With great power comes greater responsibility:

cheating with bdots

Last compiled: Tuesday 14th February, 2023 at 11:54

Abstract

Something something high density time series collected to estimate group population curves and compare those for temporal differences while controlling FWER. Original `bdots` made strict assumptions which are likely to not hold in general, resulting in truly disastrous type I error rates. My modify the original 2017 algorithm to introduce an alternative bootstrapping scheme, along with a modified permutation test to examine differences between groups. Our results demonstrate comparable control of FWER and power under the original assumptions while also proving far more robust to divergences in both the mean and error structures of the observed data.

1 Introduction

Note: I don't really want this to be all "oh this is bdots and look at what it does and how it's used" because that is all covered in chapter 1. Instead, I am looking for laser sharp focus on the problem we identified and how we seek to resolve it. But since we live in a polite society I will go through the tedious endeavor of creating an introduction.

The original bdots presented a method whereby the difference in time series between two groups could be analyzed using a FWER correction to the alpha based on the autocorrelation of the resulting t-statistics. While we can confirm that the published results do indeed hold for the case they presented, such a situation is likely atypical and not reflective of the typical case involving VWP data. bdots involved testing between two groups, yet assumed that all subjects within a group had identical mean structures – that is, there was no between-subject variability to be accounted for.

More likely it is the case that in the comparison of two groups, each group has a typical distribution of parameters for its subjects. God its annoying a little bit how bad and out of order this is, but free write so

its ok for now. As we show, when the initial (and restrictive) assumptions made in the original bdots doesn't hold, the resulting TIE is beyond what would be acceptable in most cases.

What we present instead is two alternatives, accommodating flexibility in two of the assumptions made in the original bdots. First, we propose a modified bootstrapping procedure that adequately accounts for observed between subject variability while retaining the novel FWER adjustment method presented for autocorrelated errors. In addition to this, we offer a play on a standard permutation test between the groups, borrowing from the insight of the original bdots in that it also captures within-subject variability as demonstrated in the standard errors in the model fits. We begin by describing the two proposed alternatives to the original bdots bootstrap. We then outline the details of the simulations in demonstrating the TIE rate across a number of experimental conditions, along with the results. Finally if there is time we consider a power analysis of the two resulting methods. I guess it's also fine to consider power in the case in which the original bdots does well. Naturally, the power is great, but then so what?

2 Detail on the original

Most generally the original bdots algorithm, which we will call the v1 bootstrap, proposed that we have empirically observe data resulting from some mean structure with an associated error, with

$$y_{it} = f(\theta_{it}) + \epsilon_{it} \tag{1}$$

where

$$\epsilon_{it} = \phi\epsilon_{i,t-1} + w_{it}, \quad w_{it} \sim N(0, \sigma). \tag{2}$$

Under this paradigm, the errors could be iid normal (with $\phi = 0$) or have an AR(1) structure, with $0 < \phi < 1$. It further assumes homogeneity of the mean structure, meaning that two subjects from the same group would have parameters $\theta_{it} = \theta_{jt}$ for all i, j. In other words, it was assumed that there was no variability in the mean structure between subjects in the same group. This is also evidenced in the original bdots algorithm:

1. For each subject, fit the nonlinear function, specifying AR(1) autocorrelation structure for model errors. Assuming large sample normality, the sampling distribution of each estimator can be approximated by a normal distribution with mean corresponding to the point estimate and standard deviation corresponding to the standard error
2. Using the approximate sampling distributions in (1.), randomly draw one bootstrap estimate for each of the model parameters on every subject
3. Once a bootstrap estimate has been collected for each parameter and for every subject, for each parameter, find the mean of the bootstrap estimates across individuals
4. Use the mean estimates to determine the predicted population level curve, which provides the average population response at each time point

.5 Perform steps (2)-(4) B times (default is $B = 1,000$ to obtain estimates of the population curves. Use these to create estimates of the mean response and standard deviation at each of the time points.

This is demonstrated in step (2.), where each subject is included in each iteration of the bootstrap.

3 Proposed Methods

As is more typically the case, and especially in scenarios under which the v1 bootstrap was considered, subjects within a group exhibit draw their individual parameters from a group distribution with a group mean and some measure of variability. In such a case, there is no presumption that $\theta_i = \theta_j$, and accounting for the between-subject variability will be critical for obtaining a reasonable distribution of the population curves.

3.1 Modified Bootstrap

A more likely case involving subjects in the VWP (or subjects within any group exhibiting between and within subject variability) is as such: suppose for example that we are considering a family of four parameter logistic curves, defined

$$f_\theta(t) = \frac{p-b}{1+\exp\left(\frac{4s}{\mathrm{p}-b}(x-t)\right)} + b \tag{3}$$

where $\theta = (p, b, s, x)$, the peak, baseline, slope, and crossover parameters, respectively. The distribution of parameters for subjects within this group may be normally distributed, with any individual subject i's parameters following the distribution

$$\theta_i \sim N(\mu_\theta, V_\theta). \tag{4}$$

In the course of collecting observed data on subject i, we may find that there is a degree of variability in our observations between trials, reflected in the standard errors derived when fitting the observed data to the functional form in Equation 3. This gives us a distribution for the observed parameter,

$$\hat{\theta}_i \sim N(\theta_i, s_i^2). \tag{5}$$

When obtaining reasonable estimates of the population curve, it is necessary to estimate both the observed within-subject variability found in each of the $\{s_i^2\}$ terms, *as well as* the between-subject variability present in V_θ. For example, let θ^*_{ib} represent a bootstrapped sample for subject i, where

$$\theta^*_{ib} \sim N(\hat{\theta}_i, s_i^2), \tag{6}$$

the distribution estimated by `gnls`. If we were to sample *without replacement*, we would obtain a bootstrapped group mean value θ^*_b, where

$$\theta^*_b = \frac{1}{n}\sum \theta^*_{ib}, \quad \theta^*_b \sim N\left(\mu_\theta, \frac{1}{n^2}\sum s_i^2\right). \tag{7}$$

Such an estimate captures the totality of the within-subject variability with each draw but fails the account for the variability in the group overall. Alternatively, we sample the subject *with* replacement, where each θ^*_{ib} follows the distribution in Equation 6, but the bootstrapped group mean now follows

$$\theta^*_b = \frac{1}{n}\sum \theta^*_{ib}, \quad \theta^*_b \sim N\left(\mu_\theta, \frac{1}{n}V_\theta + \frac{1}{n^2}\sum s_i^2\right). \tag{8}$$

The estimated value remains unchanged, but the variability is now fully accounted for. We therefore present a modified version of the bootstrap which we call the v2 bootstrap, making the following changes to the original:

1. In step (1.), the specification of AR(1) structure is *optional* and can be modified with arguments to functions in `bdots`. Our simulations show that while failing to include it slightly inflates the type I error in the v2 bootstrap when the data truly is autocorrelated, specifying an AR(1) structure can lead to overly conservative estimates when it is not.
2. In step (2.), we sample subjects *with replacement* and then for each drawn subject, randomly draw one bootstrap estimate for each of their model parameters based on the mean and standard errors derived from the `gnls` estimate.

Just as with the v1 bootstrap, these bootstrap estimates are used to create test statistics T_t at each time point, written

$$T_t = \frac{(\overline{p}_{1t} - \overline{p}_{2t})}{\sqrt{s_{1t}^2 + s_{2t}^2}}, \tag{9}$$

where $\overline{p}_{gt}$ and s^2_{gt} are mean and standard deviation estimates at each time point for groups 1 and 2, respectively. Finally, just as in Oleson 2017, one can use the autocorrelation of the T statistics to create a modified α for controlling the FWER.

[parenthetical] A paired bootstrapping can be implemented by performing this same algorithm but ensuring that at each iteration of the bootstrap the same subjects are sampled with replacement in each group. This happened by default in the original implementation as each subject was retained in each iteration of the bootstrap.

3.2 Permutation Testing

In addition to the v2 bootstrap, we also introduce a permutation method for hypothesis testing. The permutation method proposed is analogous to a traditional permutation method, but with an added step mirroring that of the previous in capturing the within-subject variability. For a specified FWER of α, the proposed permutation algorithm is as follows:

1. For each subject, fit the nonlinear function with *optional* AR(1) autocorrelation structure for model errors. Assuming large sample normality, the sampling distribution of each estimator can be approximated by a normal distribution with mean corresponding to the point estimate and standard deviation corresponding to the standard error
2. Using the mean parameter estimates derived in (1.), find each subject's corresponding fixation curve. Within each group, use these to derive the mean and standard deviations of the population level curves at each time point, denoted $\overline{p}_{jt}$ and s^2_{jt} for $j = 1, 2$. Use these values to compute a test statistic T_t at each time point,

$$T_t = \frac{|\overline{p}_{1t} - \overline{p}_{2t}|}{\sqrt{s^2_{1t} + s^2_{2t}}}. \tag{10}$$

   This will be our observed test statistic.
3. Repeat (2) P additional times, each time shuffling the group membership between subjects. This time, we fitting each subject's corresponding fixation curve, draw a new set of parameter estimates using the distribution found in (1). Recalculate the test statistics T_t, each time retaining the maximum value. This collection of P statistics will serve as our null distribution which we denote $\widetilde{T}$ (or whatever we wanna call it). Let $\widetilde{T}_\alpha$ be the 1 - $\alpha/2$ quantile of $\widetilde{T}$
4. Compare each of the observed T_t with $\widetilde{T}_\alpha$. Areas where $T_t > \widetilde{T}_\alpha$ are designated significant.

Paired permutation testing is implemented with a minor adjustment to step (3). Instead of permuting all of the labels between groups, choose one group and randomly assign each subject to either retain their current group membership or to change groups. Make the corresponding reassignment to members in the second group. This ensures that each permuted group contains one observation from each subject.

[Note sure where to put this –] When permutation testing is used (and maybe I just explain this in bdots paper, not here) we still use modified bootstrap for computing confidence intervals which are given when plotted.

4 Type I Error Simulations

When now go about comparing the type I error rate of the three methods just described. In doing so, we will establish several conditions under which the observed subject data may have been generated or fit. This includes two conditions for the mean structure, two conditions for the error structure, paired and unpaired data, and data fit with and without an AR(1) assumption. Considering each permutation of this arrangement

results in sixteen different simulations. Each simulation will then be examined for type I error using each of the three methods described (I said that twice).

4.1 Data Generation

Data was generated according to three conditions: mean structure, error structure, and paired status. We assume that each subject's data was of the general form

$$y_{it} = f_{\theta_i}(t) + \epsilon_{it}. \tag{11}$$

We further assume that each group drew subject-specific parameters from a normal distribution,

$$\theta_i \sim N(\mu_\theta, V_\theta). \tag{12}$$

Mean Structure In all of the simulations presented, the distribution used in Equation 12 was empirically determined from data on normal hearing subjects in the VWP (Farris-Trimble et al., 2014 [**?**]). Parameters used were those fit to fixations on the Target, following the functional form of Equation 3. Under the assumption of between-subject homogeneity, we set $V_\theta = 0$, assuring that each of the subjects' observations is derived from the same mean structure, differing only in their observed error.

Error Structure The error structure was of the form

$$e_{it} = \phi e_{i,t-1} + w_{it}, \quad w_{it} \sim N(0, \sigma) \tag{13}$$

where the w_{it} are iid with $\sigma = 0.025$. ϕ corresponds to an autocorrelation parameter and is set to $\phi = 0.8$ when the generated data is to be autocorrelated and set to $\phi = 0$ when we assume the errors are all independent and identically distributed.

Paired Data Finally, we consider the paired data, which differs in creation according to the mean structure. In the case in which $V_\theta \neq 0$, we simply used the same value of θ_i for the ith subject in each group, allowing the only difference to be that corresponding to the error structure. In the case when $V_\theta = 0$, however, it was already the case that the set of parameters were the same between subjects in each group (and indeed for all subjects in both groups). As such, letting the observed data for subject i in group A be denoted y_{iA}, we

set

$$y_{iB} = y_{iA} + N(0, \sigma)$$

so that the only difference between paired subjects was uncorrelated normal noise at each time point. (I get that this is questionable, but what is the alternative? Truly it would just be to half the degrees of freedom in the t-test, which isn't really comparing IRL paired data)

Each set of conditions generates two groups, with $n = 25$ subjects in each group, with $N = 100$ simulated trials for each subject. Columns in the tables indicate homogeneity assumption, whether or not an AR(1) error structure was used, and if correlation was specified in the fitting function. Paired and unpaired tests are in different tables. Each simulation was conducted 100 times to determine the rate of type I error.

4.2 Results

yeah, idk, here are the results by section. Updated permutation are in parentheticals. New batch of 100 running now but the hpc is physically incapable of handling more than one of these jobs at a time. No idea why. Probably because there is no god

4.2.1 FWER

Maybe I should replace TRUE/FALSE with Yes/No? Also having the paired/unpaired feels a little like a distraction , especially when there isn't all that much to say. I might consider moving those to appendix or something.

Table 1 demonstrates the FWER across a number of scenarios. Note that the first two lines replicate the simulation found in Oleson 2017 (with different parameters), confirming that under the assumption of homogeneity and AR(1) error, failing to account for autocorrelation in the fitting function dramatically inflates the family-wise error rate. We see a similar phenomenon with the v2 bootstrap and permutation tests, though to a much smaller degree.

Most notably here are the results under the assumption of homogeneity, where the FWER of the v1 bootstrap increases dramatically to the point of being unusable. We should consider how to best handle the fact that researchers have spent two years conducting studies in the VWP using this truly fucking awful method. Alternatively, the v2 bootstrap and permutation test perform as expected, with observed FWER below the nominal level.

Heterogeneity	AR(1)	AR(1) Specified	V1 Bootstrap	V2 Bootstrap	Permutation (updated)
FALSE	TRUE	TRUE	0.06	0.01	0.21 (0.05)
FALSE	TRUE	FALSE	0.87	0.08	0.18 (0.11)
FALSE	FALSE	TRUE	0.08	0.00	0.14 (0.03)
FALSE	FALSE	FALSE	0.15	0.02	0.21 (0.04)
TRUE	TRUE	TRUE	0.92	0.03	0.03
TRUE	TRUE	FALSE	0.96	0.02	0.04
TRUE	FALSE	TRUE	0.99	0.05	0.01
TRUE	FALSE	FALSE	1.00	0.05	0.03

Table 1: FWER for empirical parameters (unpaired)

Paired data is given in Table 2. Matching the conclusions drawn from Table 1, we only note here that both the v2 bootstrap and permutation test remain valid under the paired assumption.

Heterogeneity	AR(1)	AR(1) Specified	V1 Bootstrap	V2 Bootstrap	Permutation
FALSE	TRUE	TRUE	0.12	0.02	0.03
FALSE	TRUE	FALSE	0.86	0.08	0.03
FALSE	FALSE	TRUE	0.09	0.01	0.01
FALSE	FALSE	FALSE	0.14	0.01	0.03
TRUE	TRUE	TRUE	0.49	0.02	0.01
TRUE	TRUE	FALSE	0.94	0.03	0.02
TRUE	FALSE	TRUE	0.72	0.02	0.00
TRUE	FALSE	FALSE	0.74	0.04	0.00

Table 2: FWER for empirical parameters (paired)

4.2.2 Median per comparison error rate

We present also a table of the median per comparison error rate. Make of this what you will. There is also the collection of plots in Figure 1, but as it doesn't really convey much, I might just delete it all together.

Heterogeneity	AR(1)	AR(1) Specified	V1 Bootstrap	V2 Bootstrap	Permutation (updated)
FALSE	TRUE	TRUE	0.01	0.00	0.04 (0.00)
FALSE	TRUE	FALSE	0.31	0.00	0.04 (0.02)
FALSE	FALSE	TRUE	0.00	0.00	0.02 (0.00)
FALSE	FALSE	FALSE	0.00	0.00	0.03 (0.00)
TRUE	TRUE	TRUE	0.51	0.01	0.01
TRUE	TRUE	FALSE	0.76	0.01	0.00
TRUE	FALSE	TRUE	0.86	0.01	0.00
TRUE	FALSE	FALSE	0.81	0.01	0.00

Table 3: median per comparison error rate (unpaired)

Heterogeneity	AR(1)	AR(1) Specified	V1 Bootstrap	V2 Bootstrap	Permutation
FALSE	TRUE	TRUE	0.03	0.00	0.00
FALSE	TRUE	FALSE	0.26	0.00	0.00
FALSE	FALSE	TRUE	0.00	0.00	0.00
FALSE	FALSE	FALSE	0.01	0.00	0.00
TRUE	TRUE	TRUE	0.13	0.00	0.00
TRUE	TRUE	FALSE	0.52	0.02	0.00
TRUE	FALSE	TRUE	0.38	0.01	0.00
TRUE	FALSE	FALSE	0.44	0.01	0.00

Table 4: median per comparison error rate (paired)

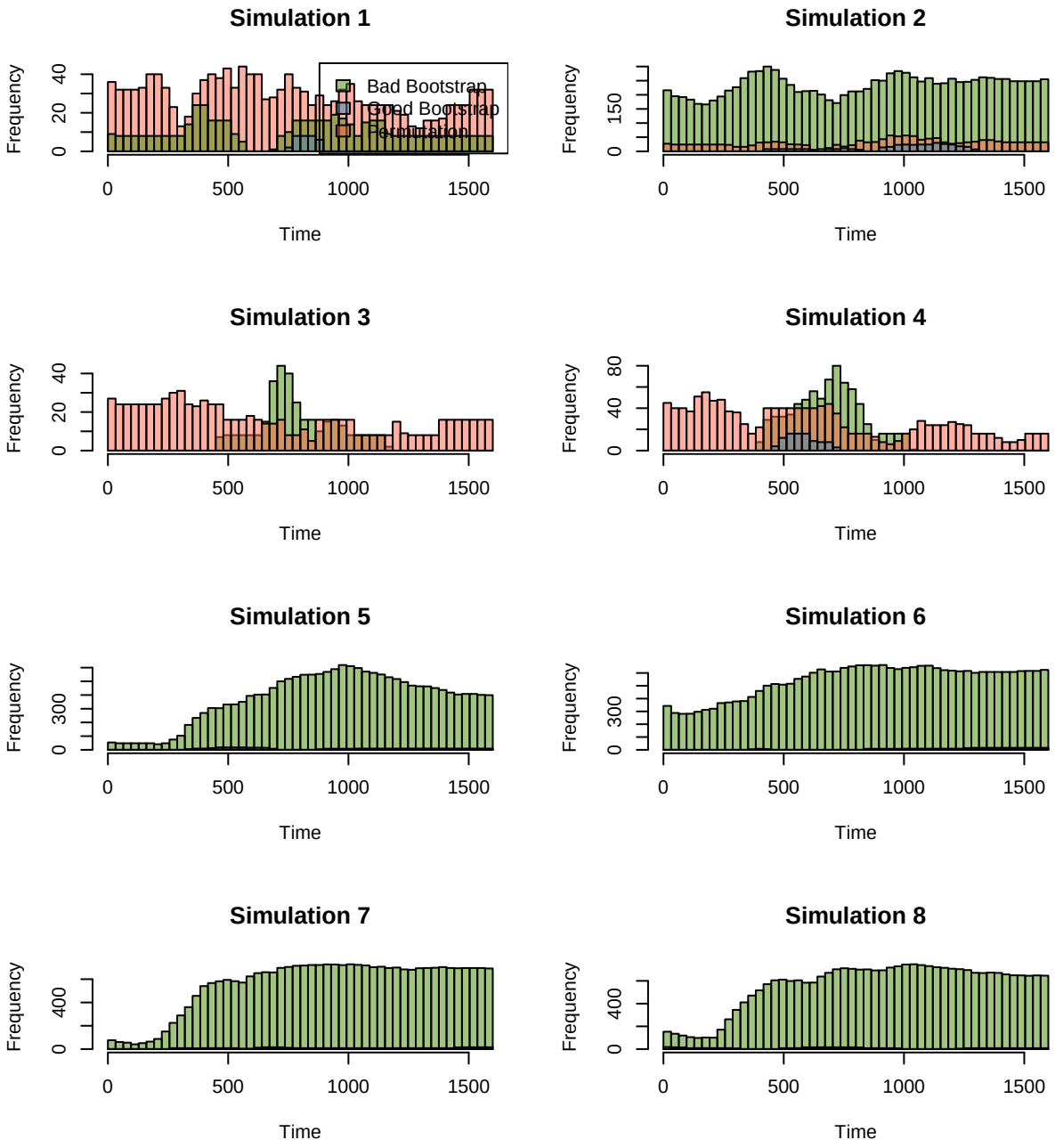

Figure 1: quick put together of what this might look like, frequency of TIE at each spot. Since there is no obvious trend here, though, I may just omit this entirely

4.3 FWER Discussion

The results here are pretty conclusive and speak for themselves. The v1 bootstrap performs as advertised in its original publication, but even then, not really any better than the two new methods proposed. And

in the far more likely case of heterogeneous subject parameters, the v1 bootstrap is a trainwreck.

Transition sentence to power simulations

5 Power Simulations

All this talk on the type I error rate sure is interesting, but what good is having a low type I rate if we are just trading it in for type II? In this hard hitting piece of investigative journalism, we set out to determine the empirical power of the proposed methods under a variety of conditions, similar to those above but excluding the case of paired observations. Maybe if there's time.

To determine power, two experimental groups were simulated with mean structures of the following form:

$$y = \begin{cases} b & x < 0 \\ mx + b & x \geq 0 \end{cases} \tag{14}$$

The first simulated group was "No Effect", with intercept and slope parameters normally distributed and standard deviation σ. The second group, the "Effect" group, was similarly distributed, but with the slope parameter having mean value of $\mu = 0.025$. Absolute values were taken to ensure that all of the parameters were positive, and simulations were run with group variability values of $\sigma = 0.005$ and $\sigma = 0.025$. The error structure was identical to that in the FWER simulations, with both an AR(1) error structure and independent noise included. Finally, we limited consideration to three possible scenarios: first, we assumed the conditions presented in Oleson 2017, assuming homogeneity between subject parameters and an AR(1) error structure, with the model fitting performed assuming autocorrelated errors. For the remaining scenarios, we assumed heterogeneity in the distribution of subject parameters, simulated with and without an AR(1) error structure. In both of these scenarios, we elected to *not* fit the model assuming autocorrelated errors. This was for two reasons: first, simulations exploring the type I error rate suggested that models fit with the autocorrelation assumption tended to be conservative. Second, and given the results of the first, this makes setting the assumption of autocorrelation to FALSE in `bdots` seem like a sensible default, and as such, it would be of interest to see how the model performs in these cases.

For each subject, parameters for their mean structure given in Equation 14 were drawn according to their group memebership and fit using `bdots` on the interval (-0.1,1). Groups were then compared using each of the methods presented (names?). By including the interval (-0.1,0) where there is no true difference, we are able to mitigate the effects of over-zealous methods, and we present this information in the following way: any bootstrapped or permuted resulting identifying the region (-0.1, 0) as being significantly different was

marked as having a type I error, regardless of other regions identified. The proportions of simulations for which this occurred in given in the column labeled α (maybe we could call this FWER?). The next column, β, is the type II error rate, giving the proportion of simulations in which no differences were identified. And finally in the remaining greek letter column is $1 - \beta - \alpha$, a modified power metric giving the proportion of simulations in which only differences in the correct region were identified. The remaining columns given a partial summary of the earliest onset of detection. As the true difference occurs on the interval $t > 0$, smaller values indicate greater power in detecting differences. Finally, a base R plot of the power at each time point is given in Figure 2. This plot represents the true power, though note that this does not take into account the rate of type I errors which in all cases occur to the left of the red dashed line.

(I know I used too many colons)

100 simulations were conducted for each scenario. Here are the results.

5.1 Results

5.1.1 Bootstrap v1

We note a few things here. Considering the power results for the v1 bootstrap in Table 5, we first observe that under the assumption of heterogeneity, the v1 bootstrap has a type I error rate so poor as to not be worthy of further consideration. Even then, under the homogeneity assumption, the v1 bootstrap really performs no better than the v2 bootstrap or permutation test, with perhaps slightly greater power than the v2 bootstrap and overly conservative type I error rate compared to the permutation test. This is best illustrated in the first row of Figure 2.

Heterogeneity	AR(1)	σ	α	β	$1 - \beta - \alpha$	1st Qu.	Median	3rd Qu.
FALSE	TRUE	0.005	0.01	0.00	0.99	0.247	0.303	0.339
FALSE	TRUE	0.025	0.00	0.00	1.00	0.265	0.324	0.347
TRUE	FALSE	0.005	0.90	0.00	0.10	0.008	0.017	0.030
TRUE	FALSE	0.025	0.97	0.00	0.03	0.012	0.015	0.016
TRUE	TRUE	0.005	0.89	0.00	0.11	0.048	0.059	0.066
TRUE	TRUE	0.025	0.91	0.00	0.09	0.023	0.032	0.037

Table 5: Power for v1 bootstrap

5.1.2 Bootstrap v2

Performs comparably in homogeneous means. Low type I error, type I error hovering around 10%-20%

Heterogeneity	AR(1)	σ	α	β	$1-\beta-\alpha$	1st Qu.	Median	3rd Qu.
FALSE	TRUE	0.005	0.00	0.00	1.00	0.296	0.346	0.392
FALSE	TRUE	0.025	0.00	0.00	1.00	0.316	0.364	0.395
TRUE	FALSE	0.005	0.00	0.12	0.88	0.373	0.576	0.756
TRUE	FALSE	0.025	0.01	0.17	0.82	0.423	0.555	0.734
TRUE	TRUE	0.005	0.01	0.20	0.79	0.357	0.522	0.691
TRUE	TRUE	0.025	0.00	0.12	0.88	0.444	0.555	0.737

Table 6: Power for v2 bootstrap

5.2 Permutation

Comparable in homogeneous means case with closer to nominal type I error. This has most reasonable balance (it seems) of controlling type I error and achieving power. Cool

Heterogeneity	AR(1)	σ	α	β	$1-\beta-\alpha$	1st Qu.	Median	3rd Qu.
FALSE	TRUE	0.005	0.05	0.00	0.95	0.259	0.299	0.332
FALSE	TRUE	0.025	0.10	0.00	0.90	0.275	0.302	0.334
TRUE	FALSE	0.005	0.03	0.05	0.92	0.454	0.609	0.748
TRUE	FALSE	0.025	0.08	0.08	0.84	0.490	0.610	0.721
TRUE	TRUE	0.005	0.08	0.12	0.80	0.418	0.578	0.743
TRUE	TRUE	0.025	0.00	0.06	0.94	0.493	0.619	0.746

Table 7: Power for permutation

5.2.1 Summary of methods

A general estimate of how well each of these methods does in a variety of conditions can be seen by taking the mean of the summary statistics across each of the trials, given in Table 8.

Method	α	β	$1-\beta-\alpha$	1st Qu.	Median	3rd Qu.
Bootstrap V1	0.613	0.000	0.387	0.101	0.125	0.139
Bootstrap V2	0.003	0.102	0.895	0.368	0.486	0.617
Permtuation	0.057	0.052	0.892	0.398	0.503	0.604

Table 8: Summary of methods for Type II error. It's worth considering presenting the means separately depending on the $V \neq 0$ assumption, because that fucking tanks the Type II error for all of them

As we can see, the permutation method is the most canonical of the methods considered, with a type I error rate close to the nominal $\alpha = 0.05$ and a type II error rate of $\beta = 0.195$, corresponding to approximately 80% power. The V2 bootstrap, alternatively, is rather conservative, trading a portion of its power for controlling the type I error rate close to zero. Finally, the V1 bootstrap is a poor contender for identifying differences in time series, with its utility limited to the strict assumptions under which it was originally presented. Even then, it performs generally no better than the other methods, but with substantially greater risk should the underlying assumptions not hold. I believe that this is conclusive enough evidence to make this not even an option in `bdots`, even if somebody could find a limited case in which they knew for certain the assumptions held. But idk. Maybe it can be an easter egg. Or require the user to type a full paragraph acknowledging the risks before it will run. But really I think we should just remove it.

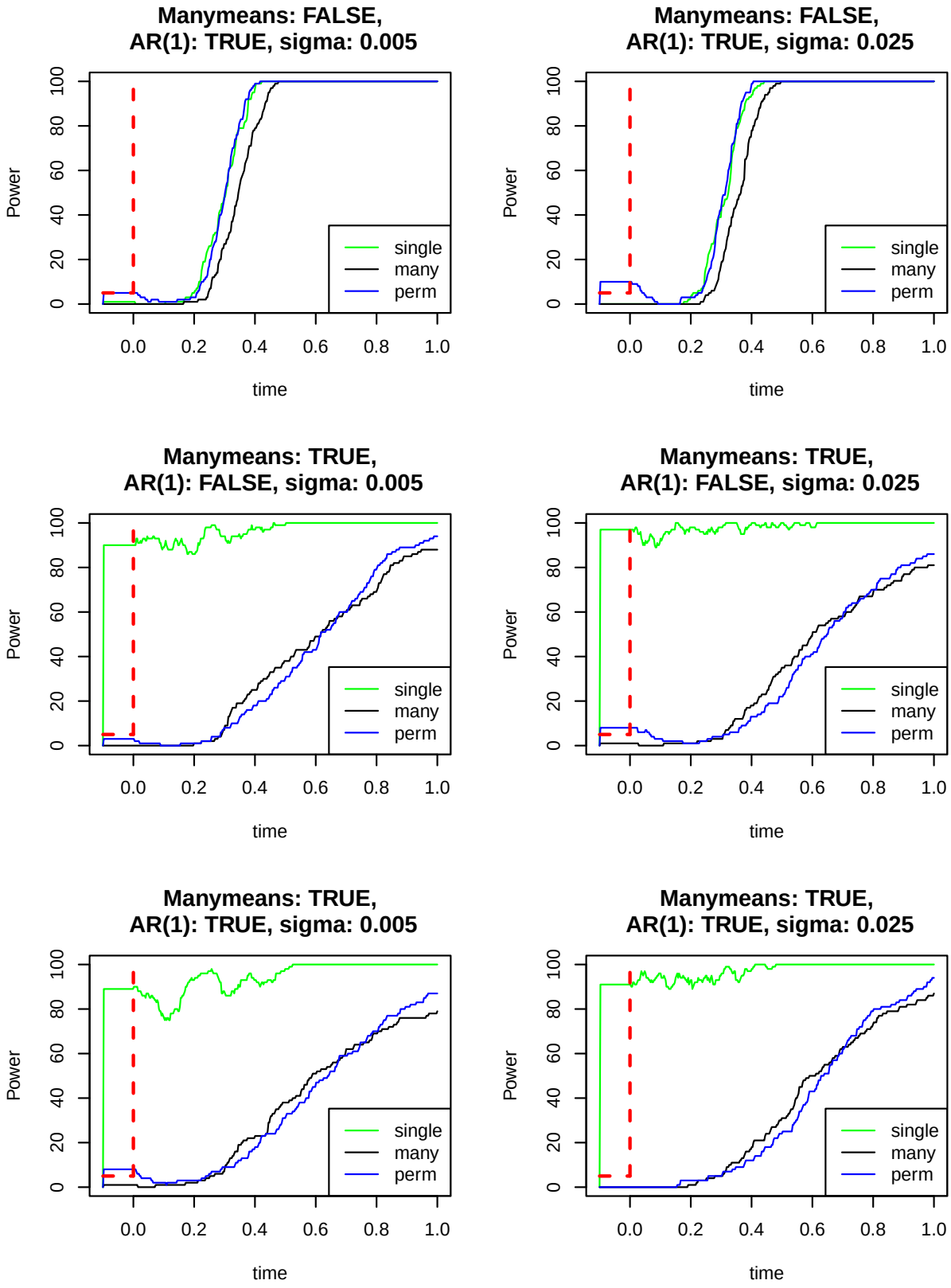

Figure 2: check out this badass plot in base R. Note that the bump in type I error before flattening out around 0.1 is a consequence of simulations in which the no effect group, on average, has a larger intercept than the effect group. When this occurs, the lines "cross" in the 0 to 0.2 range, making it seem like there is no effect there

6 Discussion and concluding remarks

We set out to interrogate the validity of the v1 bootstrap assumptions and to propose two alternative methods that would be more robust under a greater variety of assumptions. In doing so, we demonstrated conclusively the utility of the v2 bootstrap and permutation tests while also highlighting a major shortcoming of the v1 bootstrap. It's worth noting, however, that the FWER adjustment proposed in [**?**] is still valid, if not slightly conservative, and with power similar to that of the permutation method, and this will remain an option in version 2.0 of `bdots`.

There are a few limitations to the current paper that are worthy of investigation. First, limited consideration was given to the effect of sample density on the observed type I error rate or power. As the fitting function in `bdots` simply returns a set of parameters, one could conceivably perform any of the methods presented on any arbitrary collection of points, whether or not any data were observed there. This extends itself to the condition in which subjects were sampled at heterogeneous time points, as may be the case in many clinical settings. What impact this may have or how to best handle these cases remains investigated. The current implementation of `bdots` takes the union of observed time points, though this runs the risk of extrapolating many subjects past what they were ever observed. It would be of interest to know if either the permutation or v2 bootstrap perform better in these situations, and if both retain their validity under increasingly suspect conditions. Finally, in noting the rather conservative FWER estimates for both the v2 bootstrap and the permutation test, it would be worthwhile investigating if *not* resampling subject-specific parameters from the distribution provided by `gnls` would retain an acceptable FWER while increasing power.

We conclude pretty much by noting that even in the best case presented in Oleson 2017, these other methods do an identical job, and in situations in which these assumptions are wrong, it is a veritable train wreck. It seems that the $1 - \beta - \alpha$ (whatever this is called) is nearly identical between the v2 bootstrap and permutation, whereas the type II error is much greater in the bootstrap. This seems justification enough for making the permutation method the new default in `bdots` or, should i say PDOTS. It is conceivable that the assumptions presented in Oleson 2017 would have their place, say repeated observation from the same mechanism (i.e., not vwp data), in which case v1 bootstrap has optimal performance. Still, users of `bdots` will need to go out of their way in order to do so, possibly with a warning. Because really, even in that case, it hardly does any better than permutation, perhaps with a bit smaller of a type I error.

The End.

Appendix

Could include oleson 2017 parameters just to say we did it and verifying the two results that they had previously found (i.e., we have implemented this correctly). Commented out for now