With great power comes greater responsibility:

cheating with bdots

Last compiled: Sunday 19th February, 2023 at 16:05

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

A standard problem in psycholinguistics, and the cognitive sciences in general, is that of statistically analyzing a process unfolding in time. Particularly in the case of comparing a process between experimental groups in time, techniques appropriate for demonstrating showing that differences *exist* offer little towards the goal of identifying *when* they exist when the time windows are not specified in advance, with the area under a curve (AUC) representing an example of this. One may consider instead testing two time series at each point in time, using a method such a cluster based permutation testing [**?**] in which test-statistics are computed at each time, with adjacent significant tests being combined into single clusters. This is in an effort to address the problem of multiple comparisons and, by extension, control for the family wise error rate: by reducing adjacent test statistics into a single cluster, researchers work to control the FWER by simply reducing the number of tests. More recently has been the bootstrapped differences of time series (BDOTS) [**?**], using bootstrapped fits of subject-level curves in conjunction with a modified Bonferroni correction to the significance level to control the family wise error rate in the presence of highly correlated test statistics. A version of this was introduced in the R package `bdots` in 2018 [**?**].

A closer look at the specific conditions under which the original iteration of bdots was presented raises concerns, however, involving quite restrictive assumptions that are unlikely to be met in many, if not most, situations. This includes data typically collected in the Visual World Paradigm (VWP), context in which the underlying methodology was first proposed. Specifically, it assumed a homogeneous mean structure within each of the groups being compared, with no between-subject variability to be accounted for. Empirical data collected in a variety of (VWP) contexts can be used to show that such an assumption is unlikely to be true, with the resulting type I error rate being unacceptably high.

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 FWER adjustment method presented for autocorrelated errors. In addition, we offer a 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 type I error rate across a number of experimental conditions, and finally we conclude with a demonstration of power in the competing methods.

2 Detail on the original

Most generally the original bdots algorithm, which we will call the *homogeneous bootstrap*, proposed that we have empirically observed data in time resulting from a parametric function f with an associated error:

$$y_t = f(t|\theta) + \epsilon_{it} \tag{1}$$

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) 1000 times 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.

Population means and standard deviations at each time point for each of the groups were used to construct a series of (correlated) test statistics, where the family wise error rate was controlled by using the modified Bonferonni correction introduced in Oleson et. al., 2017 to test for significance.

3 Proposed Methods

As is more typically the case, subjects within a group demonstrate considerable variability in their mean parameter estimates. In such a case, there is no presumption that $\theta_i = \theta_j$, and accounting for between-subject variability within a group 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:

The distribution of parameters for subjects $i = 1, \ldots, n_g$ in group $g = 1, \ldots, G$ follows the distribution

$$\theta_i \sim N(\mu_g, V_g). \tag{3}$$

The uncertainty present in our estimation of θ_i can be accounted for by treating the standard errors derived when fitting the observed subject data to Equation 1 as estimates of their standard deviations. This gives us a distribution for the subject's estimated parameter,

$$\hat{\theta}_i \sim N(\theta_i, s_i^2). \tag{4}$$

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_g. For example, let θ_{ib}^* represent a bootstrapped sample for subject i in bootstrap $b = 1, \dots, B$, where

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

If we were to sample *without replacement*, we would obtain a homogeneous mean value from the bth bootstrap for group g, $\theta_{bg}^{(hom)}$, where

$$\theta_{bg}^{(hom)} = \frac{1}{n_g} \sum \theta_{ib}^*, \quad \theta_{bg}^{(hom)} \sim N\left(\mu_g, \frac{1}{n_g^2} \sum s_i^2\right). \tag{6}$$

Such an estimate captures the totality of the within-subject variability with each draw but fails to account for the variability in the group overall. For this reason, we sample the subjects *with* replacement, creating the heterogeneous bootstrap mean $\theta_{bg}^{(het)}$, where again each θ_{ib}^* follows the distribution in Equation 5, but the heterogeneous bootstrapped group mean now follows

$$\theta_{bg}^{(het)} \sim N\left(\mu_g, \frac{1}{n_g} V_g + \frac{1}{n_g^2} \sum s_i^2\right). \tag{7}$$

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

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 homogeneous bootstrap, these bootstrap estimates are used to create test statistics T_t at each time point, written

$$T_t^{(b)} = \frac{(\overline{p}_{1t} - \overline{p}_{2t})}{\sqrt{s_{1t}^2 + s_{2t}^2}}, \tag{8}$$

where $\overline{p}_{gt}$ and s_{gt}^2 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 $T_t^{(b)}$ statistics to create a modified α for controlling the FWER.

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 heterogeneous 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:

need to add math here

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_{jt}^2 for $j = 1, 2$. Use these values to compute a test statistic T_t at each time point,

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

   This will be our observed test statistic.

3. Repeat (2) P additional times, each time shuffling the group membership between subjects. This time, when constructing each subject's corresponding fixation curve, draw a new set of parameter estimates using the distribution found in (1). Recalculate the test statistics $T_t^{(p)}$, retaining the maximum value from each permutation. This collection of P statistics will serve as our null distribution which we denote $\widetilde{T}$. Let $\widetilde{T}_\alpha$ be the 1 - α quantile of $\widetilde{T}$

4. Compare each of the observed $T_t^{(p)}$ with $\widetilde{T}_\alpha$. Areas where $T_t^{(p)} > \widetilde{T}_\alpha$ are designated significant.

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

4 Type I Error Simulations

We now go about comparing the type I error rate of the three methods just described. In doing so, we will consider 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 settings. Each simulation will then be examined for type I error using each of the three methods described.

4.1 Data Generation

Data was generated according to Equation 1, with the parametric function $f(t|\theta)$ belonging to the family of four-parameter logistic curves defined:

$$f_\theta(t) = \frac{p-b}{1+\exp\left(\frac{4s}{\text{p}-b}(x-t)\right)} + b \tag{10}$$

where $\theta = (p, b, s, x)$, the peak, baseline, slope, and crossover parameters, respectively.

We further assume that each group drew subject-specific parameters from a normal distribution, with subject $i = 1, \ldots, N$ in group $g = 1, \ldots, G$ following the distribution in Equation 3.

Could also expression this $\theta_i^{(g)}$? Though that may be cumbersome

Mean Structure In all of the simulations presented, the distribution used in Equation **??** 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 10. Under the assumption of between-subject homogeneity, we set $\theta_i = \theta_j$ for all subjects i, j, assuring that each of the subjects' observations is derived from the same mean structure, differing only in their observed error. We will call this the homogeneous means assumption.

Error Structure The error structure is of the form

$$e_{it} = \phi e_{i,t-1} + w_{it}, \quad w_{it} \sim N(0, \sigma) \tag{11}$$

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 considered paired data, though this is only a sensible condition under the assumption of heterogeneous means. In order to construct paired data, we simply used the same parameter estimates for each subject between groups, with the observed data between subjects differing only in the observed error.

Each set of conditions generates two groups, with $n = 25$ subjects in each group, with timepoints $t = 0, 4, 8, \ldots, 1600$ in each trial and with 100 simulated trials for each subject. Columns in the tables indicate homogeneity of means assumption, whether or not an AR(1) error structure was used in constructing the data, and if autocorrelation was specified in the fitting function. The last conditions helps assess the impact of correctly identifying the type of error when conducting an analysis in `bdots`. Each simulation was conducted 100 times to determine the rate of type I error.

4.2 Results

We consider the efficacy the methods under each of the simulation settings with an analysis of the family wise error rate (FWER) and the median per-comparison error rate. The first of these details the proportion of simulations under each condition that marked *at least* one time point as being significantly different between the two groups. This is critical is understanding each method's ability to correct adjust for the multiple testing problem associated with testing each of the observed time points. These are presented in Table 1 and Table 2 for unpaired and paired data, respectively.

Complimenting the FWER estimate is an estimate of the median per-comparison rate. For each time point across each of the simulations, we computed the proportion of times in which that time was determined significant. The median of these values across all time points is what is considered. This metric gives a sense of magnitude to the binary FWER; for example, a situation in which there was a high FWER and low median per-comparison rate would indicate that the type I error within a particular time series would be sporadic and impact limited regions. Large median per-comparison rates indicate that large swaths of a time series frequently sustain type I errors. The median per-comparison rates for unpaired and paired simulations are presented in Table 3 and Table 4.

4.2.1 FWER

There are a few things of immediate note when considering the results of Table 1. First, we see from the first two settings of the unpaired simulations that the type I error rates for the homogenous bootstrap are consistent with those presented in [**?**], confirming the importance of specifying the existence of autocorrelation in the `bdots` fitting function when autocorrelated error is present. By contrast, this is far less of a concern when using the heterogeneous bootstrap or permutation testing, both of which maintain a FWER near the nominal alpha, regardless of whether or not the error structure was correctly identified. This continuous to be true under the homogenous mean assumption when the true error structure is not autocorrelated. Interestingly, the performance of the homogeneous bootstrap falters here despite theoretical consistency with the simulation settings. I am rerunning this condition again now to make sure there wasn't an error.

The most striking results of this, however, appear when the data generation assumes a heterogeneous mean structure. While both the heterogeneous bootstrap and the permutation test maintain a FWER near the nominal alpha, the homogeneous bootstrap fails entirely, with a FWER > 0.9 in all cases.

Heterogeneity	Autocorrelated	AR(1)	Homogeneous	Heterogeneous	Permutation
Assumption	Error	Specified	Bootstrap	Bootstrap	
No	Yes	Yes	0.06	0.01	0.08
No	Yes	No	0.87	0.08	0.00
No	No	Yes	0.08	0.00	0.06
No	No	No	0.15	0.02	0.01
Yes	Yes	Yes	0.92	0.03	0.05
Yes	Yes	No	0.96	0.02	0.08
Yes	No	Yes	0.99	0.05	0.03
Yes	No	No	1.00	0.05	0.06

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 permutation test and heterogeneous bootstraps maintain a valid FWER under the assumption of paired data.

Heterogeneity	Autocorrelated	AR(1)	Homogeneous	Heterogeneous	Permutation
Assumption	Error	Specified	Bootstrap	Bootstrap	
Yes	Yes	Yes	0.49	0.02	0.01
Yes	Yes	No	0.94	0.03	0.02
Yes	No	Yes	0.72	0.02	0.00
Yes	No	No	0.74	0.04	0.00

Table 2: FWER for empirical parameters (paired)

4.2.2 Median per comparison error rate

We next consider the median comparison rate, which offers some insight into the FWER. In particular, consider the situation in which in Table 3, in the fourth row we see a median per-comparison error rate of 0.00 for the homogeneous bootstrap, despite Table 1 indicating a FWER of 0.15. This is a consequence of the majority of the type I errors occuring in a relatively limited region. In contrast, the median per-comparison error rate of the homogeneous bootstrap under the assumption of heterogeneity suggests that the type I errors are widespread and not limited to any particular area.

It is also worth commenting on the permutation test median per-comparison error rate in Table 3; combined with a FWER near the nominal 0.05, these values suggest that errors are likely distributed across the entire range rather than limited to a small area (which would result in a MPCR of 0).

I confirmed this by looking both at this histograms and inspecting the data manually

Heterogeneity	Autocorrelated	AR(1)	Homogeneous	Heterogeneous	Permutation
Assumption	Error	Specified	Bootstrap	Bootstrap	
No	Yes	Yes	0.01	0.00	0.04
No	Yes	No	0.31	0.00	0.04
No	No	Yes	0.00	0.00	0.02
No	No	No	0.00	0.00	0.03
Yes	Yes	Yes	0.51	0.01	0.01
Yes	Yes	No	0.76	0.01	0.00
Yes	No	Yes	0.86	0.01	0.00
Yes	No	No	0.81	0.01	0.00

Table 3: median per comparison error rate (unpaired)

Heterogeneity	Autocorrelated	AR(1)	Homogeneous	Heterogeneous	Permutation
Assumption	Error	Specified	Bootstrap	Bootstrap	
Yes	Yes	Yes	0.13	0.00	0.00
Yes	Yes	No	0.52	0.02	0.00
Yes	No	Yes	0.38	0.01	0.00
Yes	No	No	0.44	0.01	0.00

Table 4: median per comparison error rate (paired)

4.3 Discussion

...

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:

$$y = \begin{cases} b & x < 0 \\ mx + b & x \geq 0 \end{cases} \tag{12}$$

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 12 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 Yes 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 1. 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

Method	Heterogeneity	AR(1)	α	β	1 - α - β	1st Qu.	Median	3rd Qu.
Hom. Boot	No	Yes	0.00	0.00	1.00	0.14	0.17	0.21
Hom. Boot	Yes	No	0.96	0.00	0.04	0.02	0.02	0.02
Hom. Boot	Yes	Yes	0.89	0.00	0.11	0.01	0.01	0.03
Het. Boot	No	Yes	0.00	0.00	1.00	0.20	0.23	0.28
Het. Boot	Yes	No	0.00	0.13	0.87	0.38	0.46	0.56
Het. Boot	Yes	Yes	0.00	0.08	0.92	0.38	0.49	0.68
Perm	No	Yes	0.01	0.00	0.99	0.14	0.18	0.22
Perm	Yes	No	0.04	0.07	0.89	0.36	0.44	0.58
Perm	Yes	Yes	0.02	0.02	0.96	0.36	0.50	0.67

Table 5: Power for methods (possibly reorder?)

Method	Type I Error	Type II Error	Power	1st Qu.	Median	3rd Qu.
Hom. Bootstrap	0.617	0.000	0.383	0.056	0.070	0.085
Het. Bootstrap	0.000	0.070	0.930	0.321	0.395	0.503
Permtuation	0.023	0.030	0.947	0.287	0.373	0.491

Table 6: Summary of methods for Type II error

5.1.1 Bootstrap v1

We note a few things here. Considering the power results for the v1 bootstrap in Table **??**, 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 1.

5.1.2 Bootstrap v2

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

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

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 7.

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 7: 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.

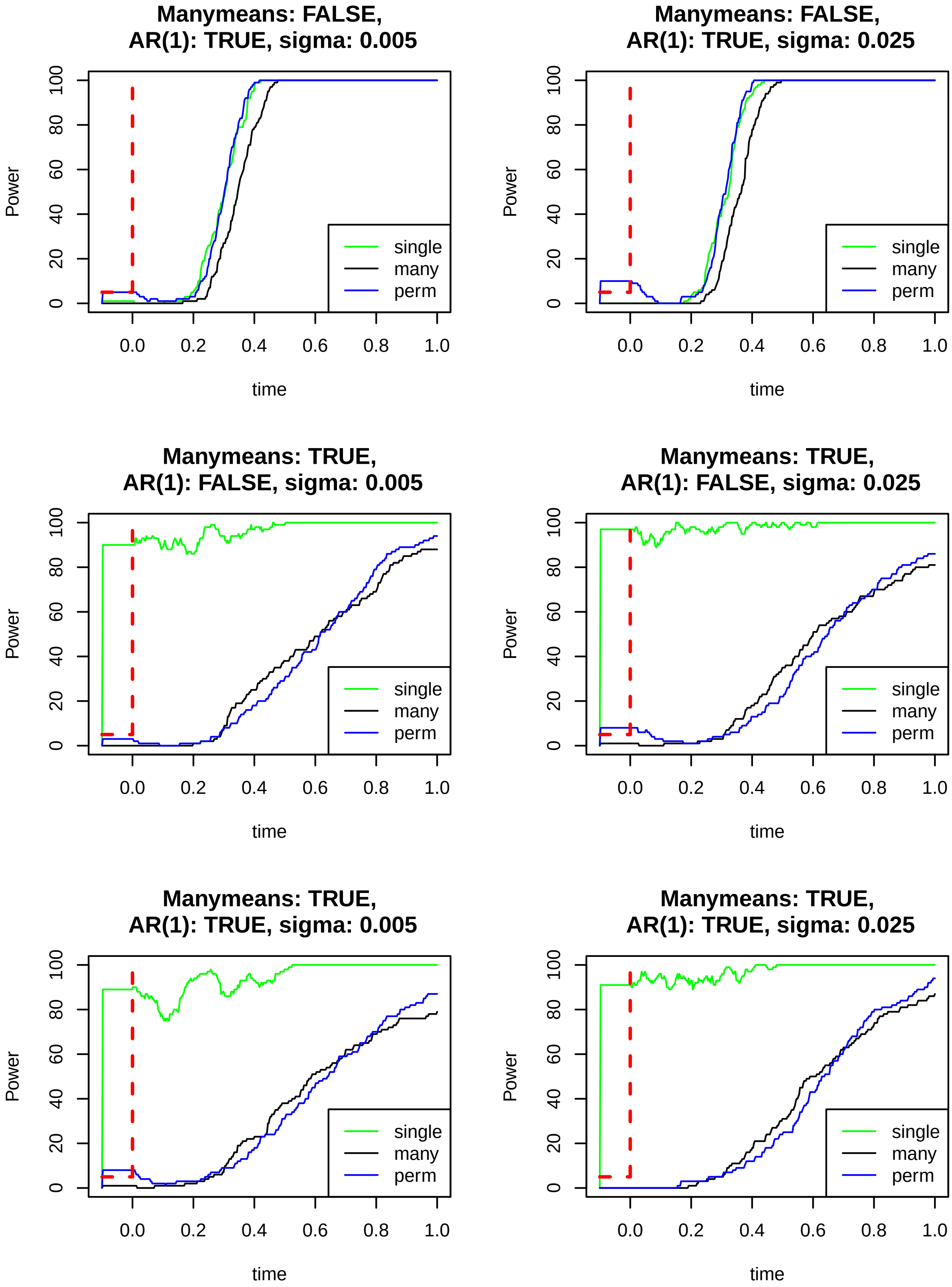

Figure 1: check out this badass plot in base R. Note that the bump in type I error before flattening out around 0.1 is a consequence of simulations in which the no effect group, on average, has a larger intercept than the effect group. When this occurs, the lines "cross" in the 0 to 0.2 range, making 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