoretically interesting scenarios, from highly consistent and structured sinusoidal patterns to highly idiosyncratic and unstructured 'random walk' patterns. Because real data is likely to have multiple sources of autocorrelation operating at distinct time scales, we also considered scenarios that were a mixture of the sinusoidal and random walk patterns. Figure 5 shows example sequences from each scenario. The residuals were generated using the `errsim` function of **`autocorr`**. In all scenarios, we normalized each residual time series so that the mean was 0 and the standard deviation was 1. We also compare the autocorrelation scenario to a baseline scenario without autocorrelation (also with $SD = 1$). Thus, the total variation is identical across the scenarios; all that varies is how this variation is distributed over time.

Scenario 1 and 2: Sinusoidal. The first two scenarios represent relatively slow processes unfolding over the course of an experimental session, such as fatigue or practice effects. For simplicity, we represent such processes as a single bandwidth of a sinusoidal function. A real-world example might be a reaction time experiment where a participant speeds up response times at the beginning of a session as they gain experience with the task, followed by a gradual slowing due to fatigue, and ending with a final speed up as the participant

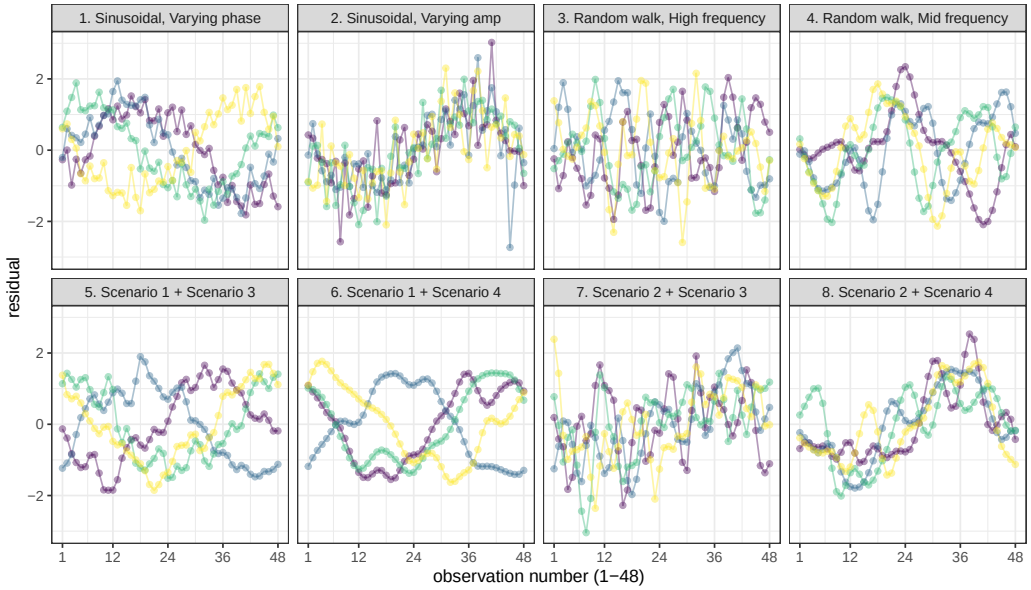

Figure 5. Randomly generated sequences exemplifying the eight autocorrelation scenarios. Each line represents the residuals for a single hypothetical participant.

rallies toward the end. Another participant might show an opposite (anti-phased) pattern, gradually slowing at the start as they learn how to maximize accuracy, followed by a gradual speed with practice, and slowing towards the end as fatigue sets in.

Reflecting this type of scenario, the first scenario we consider is one in which participants exhibit a common sinusoidal form offset by random phase angles, which we call the *varying phase* scenario. In these simulations, the phase angle g_i for the ith participant was drawn from a uniform distribution, $g_i \sim U(-\pi, \pi)$. Each simulated participant having a different random phase angle implies no time-varying structure at the population level, since the expected value of the sum of a set of sine waves with fixed amplitude and frequency but with phase offsets randomly drawn from a uniform distribution is a flat line (i.e., the waves tend to cancel one another out).

Of course, we also assume the presence of noise (without any temporal autocorrelation) superimposed upon this overall pattern. Specifically, for this first scenario, we assume 90% of the total variance is driven by the sinusoidal pattern and the remaining 10% is trial-level noise.

It may be too extreme to assume no structure to the autocorrelation pattern at the population level. Thus, our second *varying amplitude* scenario considers the opposite extreme: What if everyone showed the exact same pattern, but with varying strength? For Scenario 2, the residuals for each participant were represented as a sinusoidal pattern with frequency and phase fixed, but with amplitude determined by A_i for the ith participant drawn from a uniform distribution, ${A_i}^2 \sim U(.2, .8)$ (Figure 5, second panel of the first row). Between 20% to 80% of the residual variance is driven by the sine wave signal, and the remaining variance is Gaussian noise ($\sigma^2 = 1 - {A_i}^2$). Here, rather than canceling, the phase-locked

patterns combine to yield a population-level effect.

Scenarios 3 and 4: Random walk. The two above scenarios assume that the time-varying patterns on the dependent variable have a common sinusoidal form across participants. But it is also of interest to consider scenarios where the functional forms are idiosyncratic. To this end, we randomly generated time series from an autocorrelated Gaussian process, following the numerical method developed in Shinozuka and Deodatis (1991), which we implemented in the `stat_gp()` function in the **autocorr** package. Technical details are provided in the Appendix. The function takes two arguments, σ and γ, determining the standard deviation and the correlation length, respectively. For Scenarios 3 and 4, σ was fixed at 1, while γ was 1 for Scenario 3, and 2 for Scenario 4, such that the oscillatory patterns in Scenario 3 were higher frequency than in Scenario 4 (see the third and fourth panels in the top row of Figure 5). Because we are interested in patterns occurring on multiple time scales, these two values of γ were selected so that the resulting oscillation frequencies would be higher than the sine wave oscillations considered in Scenarios 1 and 2. We refer to Scenario 3 as "high frequency" and Scenario 4 as "mid frequency" random walks, in contrast with the lower-frequency sine waves.

Note that the idiosyncratic nature of these time series implies that, like the varying phase scenario described above, the series will tend to cancel across subjects such that the dependent variable would show no population-level effect.

Scenarios 5–8: Mixed timescales. The four remaining scenarios (Scenarios 5–8) reflect time series with autocorrelation occurring on the slow, sinusoidal time scale as well as on the faster, random walk time scales. Each scenario is simply the sum of one of the two sine wave scenarios and one of the two random walk scenarios, as described below.

Scenario 5 and Scenario 6 had a the varying-phase sine wave (Scenario 1) mixed with the high- and mid- frequency random walk patterns, respectively. Each vector of residuals reflected a mix of 90% of a randomly generated sine wave pattern with 10% of the corresponding random walk, with results exemplified in the first two panels of the second row in Figure 5.

Scenario 7 and Scenario 8 had the varying-amplitude sine wave (Scenario 2) mixed with the high- and mid- frequency random walk patterns, respectively. As described above for Scenario 2, the amplitude A_i for the ith participant was determined by $A_i^2 \sim U(.2, .8)$, with the remaining random walk variance scaled to comprise $1 - {A_i}^2$ of the total variance.

Analysis. In contrast to the uncertain position of the analyst, a researcher working with simulated data is in the privileged position of having complete knowledge of the process giving rise to the data. This makes it possible to analyze data either from the uncertain perspective of the analyst or from the omniscient perspective of the designer of the simulation. Because we were interested in the performance of GAMMs under ideal circumstances, we performed our analyses from the latter perspective: the GAMM models that we fit exactly matched the generative process.

We created the **autocorr** function `fit_2x2` to fit models to the data generated by `sim_2x2`. The function fits a GAMM as well as an LMEM style model using the `bam` function from the `{mgcv}` package, version 1.8.31 (Wood, 2011).

The model formula for fitting LMEM-style models in `bam` was

```
Y ~ W * B + s(id, bs = "re") + s(id, W, bs = "re")
```

which is formally equivalent to the `{lme4}` formula

```
Y ~ W * B + (1 + W || id)
```

where the 'double bar' `||` syntax in `(1 + W || id)` fixes the covariance between random intercepts and slopes to zero. The fixed part `W * B` specifies main effects for the within-factor `W` and between-factor `B` and the interaction between them. The 'smooth' terms `s(.., bs="re")` specify standard random effects where `id` is the subject identifier, with `s(id, bs="re")` specifying the random intercept and `s(id, W, bs="re")` specifying the random slope (i.e, allowing the effect of W to vary over subjects).[5]

For the three scenarios with the underlying phase-locked but varying amplitude sine pattern (scenarios 2, 7, and 8) the full GAMM model formula was

```
Y ~ W * B + s(t, bs="tp") + s(t, id, bs="fs") + s(id, W, bs="re").
```

In this formula, the `s(id, W, bs="re")` term, which also appears in the LMEM formula, specifies a random slope for the within-subject factor. The `s(t, bs="tp")` term specifies the (default) 'thin plate' smooth intended to capture the part of the sinusoidal pattern that varies over time `t` and is common to all subjects. The `bs="fs"` argument in `s(t, id, bs="fs")` term specifies a factor smooth, allowing the time-varying pattern for each subject to diverge from the overall pattern. Note that the factor smooth plays the role of the random intercept in the LMEM formula, with the difference of allowing the intercept to vary over time. The default behavior for factor smooths is to behave like a fixed effect, and so some authors have advised specifying a penalty to the linear basis function for factor smooths (`m = 1`) so that they behave more like random effects (Baayen et al., 2017, p. 211). For simplicity, we refer to this as a GAMM with a "penalized" factor smooth, and the previous version as "unpenalized."

```
Y ~ W * B + s(t, bs="tp") + s(t, id, m=1, bs="fs") + s(id, W, bs="re").
```

We included results for both the unpenalized and penalized factor smooth versions because we assume that many users will be unaware of the advice and just rely on function defaults.

The GAMM formula for the baseline "no autocorrelation" scenario as well as for the five remaining scenarios was

```
Y ~ W * B + s(t, id, bs = "fs") + s(id, W, bs = "re")
```

for the unpenalized version, and

```
Y ~ W * B + s(t, id, m = 1, bs = "fs") + s(id, W, bs = "re")
```

[5] Variable names in the text have been simplified for expository purposes and do not match the names in the datasets resulting from `sim_2x2`. The exact formulas used to fit models can be obtained by running `fit_2x2(NULL, cs = TRUE, dontfit = TRUE)` for scenarios 2, 7, and 8 and `fit_2x2(NULL, cs = FALSE, dontfit = TRUE)` for the remaining scenarios.

for the penalized version. Both forms are the same as above except the fixed smooth term, `s(t, bs="tp")`, has been omitted, since these patterns have no common time-varying structure across participants. We used the Wald z statistic to derive p-values for fixed effects, defined as the ratio of the parameter estimate to its estimated standard error.

An important aspect of GAMM modeling involves setting the type and number of basis functions. We opted to stick with the defaults provided by the `{mgcv}` package, reasoning that this is what a typical user would do. This implies that the smooth terms introduced by `s(t)` will use about 10 basis functions (see `?choose.k` and `?tprs` for details). This seems sufficiently large given there are only 48 observations in each time series.

Software. We ran the simulations and analyzed the results using R version 3.6.2 (R Core Team, 2019). Simulations were run using the function `mcsim()` in the R `{autocorr}` package (https://github.com/dalejbarr/autocorr). We bundled all necessary software infrastructure in a Singularity 3.5 software container (https://sylabs.io/singularity). The simulations can be reproduced by the command

```
singularity run library://dalejbarr/talklab/autocorr
```

which activates the shell script `acsim` inside the container. The script's default action is to invoke `mcsim()` to run 1,000 Monte Carlo simulations at each of the 54 unique combinations of Monte Carlo parameters (six effect size settings times nine scenarios, including the baseline scenario).

Results and Discussion

We completed 25,000 simulation runs at each of the 54 unique parameter settings (six effect size settings times the nine scenarios including baseline). Results from the simulations are available in the project repository at https://osf.io/cp9z8, which also includes source code files for reproducing analyses and for compiling this manuscript. The raw results are available in the `data_derived` subfolder of the repository. The repository also contains results from supplementary simulations looking at other time-varying patterns as well as additional AR(1) modeling for some scenarios, and provides instructions on how to run further simulations based on user-defined functions.

False positive rate. We calculated false positive rates for GAMM and LMEM models applied to data with time-varying effects across the 48 different combinations obtained by factorially combining the three effect types (within, between, interaction) with the two presentation orders (random and blocked) and the eight autocorrelation scenarios. The p values were computed based on Wald z statistics. It is well-known that this method tends to be slightly *anti-conservative*—that is, yielding false positive rates above the α level (Baayen et al., 2008; Luke, 2017). A key question of the current investigation was whether the naïve application of LMEMs to data with known autocorrelation would yield additional anti-conservativity beyond this baseline. As Figure 6 illustrates, with some exceptions, most false positive rates were acceptably close to the nominal α level.

We compared the false positive rates to the 99% Agresti-Coull confidence interval for LMEM models applied to baseline data with temporally unstructured variation (Gaussian white

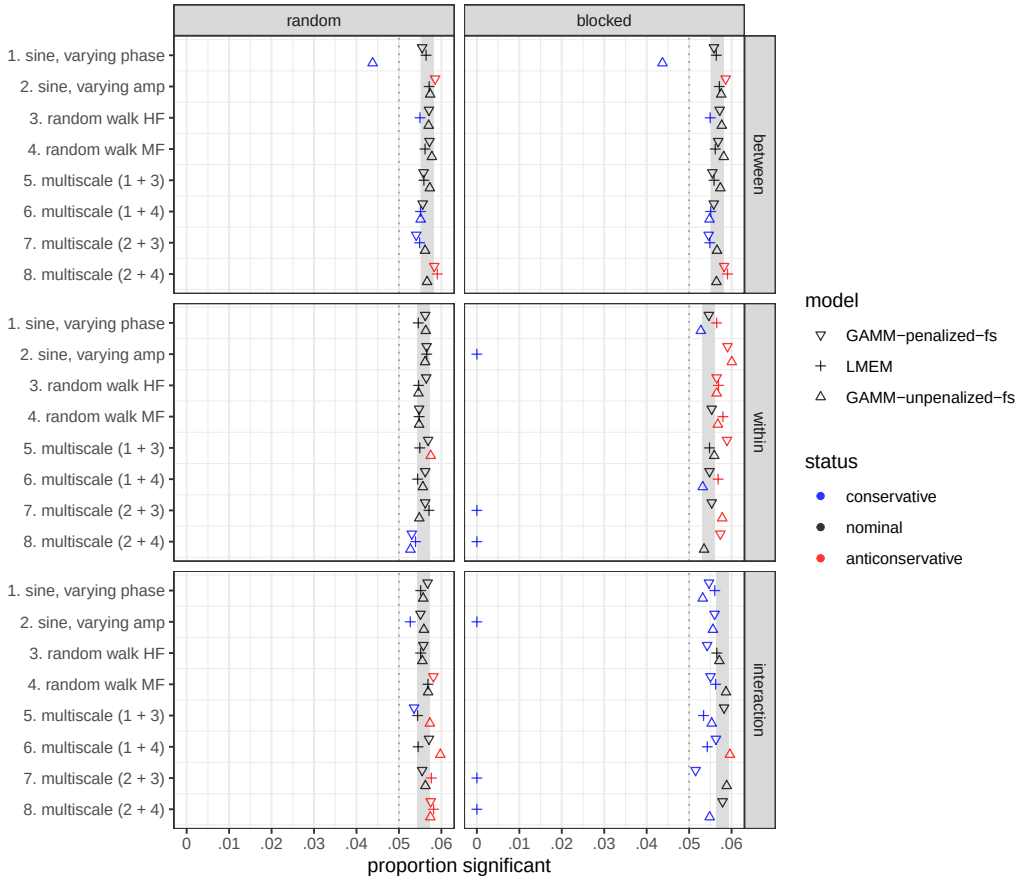

Figure 6. False positive rates. The shaded region represents the 99% Agresti-Coull confidence interval of the false positive rate obtained for LMEMs applied to data without temporal autocorrelation.

noise). Overall, false positive rates for eight of the 48 LMEM cases exceeded the confidence interval, with a maximum false positive rate of .059. This was no worse than for GAMMs, where the false positive rate exceeded the upper bound in nine cases for the unpenalized version, with a maximum false positive rate of .060. For the penalized version, the false positive rate exceeded the confidence interval in ten cases, with a maximum false positive rate of .059 In sum, for multi-level data from designs with appropriate counterbalancing of presentation order, there is no reason to think that application of LMEMs to data with trial-by-trial fluctuations increases the rate of false positives.

For designs with blocked presentation, there were three scenarios where LMEMs were extremely conservative; specifically, in Scenarios 2, 7, and 8, no false positive ever occurred for the test of the within-subject effect. GAMMs with unpenalized factor smooths exhibited moderate conservativity for tests of the between-subject factor in Scenario 1, with a false positive rate of about .044 across both presentation orders. In short, applying LMEMs to data with trial-by-trial fluctuations may lead to extreme conservativity in blocked designs, but general concerns about increased false positive rates for LMEMs find no support in

our data, despite the fact that these models had residuals that were not independent with respect to time, and violated normality assumptions in all scenarios except 3 and 4.

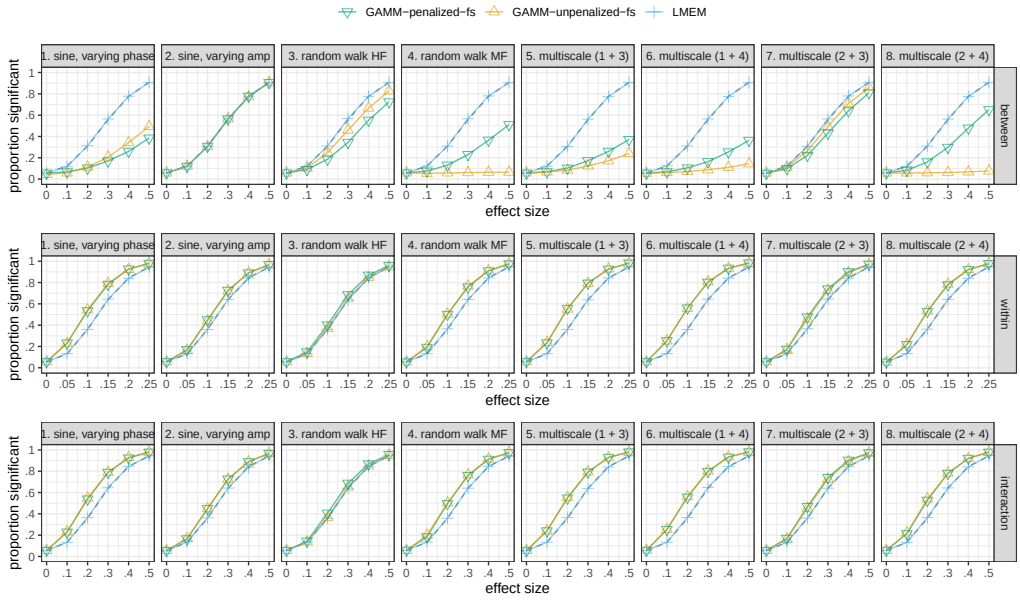

Figure 7. Power curves for the between effect (top row), within effect (middle row) and interaction (bottom row) in the randomized design. Note that the range of effect sizes considered for the within-subject effect is half that used for the between-subject and interaction effects. The dashed line in the background reflects average performance for LMEMs on data with a comparable level of Gaussian white noise.

Power for randomized designs. The power functions for GAMMs and LMEMs depend in a complex way on the type of effect (between or within), whether presentation order was random or blocked, and scenario. Figure 7 shows results for the randomized design. For comparison, in each plot the curves for the GAMM and LMEM approaches are plotted against the averaged performance of GAMMs and LMEMs in the baseline scenario where the errors contained a corresponding amount of white noise. The curves for the interaction effect patterned nearly identically to the corresponding curves for the within-subject effect, albeit with the former exhibiting exactly half of the power of the latter (note the difference in range of the x-axis scale).

Across all effects and scenarios within the random presentation order, power for LMEMs that ignored time-varying effects was identical to power for LMEMs on datasets with uncorrelated Gaussian noise. The power curves for the two sets are indistinguishable. Viewed in one way, this result is not surprising: as we noted in the Introduction, the temporal ordering of each subject's residuals is unimportant in a model where time plays no role in the model. Still, it is somewhat surprising that LMEMs performed so well even in cases where underlying sinusoidal patterns introduced non-normality into the residuals. This confirms our contention at the outset: that from the point of view of a garden variety mixed-effects model, when each participant receives a different random presentation order,

any temporal structure in the residuals is essentially irrelevant; each subjects' residuals are fully exchangeable over time.

However, when there is temporal ordering in the residuals, this can be exploited to improve power over the baseline for within-subject effects. In all scenarios with random presentation, GAMMs yielded modest improvements in power over LMEMs for within-subject factors as well as for any interactions involving these factors. We calculated the average percentage gain for GAMMs over LMEMs in each of the 8 scenarios, combining data from the within-subject and interaction effects given their near identical patterning. For GAMMs with unpenalized factor smooths, power gains as compared to LMEMs ranged from 2% (scenario 3) to 37% (scenario 6) with a median of 27%. For GAMMs with penalized factor smooths, power gains as compared to LMEMs ranged from 7% (scenario 3) to 37% (scenario 6) with a median of 25%.

However, the power gains with GAMMs for within-subject effects in randomized designs typically came at the price of impaired power for between-subject effects. Power for GAMMs on between-subject effects never exceeded power for LMEMs, and was very poor in five of the eight scenarios (1, 4, 5, 6, and 8). At best, GAMMs were equivalent to LMEMs (scenario 2) or only slightly worse (scenarios 3 and 7, where average power for GAMMs with unpenalized factor smooths was 85% and 90% of LMEM power, respectively; for GAMMs with penalized factors smooths, it was 69% and 80% of LMEM power, respectively). At worst, across the range of effect sizes examined, power for GAMMs with unpenalized factor smooths remained stuck at the α-level (.05) while LMEMs approached 100% power (scenarios 4 and 8). In these scenarios, the variance associated with the between-subject manipulation was fully absorbed by the factor smooths, rendering even extremely large effects completely undetectable.

When comparing penalized and unpenalized GAMMs for randomized designs, both versions yield identical performance for within-subject effects across scenarios, but show varying patterns for between-subject effects. Apart from Scenario 2, where between-subject power was equivalent, the unpenalized version outperformed the penalized version in Scenarios 1, 3, and 7, with gains of about 23%, 19%, and 8%, respectively. The penalized version outperformed the penalized version in Scenarios 4, 5, 6, and 8, with gains of about 335%, 43%, 102%, and 419%, respectively.

Power for blocked designs. Turning now to power for designs with blocked presentation order (Figure 8), where variation of the within-subject effect is confounded with time-varying effects, GAMMs show drastically improved detection of within-subject effects relative to LMEMs for all but the pure random walk cases (Scenarios 3 and 4). In Scenarios 2, 7, and 8, which included a common underlying sinusoidal pattern, power for detecting the within-subject effect or interaction effect with LMEMs remained at floor over the full range, while both GAMM versions performed vastly better, matching (or nearly matching) baseline power in Scenarios 1, 2, 5, and 6. In Scenarios 7 and 8, power for GAMMs was impaired relative to baseline, but still vastly superior to LMEM power. Unexpectedly, the relationship was reversed in random walk Scenarios 3 and 4, where LMEMs outperformed both GAMM versions. Relative to unpenalized GAMMs, LMEMs improved power by 33% and 43% respectively; relative to penalized GAMMs, LMEMs improved power by 85% and 59%. In sum, power for within-subject effects in designs with blocked presentation tended to be far

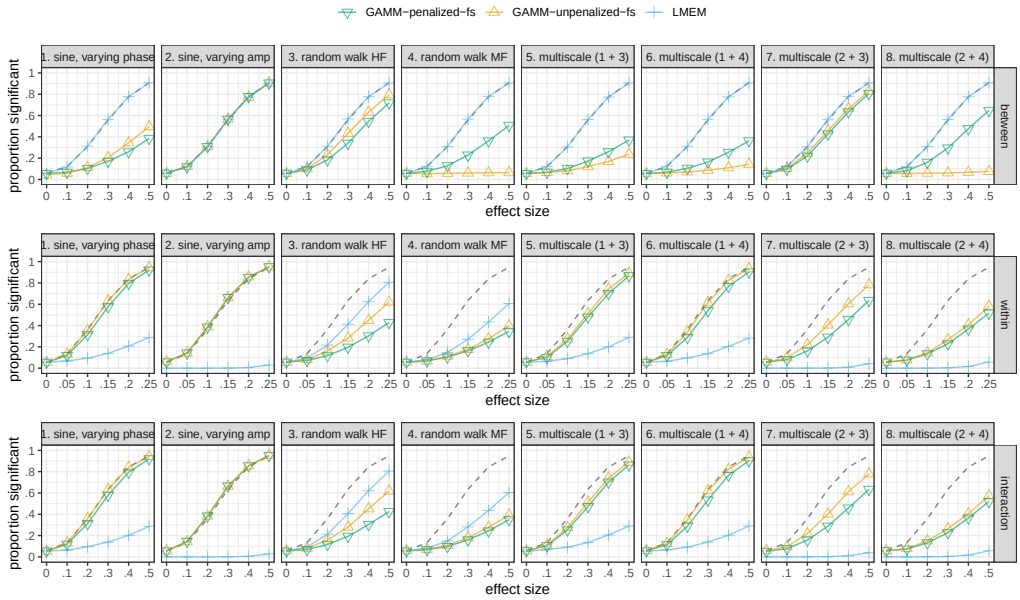

Figure 8. Power curves for the between effect (top row), within effect (middle row) and interaction (bottom row) in the blocked design. Note that the range of effect sizes considered for the within-subject effect is half that used for the between-subject and interaction effects. The dashed line in the background reflects average performance for LMEMs on data with a comparable level of Gaussian white noise.

superior for GAMMs except in purely random walk scenarios, where LMEMs were superior.

For the between-subject factor under the blocked presentation, LMEMs matched baseline performance, while GAMMs showed an unacceptable degree of conservatism. Indeed, the curves were equivalent to those from the random presentation condition.

Comparing the performance of penalized and unpenalized GAMMs on power for the within-subject effect in blocked designs showed variation across scenarios. Apart from Scenario 2, where performance was equivalent, unpenalized GAMMs always outperformed the penalized versions. Average gains in power ranged from 5% (Scenario 5) to 43% (Scenario 3).

The relative performance of penalized and unpenalized GAMMs on power for the between-subject effect showed exactly the same patterns as for the randomized designs.

Bias and precision. To assess the bias and precision of the parameter estimates, we calculated the difference between each parameter estimate and the true population value, and then formed distributions (Figure 9). All of the distributions are centered at zero, which indicates no bias across any effects under either presentation order and across all scenarios. The precision data closely mirrors the main findings from the power curves: (1) estimates for between-subject effects were imprecise under GAMMs relative to LMEMs, especially for the unpenalized versions; (2) GAMMs improved precision for within-subject effects and interactions under the random presentation order; and (3) under the blocked presentation order, LMEM estimates for within-subject effects and interactions tended to be imprecise

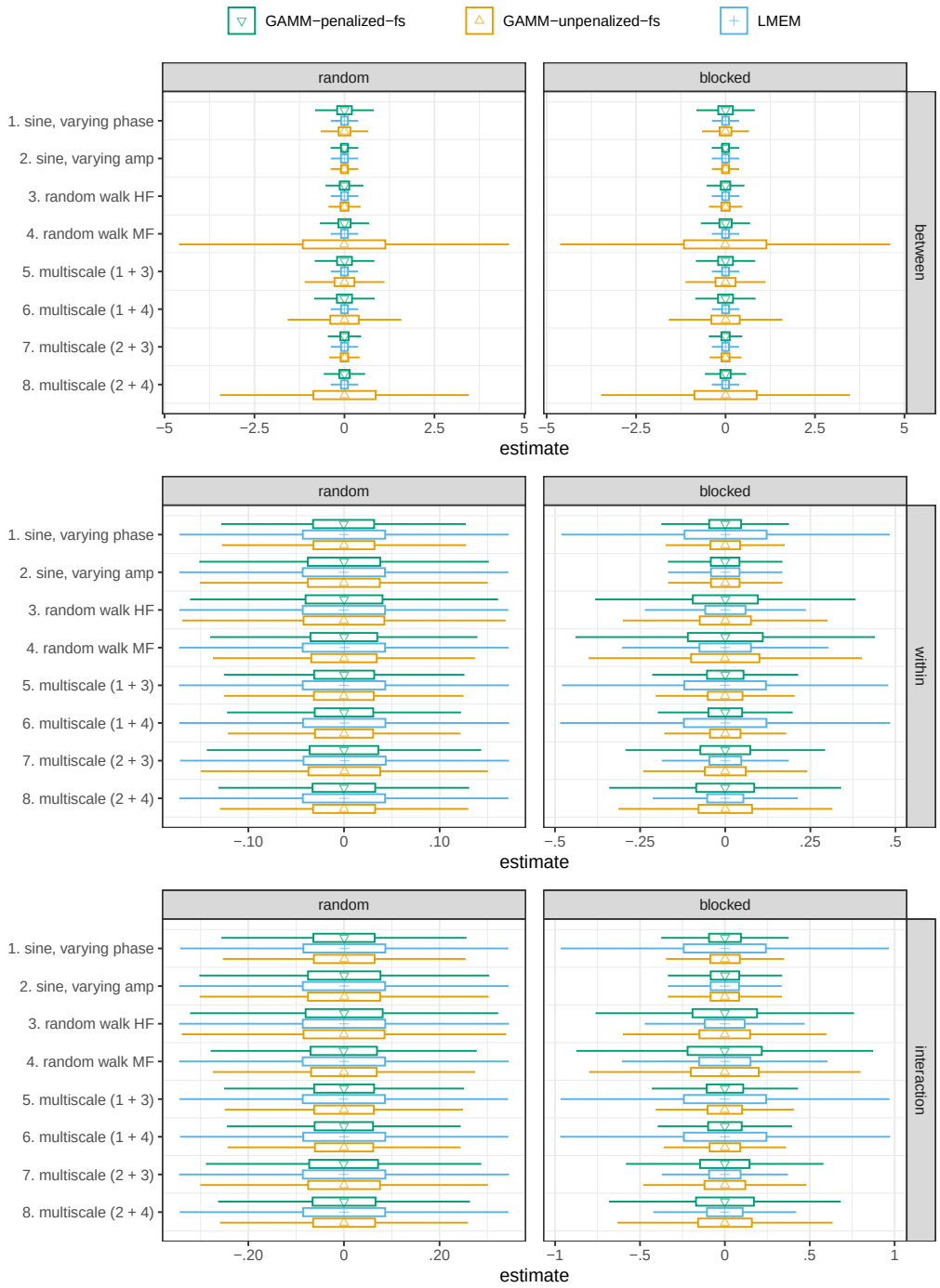

Figure 9. Bias and precision of parameter estimates.

compared to the two GAMM versions, except in the two random walk scenarios (Scenarios 3 and 4).

Simulation Set 2: Many Labs 3 Stroop Dataset

The previous simulations illuminated properties of LMEMs and GAMMs by challenging them with a variety of artificial time-varying patterns in the residuals. The patterns were based on theoretical criteria and were not intended to form a representative sample of real-world patterns. The findings lend overwhelming support to our claim that ignoring trial-by-trial fluctuations in studies with randomized presentation order does *not* increase false positive rates. They also suggest scenarios where the use of GAMMs will enhance power for within-subject effects or impair power for between-subject effects. To show these findings have external validity, it would be advantageous to reproduce them with real rather than artificial data.

For this second set of simulations, residuals were drawn directly from a real dataset containing practice effects. The American Psychological Association's online dictionary defines practice effects as "any change or improvement that results from practice or repetition of task items or activities" (https://dictionary.apa.org/practice-effect). In the case of repeated measures designs, participants are likely to respond more quickly and accurately on later trials than on earlier trials as they master task demands and become familiar with stimuli. While we know of no overview study estimating their prevalence, the sheer volume of literature on this topic suggests they are a common feature in datasets with repeated measurements, appearing across a variety of measurement types and time scales (e.g., Keuleers, Diependaele, & Brysbaert, 2010).

As source data, we used the Stroop Task dataset from the Many Labs 3 mega-study (Ebersole et al., 2016). This is a very large dataset containing response latencies from 3,337 distinct participants performing 63 trials of a version of the Stroop task (Stroop, 1935), for a total of 210,231 observations. In the basic Stroop task, participants must identify the font color of a word. Although the actual identity of a word is irrelevant for reporting its color, the basic finding is that people are slower and less accurate in reporting the color of words whose semantics are incongruent with the color (responding "green" to the word RED presented in a green font) relative to when the semantics of the word are congruent (responding "green" to the word GREEN presented in a green font).

In the Many Labs 3 version of the task, participants saw the color words RED, GREEN, and BLUE presented in red, green, or blue font, with each color word appearing 21 times, seven times in each font color. They identified the font color by pressing one of three assigned response keys, and the response and its latency (in milliseconds) were recorded. Details about randomization are not provided, but to all appearances, each participant received the stimuli in a different random ordering. Further details about the procedure are available in the Many Labs 3 repository at https://osf.io/csj7r/.

We downloaded the data from the Many Labs 3 repository at https://osf.io/n8xa7/, preprocessed it and bundled it as part of the accompanying `{autocorr}` R package, available as the object `stroop_ML3`. Consistent with the Many Labs 3 procedures, we removed any

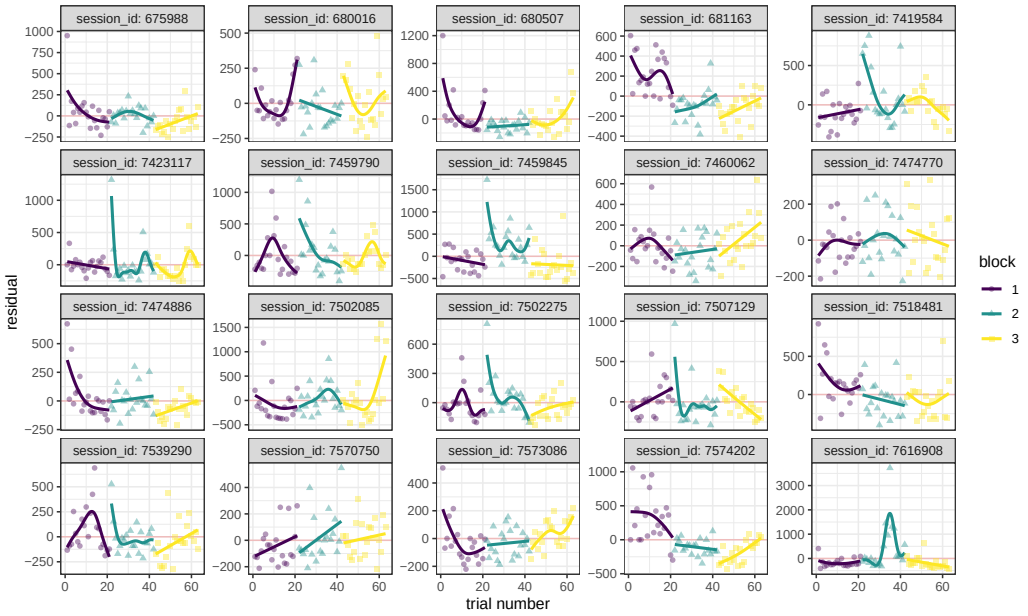

Figure 10. Trial-by-trial patterns in residuals for 20 participants in the Many Labs 3 Stroop dataset, chosen to illustrate the variety of patterns.

incorrect responses as well as response latencies greater than 10 seconds (replacing them with `NA` values). To estimate parameters for data generation and extract residuals, we fit a linear mixed effects model containing a single fixed predictor for *congruency* (the congruency of the color word with the font color) and by-subject random intercepts and a random slope for congruency. The parameter estimates and residuals are stored in the object `stroop_mod` in the `{autocorr}` package.

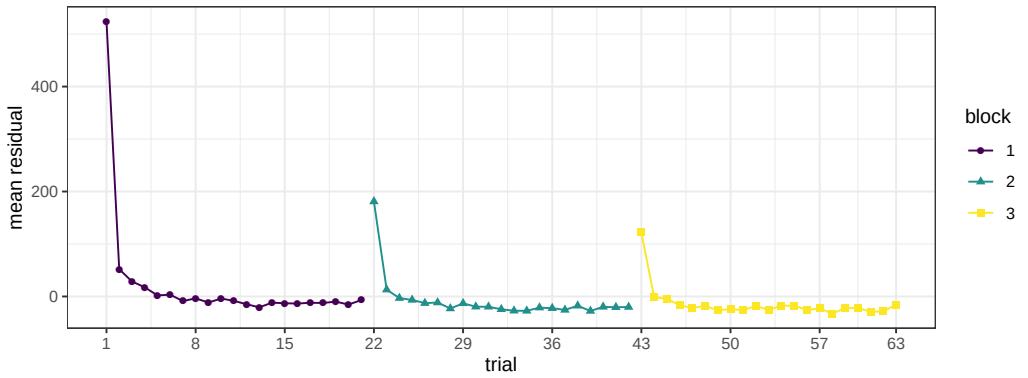

Figure 11. Mean patterns in the residuals of the Many Labs 3 Stroop dataset.

Figure 10 presents a sample of residuals from 20 participants, chosen to illustrate the variety of patterns observed in the data, while Figure 11 shows the overall pattern averaged across all 3,337 participants. Although this was not reported in the Many Labs 3 documention, discontinuities in the residuals suggest that the 63 trials were divided into three blocks of 21 trials, with a practice effect at the start of each. Inclusion of data with such discontinuities

could cause problems for GAMMs. To keep our models simple, we opted to use only those residuals from the first block.

Method

The data generating process and the parameter value distributions were identical to the first set of simulations except for the following differences. Given the identical patterning of the interaction and within-subject effects in the previous simulation, we set the interaction effect to zero and excluded it from our models. All simulated datasets had 48 participants and 20 trials, 10 of which were in the congruent condition, and 10 of which were in the incongruent condition.

The errors for each participant in each simulated dataset were created by sampling from the set of residuals from a linear mixed-effects model fit to the Stroop dataset. The model syntax was

```
lmer(latency ~ cong + (cong || session_id), stroop_ML3)
```

where `latency` is the response latency and `cong` is a deviation-coded predictor for congruency. We used the estimates of variance components from the model (by-subject random intercept and random slope for `cong`) as generative parameters. The random intercept standard deviation, τ_0, was estimated as 165 and the random slope standard deviation, τ_1, as 18. The random correlation was fixed to zero.

After randomly generating the individual subject random effects for each dataset, these were combined with the fixed effects to calculate the fitted values, and we then 'grafted' a randomly sampled set of 48 real residual vectors onto these fitted values to calculate response latencies. Each set of 48 vectors were sampled from the 3,337 vectors of residuals without replacement, so that the same residual vectors would not appear twice in the same dataset. The logic is implemented by `simulate_stroop()` in `{autocorr}`. We only used the first 20 residual values from each vector to avoid including the discontinuities noted above, and because the design required an even number of trials.

After generating each dataset, we fit the same three models to the data as in the first set of simulations: two GAMM models, one with penalized and one with unpenalized factor smooths, and one linear mixed effects model. Both GAMMs included a 'common smooth' to account for shared variance in practice effects across participants. The code is available in `fit_stroop()` in `{autocorr}`.

To estimate power and Type I error, we ran 10,000 Monte Carlo simulations at each of six parameter settings for the fixed effects: at $\beta_1 \in \{0, 14, 28, 42, 56, 70\}$ and $\beta_2 \in \{0, 44, 88, 132, 176, 220\}$ where β_1 and β_2 are the within-subject and between-subject fixed effects, respectively.

	model	within	between
1	GAMM-penalized-fs	0.038	0.055
2	GAMM-unpenalized-fs	0.038	0.055
3	LMEM	0.040	0.053

Table 3

Type I error rates for within-subject and between-subject effects by model, Stroop simulations.

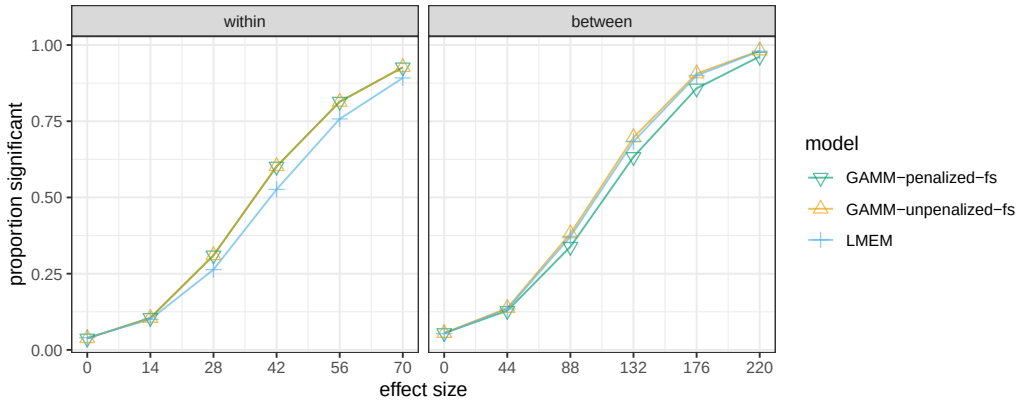

Figure 12. Power curves for LMEMs and GAMMs with and without penalized factor smooths, for within-subject and between-subject effects.

Results and Discussion

As with the previous simulations, there was no evidence that LMEMs applied to data with trial-by-trial fluctuations inflated false positive rates beyond the specified α level (Table 3). All error rates were close to α. Also consistent with the previous simulations, GAMMs boosted within-subject power while impairing between-subject power (Figure 12). The enhancement to within-subject power did not depend upon whether factor smooths were penalized or unpenalized; on average, GAMMs boosted power by 10%, with a maximum of 18%. LMEMs showed superior power for between-subject effects relative to GAMMs with

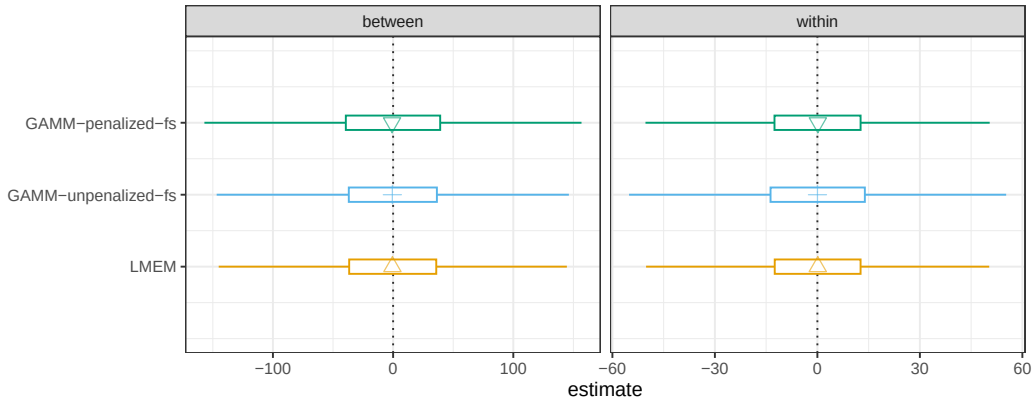

Figure 13. Bias and precision for within-subject and between-subject effects by model.

the recommended penalized factor smooths, with an average gain of 6% and a maximum of 10%. Between-subject power for GAMMs with unpenalized factor smooths showed no such impairment. These main patterns are echoed in the plot of bias and precision of parameter estimates (Figure 13).

General Discussion

Experts have suggested that fitting standard linear mixed-effects models to data with residual autocorrelation inflates false positive rates. We argued that this would not be the case for experiments with randomized presentation order, and our simulations uniquivocally confirm this. For randomized experiments, it is perfectly safe to ignore residual autocorrelation. Finding structure when plotting residuals against the irrelevant variable of time merely implies the possibility of increasing power and precision; it does not imply that a model that ignores time is inadequate, nor that it violates statistical assumptions. That power could be improved by using GAMMs does not entail that one *should* use them, since using complex, advanced techniques that are not fully understood itself has costs: costs of acquiring the relevant technical skills; costs of having to estimate a more complex model and the increased computing time and potential convergence issues this entails; costs to reproducibility due to greater analytic flexibility; and the potential hidden costs of using highly complex techniques whose proper use and potential side-effects are not yet fully understood.

We investigated the costs and benefits of modeling by-trial fluctuations using Monte Carlo simulations, comparing the performance of GAMMs versus conventional LMEMs in the analysis of multi-level data with a variety of types of by-trial fluctuations. According to our findings, using GAMMs to remove time-varying effects in residuals is unnecessary and sometimes even counterproductive. Although GAMMs with factor smooths offer minor improvements to power for within-subject effects, this benefit was often accompanied by unacceptably low power for between-subject effects.

For experiments with random presentation where time is not a variable of interest, it is unnecessary and potentially misleading for researchers to look for temporal structure in the residuals during the model checking stage. By-subject autocorrelelograms may be especially misleading, because they assume that the underlying process is stationary over time; in other words, for any given lag k, they assume the same correlation between residuals for trials i and $i + k$ as between those for trials m and $m + k$. For instance, in an experiment with n trials this assumes that the correlation of residuals for trials 1 and 2 be the same as for any other two trials separated by a lag of 1. Stationarity seems extremely unlikely in human data due to learning, fatigue, and mind-wandering effects. Autocorrelation plots are therefore likely to be deceptive. If one is looking for temporal structure, the best place to look is in a simple plot of the residuals by trial (or time).

When deciding whether or not to account for by-trial fluctuations in data with GAMMs, considerations of experimental design are crucial. Our analysis suggests that designs where all factors are administered within sampling units (typically, subjects or stimuli), GAMMs could moderately increase power. However, if any factors have been administered between

sampling units, results from GAMMs for these factors may be misleading, due to the potential for increased false negatives. Applying GAMMs to a 2x2 experiment where both factors are between-subjects could be potentially catastrophic for power. For designs with mixed between and within factors, it would seem prudent to analyze between subjects factors and within subject factors in separate models, one using factor smooths to improve power for within-subject effects, and one without them to test between-subject effects without compromising to power.

With blocked presentation order, the nuisance perspective on by-trial fluctuation is forced upon us, because such designs naturally confound treatment variation with by-trial variation. Although LMEM performance was extremely poor for within-subject effects with sinusoidal patterns, we must note that the simulations presented a worst-case situation wherein the frequency of the sine wave was identical to the frequency of the condition indicator variable (considered as a square wave). For the fixed phase/varying amplitude scenarios, by-subject random slope variation becomes nearly completely confounded with the time-varying effects. In real situations, such near-perfect confounding is extremely unlikely, and so our results dramatically overestimate the impairment to within-subject power. A further observation that is of interest is that GAMMs actually performed worse than LMEMs on blocked designs with the random walk patterns even for within-subject effects. Thus, GAMMs are not guaranteed to do better than LMEMs on data from experiments with blocked presentation.

All our simulated experiments conformed to an ideal experimental design, with full counterbalancing, a presentation order that was randomized independently for each subject (or blocked but perfectly counterbalanced), and no missing data. Real datasets often—perhaps usually—fall short of this ideal. Researchers sometimes do not randomize independently for each subject, but instead re-use a small number of random orders, or randomize only at the level of the presentation list. Missing data—a factor less under the researcher's control—may also give rise to imbalances in the design. Our investigation suggests that if trial-by-trial fluctuations in performance are present, partial randomization and unbalanced data may be problematic, at least to the extent that they confound trial-by-trial fluctuations with variability from independent variables. However, from the current results it is not entirely clear how much this should be cause for concern, nor how well using GAMMs would ameliorate these problems. To gain better insight, more simulations are needed that vary a greater range of factors. For now, we recommend that researchers randomize trial order independently at the subject level to minimize the potential impact of by-trial fluctuations.

We did not intend for the autocorrelation scenarios we explored to comprise a representative sample of the set of time-varying nuisance patterns found in real datasets. Although we sought to make the ratio of residual variance to random effect variance realistic, our choice of autocorrelation patterns in the first set of simulations was intended to map out the space of possibilities and highlight properties of GAMMs. We confirmed the general pattern with real data in the second set of simulations, but it would be worthwhile examining additional natural datasets. Also, we have not considered the effects of subject sample size nor number of trials. We would expect the problems we uncovered to be exacerbated with smaller samples, where estimation is more difficult.

To simplify our investigation, our simulated data and the models we fit only assumed that fluctuations affected the time-varying intercept, but was unrelated to the expression of effects of the independent variables. In other words, we assumed time-varying random intercepts but time-independent random slopes. This is similar to the models that Baayen et al. (2017) fit to real datasets, which also assumed static random slopes. However, any manipulation that depends on subjects' attention to some distinction may show an effect size that fluctuates along with fatigue. The logic of the analysis we presented in the introduction suggests this is still not a problem—so long as presentation order is random, it is entirely valid to estimate the *mean* effect and ignore fluctuations in its manifestation over time. Although precision may be improved by estimating such effects, we see no basis for assuming that assumptions would be violated by not doing so.

We have also only considered univariate data where there is a single observation per trial. This excludes from consideration many types of research such as visual world eyetracking, EEG/MEG, pupillometry, and longitudinal studies. Our intention is not to dampen enthusiasm for the use of GAMMs in analyses where time is a critical variable. Indeed, GAMMs strike us as a promising approach in these contexts (van Rij et al., 2019).

GAMM experts may be able to improve model performance for some of the cases we examined, or perhaps even eliminate altogether some of the problems we have diagnosed by using values other than the `{mgcv}` defaults. However, this would not undermine our main contention that trial-by-trial fluctuations can usually be safely ignored. Moreover, this would reinforce our point that GAMMs are probably too risky for the typical user. A far simpler and more broadly accessible defense against statistical artifacts from trial-by-trial fluctuations is to exert care over randomization and counterbalancing when conducting an experiment.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*, 163–181.

Amaro, E., & Barker, G. J. (2006). Study design in fMRI: basic principles. *Brain and Cognition*, *60*, 220–232.

Baayen, H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baayen, H., van Rij, J., de Cat, C., & Wood, S. (2018). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In *Mixed-Effects Regression Models in Linguistics* (pp. 49–69). Springer.

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bence, J. R. (1995). Analysis of short time series: correcting for autocorrelation. *Ecology*, *76*, 628–639.

Blake, M. J. F. (1967). Time of day effects on performance in a range of tasks. *Psychonomic Society*, 349–350.

Briggs, W. L., & Henson, V. E. (1995). *The DFT: An Owner's Manual for the Disrete Fourier Transform*. SIAM.

Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., ... Sham, P. (1996). Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, *35*, 261–277.

Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, *44*, 32–61.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68-82.

Forbach, G. B., Stanners, R. F., & Hochhaus, L. (1974). Repetition and practice effects in a lexical decision task. *Memory & Cognition*, *2*, 337–339.

Griffiths, W., & Beesley, P. (1984). The small-sample properties of some preliminary test estimators in a linear model with autocorrelated errors. *Journal of Econometrics*, *25*, 49–61.

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods*, *3*(1), 104.

Jones, M., Curran, T., Mozer, M. C., & Wilder, M. H. (2013). Sequential effects in response time reveal learning mechanisms and event representations. *Psychological Review*, *120*(3), 628.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice Effects in Large-Scale Visual Word Recognition Studies: A Lexical Decision Study on 14,000 Dutch Mono- and Disyllabic Words and Nonwords. *Frontiers in Psychology*, *1*, 174.

Kleitman, N. (1963). Sleep and wakefulness.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*, 1494–1502.

McVay, J. C., & Kane, M. J. (2009). Conducting the train of thought: working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 196.

Mirman, D. (2016). *Growth curve analysis and visualization using R*. CRC press.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*, 134–140.

Olszowy, W., Aston, J., Rua, C., & Williams, G. B. (2019). Accurate autocorrelation modeling substantially improves fMRI reliability. *Nature Communications*, *10*, 1–11.

Papoulis, A., & Pillai, S. U. (2002). *Probability, Random Variables and Stochastic Processses* (Fourth ed.). Boston: McGraw Hill.

Purdon, P. L., & Weisskoff, R. M. (1998). Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Human Brain Mapping*, *6*(4), 239–249.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Shinozuka, M., & Deodatis, G. (1991). Simulation of stochastic processes by spectral representation. *Applied Mechanics Reviews*, *44*(4), 191–204.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. *arXiv preprint arXiv:1703.05339*.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. (2019). Analyzing the time course of pupillometric data. *Trends in Hearing*, *23*, 1–22.

Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution*, *1*, 7-18.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*,

3-36.
Wood, S. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC.
Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage*, *14*, 1370–1386.
Worsley, K. J., & Friston, K. J. (1995). Analysis of fMRI time-series revisited—again. *Neuroimage*, *2*, 173–181.

Appendix

Generating correlated residuals

We use the numerical method developed in Shinozuka and Deodatis (1991) to generate correlated residuals. Residuals are modeled as a zero-mean stationary stochastic process $X(t)$, which we express in its temporally discretized spectral form as (Papoulis & Pillai, 2002)

$$X\left(p\Delta t\right) = \mathrm{Re}\left\{\sum_{n=0}^{M-1} B_n \exp\left[i(n\Delta\omega)(p\Delta t)\right]\right\}, \tag{1}$$

for $p = 0, \ldots, M-1$ and $B_n = \sqrt{2}A_n \exp(i\alpha_n)$ for $n = 0, \ldots, M-1$. The coefficients A_n are given by $A_n = \sqrt{2S(n\Delta\omega)\Delta\omega}$ with α_n uniformly distributed on $[0, 2\pi]$ and $A(0) = 0$ for $S(0) \neq 0$. $S(\omega)$ denotes the spectrum (see below). We discretize time and frequency in steps of Δt and $\Delta\omega$, respectively.

The spectrum $S(\omega)$ is the Fourier transform of the autocorrelation function (ACF) $R(t)$ and is given by

$$S(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} R(t)\mathrm{e}^{-i\omega t}\mathrm{d}t\,. \tag{2}$$

In the present study, we empployed the squared-exponential function

$$R(t) = \sigma^2 \exp\left(-\frac{t^2}{2\gamma^2}\right). \tag{3}$$

For practical reasons, we cannot evaluate the integral in (2) for limits at infinity, but require some finite limit $T_{\max}$. Since realistic ACFs decay towards zero for large times, we choose $T_{\max}$ such that $|\gamma(t)| < \delta$ for $t > T_{\max}$ and some δ with $0 \leq \delta \ll 1$. In other words, the ACF is very small for times larger than $T_{\max}$, so that we make a vanishingly small error by excluding these values in the evaluation of the integral. ACFs for stationary stochastic processes are symmetric around $t = 0$, so that the entire time domain for the integral in (2) is $[-T_{\max}, T_{\max}]$.

As Eq. (1) shows, there is a maximal frequency $(M-1)\Delta\omega$ — which results from setting $n = M-1$ in the sum — for which we need to compute the spectrum. In other words, there is a critical cut-off frequency ω_c for the spectrum, which we determine by demanding that

$$\int_0^{\omega_c} S(\omega)\mathrm{d}\omega = (1-\epsilon)\int_0^{\infty} S(\omega)\mathrm{d}\omega\,, \tag{4}$$

for $0 < \epsilon \ll 1$. Equation (4) states that we choose ω_c in such a way that we only lose a small fraction ϵ of the total power, i.e. the integral over the spectrum. The practical steps for determining ω_c are as follows. We first divide the time domain $[T_{\max}, T_{\max}]$ into $2N_t$ grid points and then use the fast Fourier transform (FFT) to evaluate the integral in Eq. (2). The FFT yields $2N_t$ values of the spectrum evaluated at $[-N_t\Delta\omega_{\mathrm{S}}, (N_t-1)\Delta\omega_{\mathrm{S}}]$, where $\Delta\omega_{\mathrm{S}} = (2\pi)/(2T_{\max})$ denotes the fundamental frequency (Briggs & Henson, 1995). Note that we work with the angular frequency ω, and not the frequency f, hence a factor of 2π. The cut-off frequency ω_c follows by numerically evaluating the integrals in Eq. (4) based on the computed spectrum on the frequency domain $[-N_t\Delta\omega_{\mathrm{S}}, (N_t-1)\Delta\omega_{\mathrm{S}}]$.

With the critical frequency ω_c determined, we introduce the spectral discretisation $\Delta\omega = \omega_c/N$ and the corresponding spectral grid $[-N\Delta w, (N-1)\Delta w]$ for the the stochastic process in Eq. (1). We can choose N to be different from N_t and hence Δw may differ from $\Delta\omega_{\mathrm{S}}$. Because $X(t)$ in Eq. (1) has a period of T_0, i.e. $X(t+T_0) = X(t)$, we arrive at the constraint

$$M\Delta t = T_0 = \frac{2\pi}{\Delta\omega}\,. \tag{5}$$

In other words, we can either fix M, and obtain Δt, or we fix Δt, and hence obtain M. Because we can choose Δt, the time discretisation of $X(t)$ can be different from the time discretisation that we used for determining the cut-off frequency for the spectrum, $\Delta t_{\mathrm{S}} = T_{\max}/N_t$.

A key issue in spectral reconstruction is known as aliasing. Essentially, components with certain frequencies, i.e. certain values of n in Eq. (1), cannot be distinguished (hence the name alias). To avoid this, the critical frequency in the spectral decomposition (f_c) needs to be smaller than the so-called Nyquist frequency, which is half the sampling frequency f_s, i.e.

$$f_c < \frac{f_s}{2} \quad \Longleftrightarrow \quad \omega_c < \frac{2\pi f_s}{2} = \frac{2\pi}{2\Delta t}\,. \tag{6}$$

Here, we used the fact that the sampling frequency is the inverse of the time discretization Δt of $X(t)$: $f_s = 1/\Delta t$. Equation (6) entails that given a cut-off frequency ω_c, the time discretization of $X(t)$ needs to satisfy

$$\Delta t < \frac{2\pi}{2\omega_c}\,. \tag{7}$$

In other words, there is a maximal time step for the numerical construction of $X(t)$. Using Eq. (5) and the definition of $\Delta\omega$ from above, this relates M and N via

$$\Delta t = \frac{2\pi}{M\Delta\omega} = \frac{2\pi N}{M\omega_c} < \frac{2\pi}{2\omega_c} \quad \Longleftrightarrow \quad 2N < M\,. \tag{8}$$

This constraint needs to be observed when implementing Eq. (1). A final consideration is worth noting. As mentioned above, the spectral discretization for computing the cut-off frequency, $\Delta\omega_S$, may differ from the one for constructing $X(t)$, $\Delta\omega$. If this is the case, we need to recompute the spectrum since the computation of $X(t)$ in Eq. (1) relies on $S(n\Delta\omega)$ via the coefficients A_n, while the integration in Eq. (2) is based on $S(n\Delta\omega_{\mathrm{S}})$. In order to recompute the spectrum for the frequency domain $[-N\Delta w, (N-1)\Delta w]$ using FFT and

Eq. (2), we require a new time discretization. Neither Δt, which enters the computation of $X(t)$, nor Δt_{S} from above are appropriate. The new time discretisation follows from the frequency discretization $\Delta\omega$ and the fact that there are $2N$ grid points as

$$\Delta t \Delta f = \frac{1}{2N} \quad \Longleftrightarrow \quad \Delta t = \frac{2\pi}{2N\Delta\omega} = \frac{2\pi}{2\omega_c}\,. \tag{9}$$

As a consequence, the length of the time domain for the integral in Eq. (2) is

$$2N\Delta t = \frac{2\pi N}{\omega_c} = \frac{2\pi}{\Delta\omega} = T_0\,, \tag{10}$$

where we used Eq. (5). Hence, the total time over which we integrate the ACF to obtain the spectrum and the total time of $X(t)$ are identical. What differs is the time discretization.

Probability densities for sine waves

We here derive the probability density for the residuals when they are described by a sum of Gaussian white noise with either a sine function with fixed amplitude/random phase or a sine function with fixed phase/random amplitude. These two cases are labelled 1 and 2 in Figure 5. Let X denote the random variable involving the sine function and $Y \sim \mathcal{N}(0, \sigma^2)$ the Gaussian white noise. When we set $Z = X + Y$, the probability density for Z is given by (Papoulis & Pillai, 2002)

$$p_Z(z) = \int_\Omega p_X(x) p_Y(z - x)\mathrm{d}x, \tag{11}$$

where the possible values of X are collected in the set Ω. For the first case when we fix the amplitude and randomize the phase, we use $X = \sin(\omega t + \Phi)$ where $\omega = 2\pi/T$ is the frequency of the sine function and $\Phi \sim \mathcal{U}[-\pi, \pi]$ is a uniformly distributed random phase. Note that we use $\Delta = 48$ and $t \in [0, \Delta]$ throughout, see e.g. Figure 1. Because t is a deterministic variable on $[0, \Delta]$, we can also interpret it as a random variable drawn uniformly from the same interval, i.e. $T \sim \mathcal{U}[0, \Delta]$. With that, it is evident that $(2\pi T + \Phi)$ is a uniformly distributed random variable, which we can replace with $\Psi \sim \mathcal{U}[-\pi, \pi]$ because of the periodicity of the sine function. Hence, we require the probability density for $X = \sin\Psi$. Since the range of X is $[-1, 1]$, we can restrict Ψ to the interval $[-\pi/2, \pi/2]$. Due to the conservation of probability, we obtain (Papoulis & Pillai, 2002)

$$p_X(x) = p_\Psi(\psi)\left|\frac{\mathrm{d}\psi}{dx}\right| = \frac{1}{\pi}\frac{1}{\sqrt{1 - x^2}}\,, \tag{12}$$

where we used that $\mathrm{d}x = \cos\psi \mathrm{d}\psi$ and

$$\cos^2(\psi) = \cos^2(\arcsin x) = 1 - x^2\,, \tag{13}$$

since $\cos^2(x) = 1 - \sin^2(x)$. Note that the restriction of Ψ to $[-\pi/2, \pi/2]$ also insures that we can invert the sine function, since it is strictly monotonically increasing in this interval. From Eq. (11), we therefore find that

$$p_Z(z) = \frac{1}{\pi}\int_{-1}^{1} \frac{F(z - x, \sigma)}{\sqrt{1 - x^2}}\mathrm{d}x\,, \tag{14}$$

where

$$F(x,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{15}$$

is the Gaussian probability distribution. In the second case, when the phase is constant, but the amplitude is randomized, we generate the residuals with $X = A\sin(\omega t)$, where $A \sim \mathcal{U}[0,\alpha]$. Based on the considerations above, we can replace X with $X = A\sin(\Psi)$ where $\Psi \sim \mathcal{U}[-\pi/2,\pi/2]$. Hence, X is the product of two random variables A and $B = \sin(\Psi)$ with the probability density (Papoulis & Pillai, 2002)

$$p_X(x) = \int_{-1}^{1} p_B(b) p_A\left(\frac{x}{b}\right) \frac{1}{|b|} \mathrm{d}b\,. \tag{16}$$

Because of the absolute value in Eq. (16), it is instructive to consider the two cases $b \in [-1,0]$ and $b \in [0,1]$ separately. Starting with the latter, we need to evaluate

$$p_X(x) = \int_{0}^{1} p_B(b) p_A\left(\frac{x}{b}\right) \frac{1}{b} \mathrm{d}b\,. \tag{17}$$

Since $b > 0$ and $a > 0$, $x > 0$. Moreover, as $\max A = \alpha$, we require $x/b \leq \alpha$ in Eq. (17), which leads to

$$p_X(x) = \int_{x/\alpha}^{1} p_B(b) p_A\left(\frac{x}{b}\right) \frac{1}{b} \mathrm{d}b\,. \tag{18}$$

When $b \leq 0$, $x \leq 0$. Because $x/b \leq \alpha$ still holds, we now have the restriction, $b \leq x/a$, so that

$$p_X(x) = \int_{-1}^{x/\alpha} p_B(b) p_A\left(\frac{x}{b}\right) \frac{1}{b} \mathrm{d}b\,, \tag{19}$$

which can be rewritten (transforming $b \mapsto -b$) as

$$p_X(x) = \int_{|x|/\alpha}^{1} p_B(-b) p_A\left(-\frac{x}{b}\right) \frac{1}{|b|} \mathrm{d}b\,, \tag{20}$$

As $p_B(b) = p_B(-b)$ (see Eq. (12)) and $p_A = 1/\alpha$, we can combine Eqs. (18) and (20) into

$$\begin{aligned} p_X(x) &= \frac{1}{\pi\alpha} \int_{|x|/\alpha}^{1} \frac{1}{z\sqrt{1-b^2}} \mathrm{d}b\,, \\ &= \frac{1}{\pi\alpha} \left[\ln\left(\sqrt{1 - \frac{x^2}{\alpha^2}} + 1 \right) - \ln\left(\frac{|x|}{\alpha} \right) \right]. \end{aligned} \tag{21}$$

Using Eq. (11), we finally obtain the probability density for the case of a randomized amplitude as

$$p_Z(z) = \int_{-\alpha}^{\alpha} p_X(x) F(z-x) \mathrm{d}y\,. \tag{22}$$