With great power comes greater responsibility

Last compiled: Monday 17th April, 2023 at 18:15

1 Introduction

A problem ubiquitous throughout the sciences, though in the cognitive sciences especially, is that of statistically analyzing a process unfolding in time. In particular, we consider the problem of comparing a process in time as it evolves differentially between two or more experimental groups. And while there are many techniques for demonstrating *that* a difference exists, few offer any insight into *when*. Testing for temporal differences is complicated by the fact that when this process is continuous, there are often an arbitrarily large number of time points that could be compared. This is essentially a problem of multiple comparisons.

Various approaches have been proposed for addressing this issue, the most rudimentary of which involve binary tests for identifying the existence of any difference, such as the area under the curve (AUC). More sophisticated techniques involve the used of cluster-based permutation testing (?), whereby test statistics are computed at each observed time point with adjacent significant tests being combined into clusters. This controls the family-wise error rate (FWER) by reducing adjacent test statistics into a single cluster, thereby reducing the number of total tests. More recently, ? introduced a modified Bonferroni correction to a series of test statistics, using estimates of autocorrelation between statistics to make the appropriate adjustments to the significance levels to control FWER. This approach, which they named bootstrapped differences in time series, was introduced in the R package, `bdots` (?).

The modified Bonferonni correction used by `bdots` relies on the construction of estimated distributions of time series via bootstrapping for each experimental group. These distributions, in turn, are used to construct test statistics at each observed time point. A closer look at the original iteration presents concerns as it involves quite restrictive assumptions on the data that are unlikely to be met in many, if not most situations.

This includes data typically collected in the context of the Visual World Paradigm (VWP), a widely used experimental paradigm in language research that involves tracking eye movements from participants in response to spoken language. This is notable in that it was data collected in VWP research that motivated

the `bdots` methodology. There, it was maintained that the process of interest in each experimental group assumed a homogeneous mean structure, with no between-subject variability to be accounted for. Empirical data collected in a variety of (VWP) contexts can be used to demonstrate that this assumption is unlikely to be true, the consequence of which is a type I error rate that is unacceptably high. i don't think its worth mentioning here, but to address bob's concern, this does also assume no variability in the differences for a paired case, which is why hom. bootstrap still does poorly when the paired parameters aren't identical

Here, we present two alternative methods that accommodate flexibility in the assumptions made by the original bootstrapped differences in time series algorithm. First, we propose a modified bootstrapping procedure that adequately accounts for observed between-subject variability while retaining the FWER adjustment that addresses the autocorrelated test statistics. In addition, we offer a permutation test for identifying temporal differences between 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 with a mathematical description of the problem along with the proposed alternatives to the original bootstrapping algorithm. This is followed by a simulation estimating the FWER across a number of experimental conditions. Finally, we consider two separate simulations for the assessment of power among the competing methods: this includes one simulation with a simple piecewise-linear function relating power to effect sizes; the second simulation interrogates power when the magnitudes of between-subject and within-subject variability differ.

2   Methods

In each of the methods to be described, we begin with the observation of y_{it} for subjects $i = 1, \ldots, n$ over times $t = 1, \ldots, T$. Typically, these subjects fall into different groups $g = 1, \ldots, G$, with each group containing n_g subjects. We further assume that the empirically observed data follows from a parametric function f with associated error:

$$y_{it} = f(t|\theta_i) + \epsilon_{it} \tag{1}$$

where θ_i is the subject-specific parameterization of f, and with

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

Under this paradigm, the errors are permitted to be either iid normal (with $\phi = 0$) or have an AR(1) structure, with $0 < \phi < 1$. It is generally assumed that the observed data across subjects make up a

distribution for each group, specified as either a multivariate distribution of the parameters θ_i or a distribution of resulting curves, $f(t|\theta_i)$. Ultimately, it will be from a distribution of curves that we determine the temporal characteristics of each group, with the differences in these temporal characteristics being what we are interested in identifying. With the general notation addressed, we now move to the particulars of each of the methods considered.

2.1   Homogeneous Bootstrap

The original bootstrapping algorithm presented in **?** follows what we will call the *homogeneous means* assumption. Accordingly, we will call this bootstrapping algorithm the *homogeneous bootstrap*. Under the homogeneous means assumption, it is still assumed that observed data for each subject retains the mean structure given in Equation 1, but with the additional assumption that $\theta_i = \theta_j$ for all subjects i, j within the same group. In other words, there is assumed to be no variability in the mean structure between subjects within the same group. This is evidenced in the original bootstrapping differences in time series algorithm, which samples without replacement at each bootstrapping step:

1. For each subject, fit a nonlinear regression model to obtain $\hat{\theta}_i$. **?** recommends specifying an AR(1) autocorrelation structure for model errors. Assuming large sample normality, the sampling distribution of each estimator can be approximated by a multivariate normal distribution with mean for subject i, $\hat{\theta}_i$ corresponding to the point estimate and standard deviations corresponding to the standard errors, s_i.

2. Using the approximate sampling distributions in (1.), randomly draw one bootstrap estimate for each of the model parameters on every subject

$$\hat{\theta}_i^{(b)} \sim N(\hat{\theta}_i, s_i^2) \tag{3}$$

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 n_g individuals for the bth bootstrap in group g,

$$\theta_g^{(b)} = \frac{1}{n_g} \sum_{i=1}^{n_g} \hat{\theta}_i^{(b)} \tag{4}$$

4. Use the mean parameter estimates to determine a bootstrapped population level curve, which provides the average population response at each time point, $f(t|\theta_g^{(b)})$.

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. For each group $g = 1, \ldots, G$, this gives

$$\bar{p}_{gt} = \frac{1}{B} \sum_{b=1}^{B} f(t|\theta_g^{(b)}), \qquad s_{gt}^2 = \frac{1}{B-1} \sum_{b=1}^{B} f(t|\theta_g^{(b)}) - \bar{p}_{gt}, \tag{5}$$

where $\bar{p}_{gt}$ and s_{gt}^2 are mean and standard deviation estimates at each time point for group g.

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 ? to test for significance. As this correction is also used for the heterogeneous bootstrap presented next, we offer a more comprehensive review of this adjustment at the end of this section.

2.2 Heterogeneous Bootstrap

Typically, subjects within a group demonstrate considerable variability in their mean parameter estimates. In this case, we should avoid the presumption that $\theta_i = \theta_j$ as accounting for between-subject variability within a group will be critical for obtaining a reasonable distribution of the population curves. More likely, we may assume that the distribution of parameters for subjects $i = 1, \ldots, n_g$ in group $g = 1, \ldots, G$ follows the distribution

$$\theta_i \sim N(\mu_g, V_g), \tag{6}$$

where μ_g and V_g are the group-specific mean and variance values, respectively. Similar to what was presented in the homogeneous bootstrap algorithm, we can further account for uncertainty in our estimation of θ_i by $\hat{\theta}_i$ by treating the standard errors derived when fitting the observed data to the mean structure suggested in Equation 1 as estimates of their standard deviations. This gives us a multivariate normal distribution for each subject's estimated parameter,

$$\hat{\theta}_i \sim N(\theta_i, s_i^2). \tag{7}$$

As our goal remains as being able to obtain reasonable estimates of the population curves for each group, 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 $b = 1, \ldots, B$, where

$$\theta_{ib}^* \sim N(\hat{\theta}_i, s_i^2), \tag{8}$$

as was done in Step (2.) of the homogeneous bootstrapping algorithm. 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

$$\theta_{bg}^{(hom)} = \frac{1}{n_g} \sum_{i=1}^{n_g} \theta_{ib}^*, \quad \theta_{bg}^{(hom)} \sim N\left(\mu_g, \frac{1}{n_g^2} \sum_{i=1}^{n_g} s_i^2\right). \tag{9}$$

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 8, but the heterogeneous bootstrapped group mean now follows

$$\theta_{bg}^{(het)} \sim N\left(\mu_g, \frac{1}{n_g} V_g + \frac{1}{n_g^2} \sum s_i^2\right). \tag{10}$$

in the homogenous bootstrap, we sampled without replacement and this was indicated by the countable indicies on the summation. Should I include those here? Can I do so without introducing new indices?

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: I'm not sure what additional math to include here

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

$$T_t^{(b)} = \frac{(\bar{p}_{1t} - \bar{p}_{2t})}{\sqrt{s_{1t}^2 + s_{2t}^2}}, \tag{11}$$

where $\bar{p}_{gt}$ and s_{gt}^2 are mean and standard deviation estimates at each time point for groups 1 and 2, respectively. Finally, just as in **?**, one can use the autocorrelation of the $T_t^{(b)}$ statistics to create a modified α for controlling the FWER.

2.3 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:

not sure what additional math to add 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 $\bar{p}_{gt}$ and s_{gt}^2 for $g = 1, 2$. Use these values to compute a permutation test statistic $T_t^{(p)}$ at each time point,

$$T_t^{(p)} = \frac{|\bar{p}_{1t} - \bar{p}_{2t}|}{\sqrt{s_{1t}^2 + s_{2t}^2}}. \tag{12}$$

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.

2.4 Paired Data

Briefly, we attend to the issue of paired data for each of the discussed methods, as this is critical for a proper assessment of both FWER and power. Specifically, in a paired setting we note that our interest is in

determining the distribution of paired differences rather than of the respective groups.

For the homogeneous bootstrap, this is done by default: as each subject is sampled without replacement, we can be sure that each bootstrap estimate between groups contains the same subjects. For the heterogeneous bootstrap, this is done by ensuring that the same subjects sampled in each bootstrap for one group is matched identically with sampling subjects in the other. Put differently, this states that we begin by determining which subjects will be included in each bootstrap *for both groups at the same time*. Lastly, in the case of permutation testing, paired data is addressed by ensuring that each permuted group contains one observation for each subjects, so that each subject in one permuted group has its corresponding observation in the other. am i saying this weird?

2.5   Modified Bonferonni Correction

While the permutation method determines significance via comparison to an estimated null distribution, both the homogeneous and heterogeneous bootstrap construct estimates of the observed distribution against which the null hypothesis is tested. This results in a series of often highly correlated test statistics, raising concerns that are typically associated with multiple testing.

The method for controlling FWER presented in **?** begins on the assumption that adjacent test statistics in a densely sampled time series will be highly correlated. That is, it will be assumed that the test statistics have null standard normal distribution $T_t \sim N(0, 1)$, but that the sequence of statistics themselves unfolds with an AR(1) process

$$T_t = \rho T_{t-1} + \epsilon_t, \tag{13}$$

where $\epsilon_t \sim N(0, (1 - \rho^2))$. This gives the conditional distribution

$$T_t | T_{t-1} \sim N(\rho T_{t-1}, 1 - \rho^2), \tag{14}$$

with the joint distribution of T_t and T_{t-1} being bivariate normal with mean 0, a variance of 1, and correlation ρ. From this, they show that for some α^*, the true FWER α under the null hypothesis can be expressed as

$$1 - P\left(\bigcap_{t=1}^{T} I_t\right) = 1 - P(I_1) \prod_{t=2}^{T} P(I_t | I_{t-1}) = 1 - P(I_1) P(I_t | I_{t-1})^{T-1}, \tag{15}$$

where I_t is the event that $|T_t| \leq z_{\left(1 - \frac{\alpha^*}{2}\right)}$. The correction is made by finding the nominal significance α^* that produces the desired FWER of α. Of note here, no adjustment to the nominal α is needed when the tests are perfectly correlated. Conversely, when the tests are perfectly independent, Equation 15 reduces to the standard Bonferroni correction. This correction is implemented in `bdots` and is also provided outside of

the bootstrapping process with the function `bdots::p_adjust` using `method = "oleson"`.

3   FWER Simulations

We now go about comparing the family-wise 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 generating data with both a homogeneous and heterogeneous means assumption, generating data with and without autocorrelated errors, and fitting data with and without an AR(1) assumption. In addressing performance under paired and unpaired conditions, we have included two distinct but reasonable instances in which the data may be paired, which we will elaborate further on shortly. However, as paired data becomes difficult to define under the homogeneous means assumption (in which all subjects are, in a sense, "paired"), we will omit these settings from our final simulations. In all, this gives us sixteen different arrangements which we will examine for their family-wise error rates using each of the three methods previously described.

3.1   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:

$$f(t|\theta) = \frac{p - b}{1 + \exp\left(\frac{4s}{p-b}(x - t)\right)} + b \tag{16}$$

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 multivariate normal distribution, with subject $i = 1, \ldots, N$ in group $g = 1, \ldots, G$ following the distribution in Equation 6. These parameters are then used to simulate empirical data according the the mean and error structures for each simulation setting, which we detail next.

Mean Structure   In all of the simulations presented, the distribution of parameters used in Equation 6 was empirically determined from data on normal hearing subjects in the VWP (**?**). Parameters used were those fit to fixations on the Target, following the functional form of Equation 16. We will call this the heterogeneous means assumption, as each subject will draw their own unique subject-specific parameters from this distribution. A visual depiction of the distribution of these curves is given in Figure 1(a).

Under the assumption of between-subject homogeneity, which we call the homogeneous means assump-
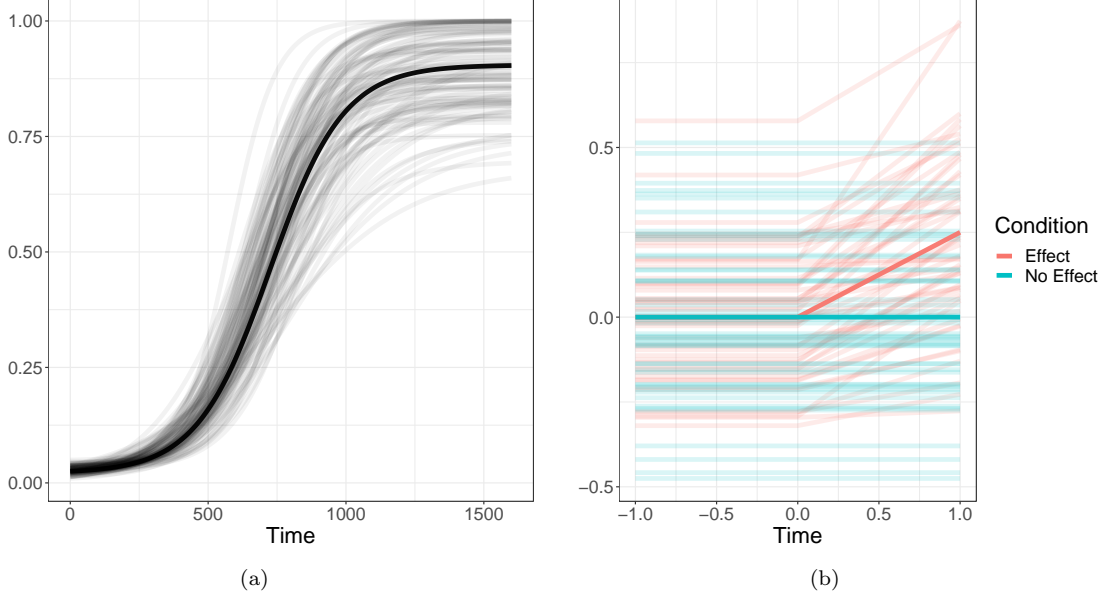
Figure 1: Distributions of various group with the mean curve in bold (a.) 50 samples from the generating distribution of the four-parameter logistic in Equation 16 used for testing FWER. (b.) 50 samples from the generating distributions of each group in Equation 19

tion, 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 structure.

Error Structure   The error structure is of the form

$$e_{it} = \phi e_{i,t-1} + w_{it}, \quad w_{it} \sim N(0, \sigma) \tag{17}$$

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   As we previously noted, paired data is only a sensible condition under the assumption of heterogeneous means, and we limit our consideration to that case. Further, we observe that in the construction of paired data, there are two methods that seem reasonable, and we employ both of them here.

In considering the construction of paired data for subject i, the first method proceeds as follows: we begin by drawing parameters θ_i from Equation 6. Denote this θ_{i1} to indicate that this is the parameter estimate for subject i in group 1. We then simulate observed data according to Equation 1. To create the paired data, we then set $\theta_{i2} = \theta_{i1}$ and again simulate observed data according to Equation 1. Under this first method, the generating parameters between groups are *identical*, with the only differences between the

simulated data being that contributed by the error term.

In the second method of obtaining paired data, we proceed as in the first, except now letting

$$\theta_{i2} = \theta_{i1} + N(0, 0.05 \cdot V_g). \tag{18}$$

This adds a small amount of random noise between paired parameters, simulating the degree of variability that may normally be found between conditions, even when there is no true effect. Accommodating this phenomenon is relevant in situations in which the data gathering mechanism has imperfect reliability, as in the case of the VWP. It is also relevant because, as we will show, the homogeneous bootstrap is highly sensitive to these assumptions, with potentially disastrous results when these conditions are not precisely met. To avoid potential confusion, the results for each of these will be presented separately.

Each set of conditions generates two groups, with $n = 25$ subjects in each group, with time points $t = 0, 4, 8, \ldots, 1600$ 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 help assess the impact of correctly or incorrectly identifying the type of error when conducting an analysis in `bdots`. Finally, results will be separated by paired status. Each simulation was conducted 1000 times, with the proportion of simulations in which a significance difference was incorrectly identified used to determine the family-wise error rate.

3.2 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. The results for the unpaired data are presented in Table 1, while those for each of the paired variations are presented in Tables 2 and 3.

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 instances in which that time was incorrectly determined to have a significant difference. 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

10

rates indicate that large swaths of a time series frequently sustain type I errors. The median per-comparison rates for unpaired simulations are presented in Table 4 and each of the paired simulations in Table 5 and Table 6.

3.2.1 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 FWER 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 continues to be true under the homogenous mean assumption when the true error structure is not autocorrelated.

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.

Het. Means	AR(1) Error	AR(1) Specified	Hom. Boot	Het. Boot	Perm.
No	Yes	Yes	0.09	0.00	0.06
No	Yes	No	0.84	0.06	0.14
No	No	Yes	0.12	0.01	0.08
No	No	No	0.14	0.00	0.05
Yes	Yes	Yes	0.94	0.05	0.05
Yes	Yes	No	0.99	0.07	0.07
Yes	No	Yes	1.00	0.08	0.05
Yes	No	No	0.99	0.05	0.04

Table 1: FWER for logistic function (unpaired)

We turn our attention now to simulations with paired data. Table 2 gives the results under the construction assuming *identical* parameters between groups. Note that despite the simulation having a heterogeneous mean structure, the FWER associated with the homogeneous bootstrap is comparable to that of the permutation test (though both with elevated FWER rates). Because these data are paired, the observed variability within each group has no effect on the distribution of paired differences. This is in stark contrast to the

results presented in Table 3, where a small amount of variability was added to the paired set of parameters. What is most relevant here is that both the heterogeneous bootstrap as well as the permutation test are robust to these small differences in the paired setting, while the results associated with the homogeneous bootstrap are radically different. This serves to demonstrate how sensitive the homogeneous bootstrap is to such a rigid set of underlying assumptions.

Het. Means	AR(1) Error	AR(1) Specified	Hom. Boot	Het. Boot	Perm.
Yes	Yes	Yes	0.10	0.00	0.12
Yes	Yes	No	0.75	0.06	0.12
Yes	No	Yes	0.12	0.00	0.11
Yes	No	No	0.11	0.01	0.13

Table 2: FWER for logistic function (paired, identical parameters)

Het. Means	AR(1) Error	AR(1) Specified	Hom. Boot	Het. Boot	Perm.
Yes	Yes	Yes	0.48	0.04	0.10
Yes	Yes	No	0.93	0.07	0.12
Yes	No	Yes	0.81	0.04	0.08
Yes	No	No	0.81	0.07	0.09

Table 3: FWER for logistic function (paired, added noise)

3.2.2 Median per-comparison error rate

We next consider the median per-comparison error rate, which offers some insight into the FWER. In particular, consider the situation in which in Table 4, 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 4; combined with a FWER near the nominal 0.05, that these values are not identically 0 suggests that errors are likely distributed across the entire range rather than limited to a small area.

Het. Means	AR(1) Error	AR(1) Specified	Hom. Boot	Het. Boot	Perm.
No	Yes	Yes	0.02	0.00	0.01
No	Yes	No	0.31	0.01	0.02
No	No	Yes	0.01	0.00	0.01
No	No	No	0.00	0.00	0.01
Yes	Yes	Yes	0.59	0.01	0.01
Yes	Yes	No	0.83	0.02	0.01
Yes	No	Yes	0.84	0.02	0.01
Yes	No	No	0.82	0.01	0.01

Table 4: Median per-comparison error rate for unpaired data

Regarding paired data, we see again a similar situation play out as we did with the FWER. Table 5 treats the paired setting with identical parameters, and we see median per-comparison error rates consistent with the lower FWER from Table 2. And again, Table 6 demonstrates a more robust performance for both the heterogeneous bootstrap and permutation test, relative to that of the homogeneous bootstrap.

Het. Means	AR(1) Error	AR(1) Specified	Hom. Boot	Het. Boot	Perm.
Yes	Yes	Yes	0.03	0.00	0.02
Yes	Yes	No	0.27	0.01	0.02
Yes	No	Yes	0.01	0.00	0.02
Yes	No	No	0.01	0.00	0.02

Table 5: Median per-comparison error rate for paired data (identical parameters)

Het. Means	AR(1) Error	AR(1) Specified	Hom. Boot	Het. Boot	Perm.
Yes	Yes	Yes	0.14	0.01	0.02
Yes	Yes	No	0.46	0.01	0.03
Yes	No	Yes	0.44	0.01	0.01
Yes	No	No	0.41	0.02	0.02

Table 6: Median per-comparison error rate for paired data (added noise)

3.3 Discussion

It was the control of the family-wise error rate in the context of densely sampled, highly correlated tests that first motivated the development of `bdots`. And while the results presented in this section indeed demonstrate control of the FWER in under a strict set of assumptions, they also highlight the consequences when these conditions are not met. With the introduction of both the heterogeneous bootstrap and the permutation test, we have offered two alternatives that are robust to a wide variety of situations while maintaining performance similar to that of the homogeneous bootstrap in the best of cases. This is but one half of the question, however. In the following section, we present two separate simulations to determine if the robustness acquired is as the cost of significant power.

4 Power Simulations

In our assessment of power among the presented methods, we examine two distinct simulations, each associated with a different aspect of the question of identifying temporal differences. The first of these seeks to identify the relationship between power and effect size, which is done by comparing two simple piecewise linear functions as we will describe shortly. The goals of the second simulation, which is done in the context of VWP data, are twofold: first, to assess and verify the power of the new methods in paired and unpaired settings, and second, to investigate the relationship between power and effect size in light of between-subject variability.

4.1 Piecewise Linear Power

To better understand the relationship between effect size and power, we have created a simple example consisting of two simulated experimental groups whose mean structure is a simple piecewise linear function defined on the interval $(-1, 1)$ as follows:

$$y = \begin{cases} b & x < 0 \\ mx + b & x \geq 0 \end{cases} \tag{19}$$

The set of parameters drawn from each subject include a baseline parameter, b, as well as a slope parameter, m, which were each drawn from a univariate normal distribution. To distinguish experimental groups, one designated as being the "Effect" group, other other being "No Effect", we set the slope parameter of the "No Effect" group to be identically $m = 0$, while the "Effect" group had a slope parameter that was normally distributed with a mean value $\mu_m = 0.25$. Both groups drew their baseline parameter, b, from a

normal distribution with mean $\mu_b = 0$. As a consequence of this, and following from Equation 19, there should be no difference between groups when $x < 0$, with an effect size for the "Effect" group being mx for all $x \geq 0$. This was done for two reasons. First, by intentionally including an interval in which there is no difference, we are able to observe power in the context of type I error as well. That is, achieving a higher power when $x \geq 0$ will be of little value if differences are equally (and incorrectly) identified when $x < 0$. The second reason for having set our problem up this way was to give us the ability to observe changes in power in conjunction with changes in effect size, mediated by increasing values in x. A visual depiction of each of these groups is given in Figure 1(b). maybe i separate those images and bring that down here. Perhaps I should make a similiar distribution plot for differences in logistic shift?

For these simulations, we limited consideration to three possible scenarios: first, we assumed the conditions presented in ?, 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 last two 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 cases in which their is autocorrelated error that is not accounted for. Simulation results for a number of other settings are handled in the appendix.

For each subject, parameters for their mean structure given in Equation 19 were drawn according to their group membership and fit using `bdots` on the interval (-1,1). Time windows in which the groups differed were identified using each the homogenous bootstrap, heterogeneous bootstrap, and permutation testing. By including the interval (-1,0) in which the null hypothesis was true, we are able to mitigate the effects of over-zealous methods in determining power, and we present the results in the following way: any tests in which a difference was detected in (-1,0) was marked as having a type I error, and the proportion of simulations in which this occurred for each method is reported as the FWER in the column labeled α. The next column, β, is the type II error rate, indicating the proportion of trials in which no differences were identified over the entire region. The last Greek-letter column is $1 - \beta - \alpha$, a modified power statistic indicating the proportion of tests in which a difference was correctly identified. The remaining columns relate to this modified power column, giving a partial summary of the earliest onset of detection. As a true difference occurs on the interval $t > 0$, smaller values indicate greater power in detecting differences. Finally, a plot giving the power at each time point is given in Figure 2. This plot represents the true power, though note that it does not take into account the rate at which these regions were identified in conjunction with a

type I error rate.

4.1.1 Results

The results of the power simulation are presented in Table 7. We begin by considering the case in which we assumed a homogeneous mean structure with autocorrelated errors, matching the conditions in which the homogeneous bootstrap was first presented. Notably, we find that the permutation method demonstrates the greater power, with the median onset time just under that of the homogeneous bootstrap. This is at the expense of a larger FWER, though still below the nominal level. Alternatively, the heterogeneous bootstrap maintains a similar FWER as the homogeneous bootstrap at the cost of power. The remaining settings tell a similar story with the exception of the homogeneous bootstrap which continues to demonstrate unacceptable FWER under the heterogeneous means assumption. As to the effect of incorrectly specifying an AR(1) error structure when comparing the last two settings, note that there tends to be very little effect between the heterogeneous bootstrap and permutation, with neither performing consistently better or worse at any of the quantiles given.

Method	Heterogeneity	AR(1)	α	β	1 - α - β	1st Qu.	Median	3rd Qu.
Hom. Boot	No	Yes	0.00	0.00	1.00	0.025	0.030	0.035
Het. Boot	No	Yes	0.00	0.00	1.00	0.035	0.040	0.045
Perm	No	Yes	0.03	0.00	0.97	0.020	0.025	0.030
Hom. Boot	Yes	No	0.95	0.00	0.05	0.005	0.008	0.010
Het. Boot	Yes	No	0.00	0.01	0.98	0.260	0.330	0.480
Perm	Yes	No	0.04	0.00	0.95	0.245	0.325	0.452
Hom. Boot	Yes	Yes	0.94	0.00	0.06	0.005	0.013	0.015
Het. Boot	Yes	Yes	0.01	0.01	0.98	0.270	0.370	0.465
Perm	Yes	Yes	0.04	0.00	0.96	0.245	0.365	0.440

Table 7: caption

The results in Table 8 are a summary of all of the methods found by taking the mean of each of the results presented. This is intended to interrogate the performance of each of these methods when underlying assumptions are unknown or unspecified. We find a robust performance for each of the new methods presented, maintaining a reasonable relationship between FWER and power. The metrics associated with homogeneous bootstrap are perhaps a bit misleading here as they appear to demonstrate exceptional power, though at the cost of unacceptable type I error.

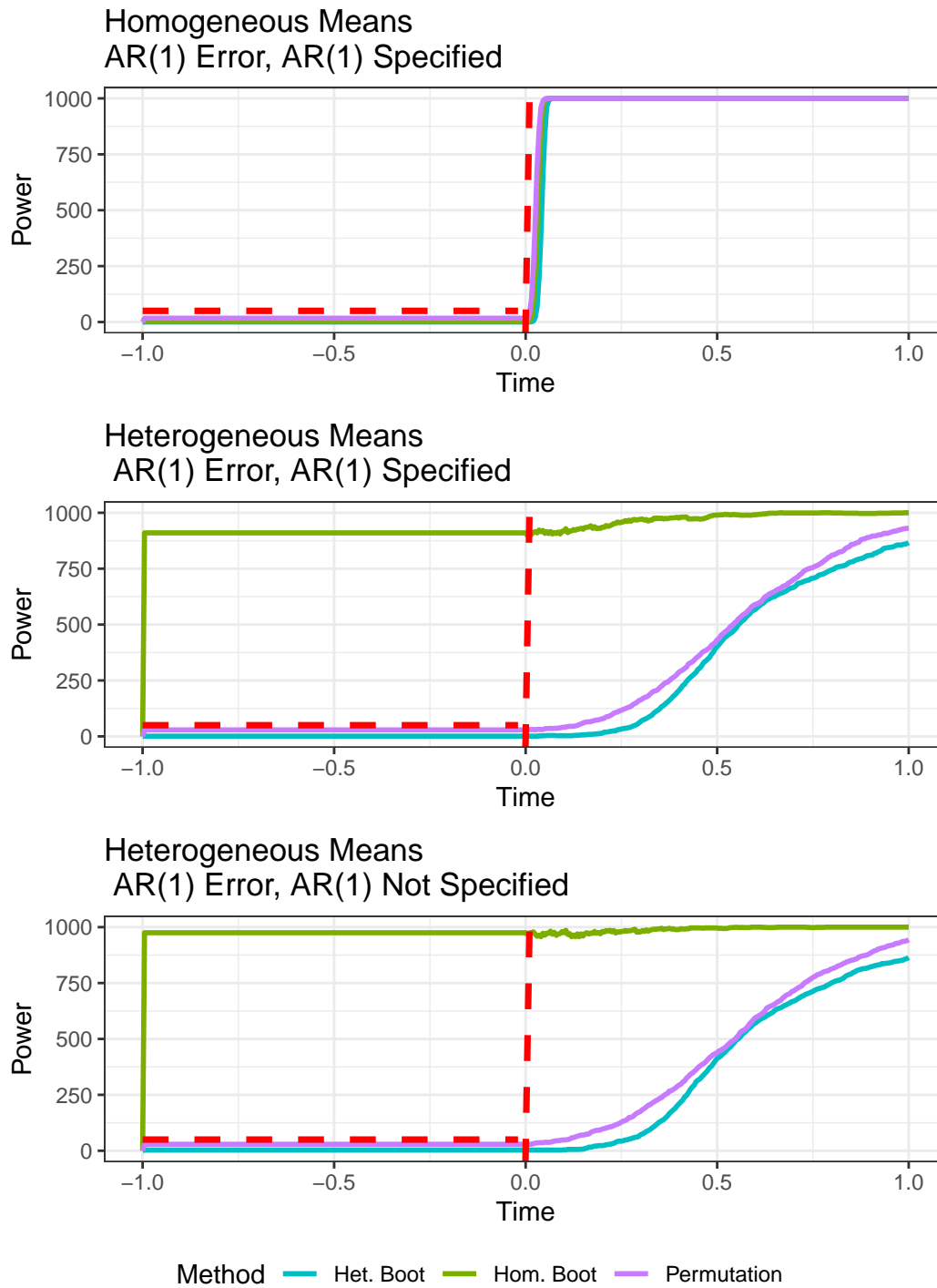| Method	α	β	$1 - \alpha - \beta$	1st Qu.	Median	3rd Qu.
Hom. Bootstrap	0.772	0.000	0.228	0.011	0.016	0.021
Het. Bootstrap	0.002	0.097	0.901	0.328	0.419	0.546
Permtuation	0.027	0.038	0.935	0.294	0.424	0.556

Table 8: Summary of methods for Type II error

Figure 2: Observed power of each of the methods at each time in (-1,1)

4.2   Logistic Shift Power

this section seems a little short, but it really is just a quick-and-dirty look, mostly confirming questions had by the committee. Notably, I don't have a defined "results" section, but thats probably ok? Because what else is there really to say besides look at what we did

The final set of simulations we consider seek to address concerns related to power in a paired setting, as well as to investigate the relationship between effect size and variability. We proceed with simulation settings similar to those in our investigation of the type I error in that each of the two groups under consideration draw from an empirical distribution of parameters associated with the four parameter logistic function. Unlike the situation in determining the FWER, here differences are introduced between groups by changing the crossover parameter, indicating the inflection point of the function. To simulate various effect sizes, we investigate changes in crossover of 50 and 150. Further, to determine the effect of between subject variability on the identification of differences, we have set the standard deviation of the crossover parameter from the empirical distribution to take values of either 60 or 120. Thus, the settings we will be presenting include small and large changes in light of small or large variability. Additionally, data was generated with iid errors and fit without specifying AR(1) correlation in `bdots`. Finally, each of these settings will be run as paired and unpaired data, with parameters for the paired data being identical between groups (i.e., no added variability).

Consider first, for example, the results presented in Figure 3, each of which demonstrate the observed power at each time point when the crossover parameter between group was shifted by 50. Predictably, we see greater power in the unpaired setting when the standard deviation of the crossover parameter is smaller, relative to the shift. By contrast, in the paired setting we little difference in power as a function of parameter variability. This is also to be expected as here we are concerned with the variability of the *difference* rather than of each of the groups, and these results confirm that this is indeed the case.

A similar set of conclusions can be seen by considering the observed power presented in Figure 4, where here instead we the results when the crossover parameter is shifted by 150. Similar to the prior case, we see greater power associated with the smaller standard deviation in the unpaired case, with power in the paired case performing similarly. And finally, by comparing the results between Figures 3 and 4 we confirm in all cases that a greater effect size is associated with greater power.

5   Discussion

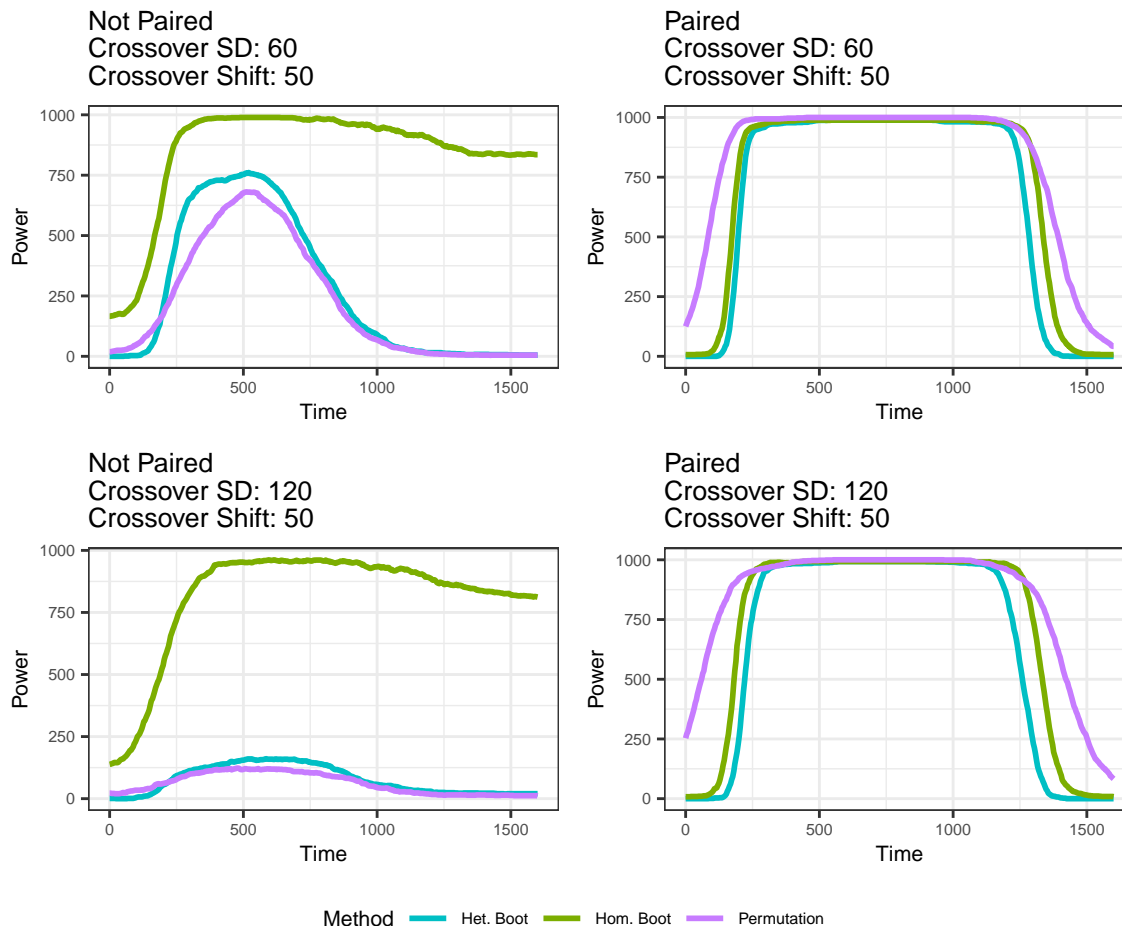i am maybe not quite happy with this conclusion section, will continue to chew on it a little bit

Figure 3: not sure that I love the titles of these plots Observed power following a small shift in the crossover parameter

We set out both to interrogate the validity of the homogeneous 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 heterogeneous bootstrap and permutation tests while also highlighting a major shortcoming of the original. 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.

In light of the results presented, one issue of concern is addressing the fact that a version of `bdots` with the homogeneous mean assumption was presented in 2018 and remained accessible on CRAN until the end of 2022. This has implications for the number of papers in which `bdots` may have demonstrated significance between groups when the underlying assumptions of homogeneous mean structure did not hold, as is likely the case in all instances related to the VWP. Concurrent with this issue is the issue of identifying current users of `bdots` of this change, as results found only a month ago will be profoundly different than what
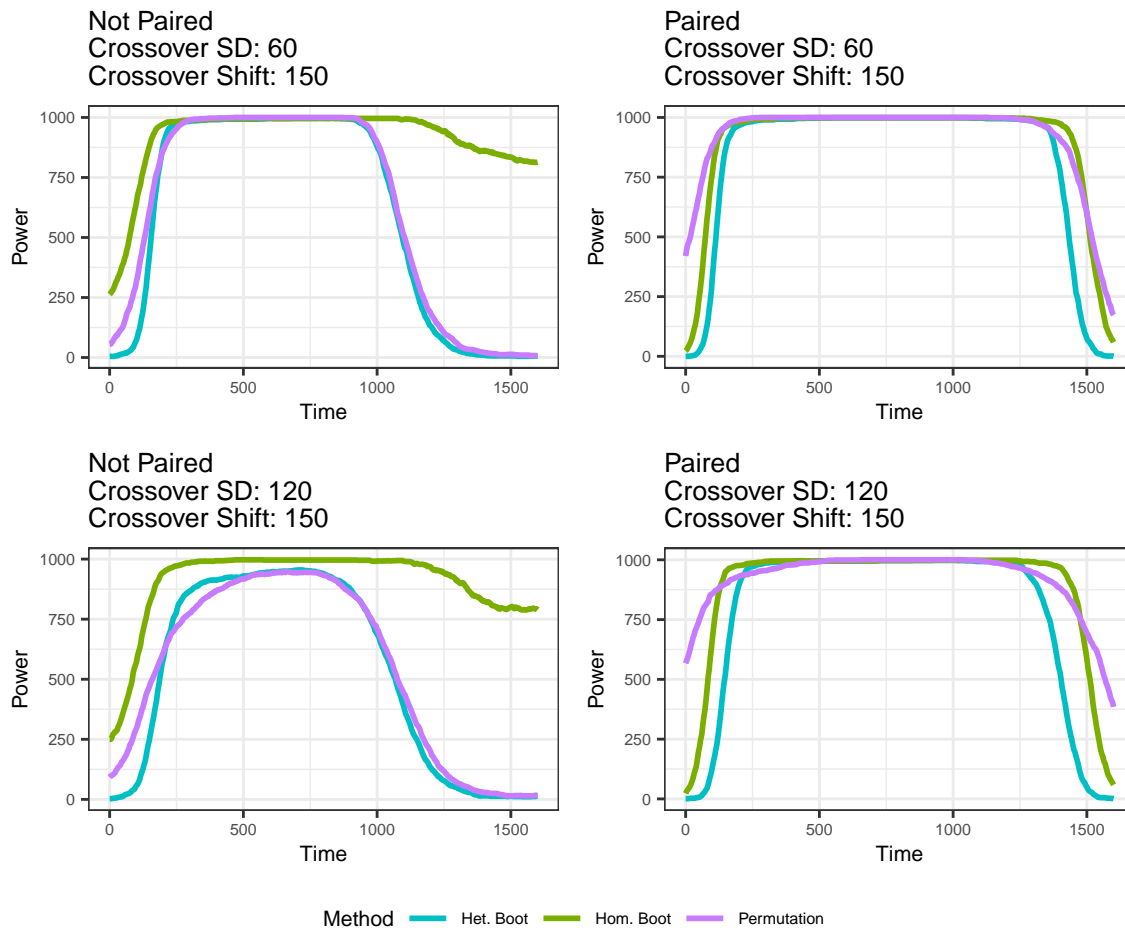
Figure 4: Observed power following a large shift in the crossover parameter

is seen today. At present, I am not sure the best way to address either of these. In either case, however, it will be prudent to remove this option from the `bdots` package all together, as there appears to be no obvious advantage to the homogeneous bootstrap over the others in terms of either controlling the FWER or obtaining power, even when the homogeneous mean structure assumption is met.

There are several limitations of the current paper that are worthy of further investigation. First, 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 open for exploration. It is also worth investigating in greater detail what impact the re-drawing of subject specific parameters from their respective distributions has on both the FWER and power, as in several of

the simulations the observed FWER was much lower than the nominal level. Particularly in the case of the permutation method which is *not* seeking to estimate the group distributions, it may be worthwhile to see if a favorable trade can be made to increase the resulting power.

We conclude by noting that `bdots` is now equipped with two methods to effectively control the FWER when assessing the differences in time series under a greater set of underlying assumptions, including those involving the presence of highly correlated test statistics. Further, both methods presented are robust to misspecification of the error structure while maintaining an acceptable FWER and adequate power.

Appendix

6   Full Piecewise Power Simulations

Here we present the full collection of the linear piecewise power simulations, which, in addition to those given in Table 7 includes cases for heterogeneous means where autocorrelation is specified when fitting with `bdots`. This is indicated in the "AR(1) Specified" column. Notably from this table, we see that in the case in which the true errors are iid, there is no measurable effect on power when an autocorrelated structure is incorrectly specified. This is similar to the opposite situation, in which the true error does have an AR(1) structure. In this case, we observe a marginal benefit to correctly specifying an AR(1) structure. This may in fact make retaining the AR(1) assumption a reasonable default in the `bdots` package.
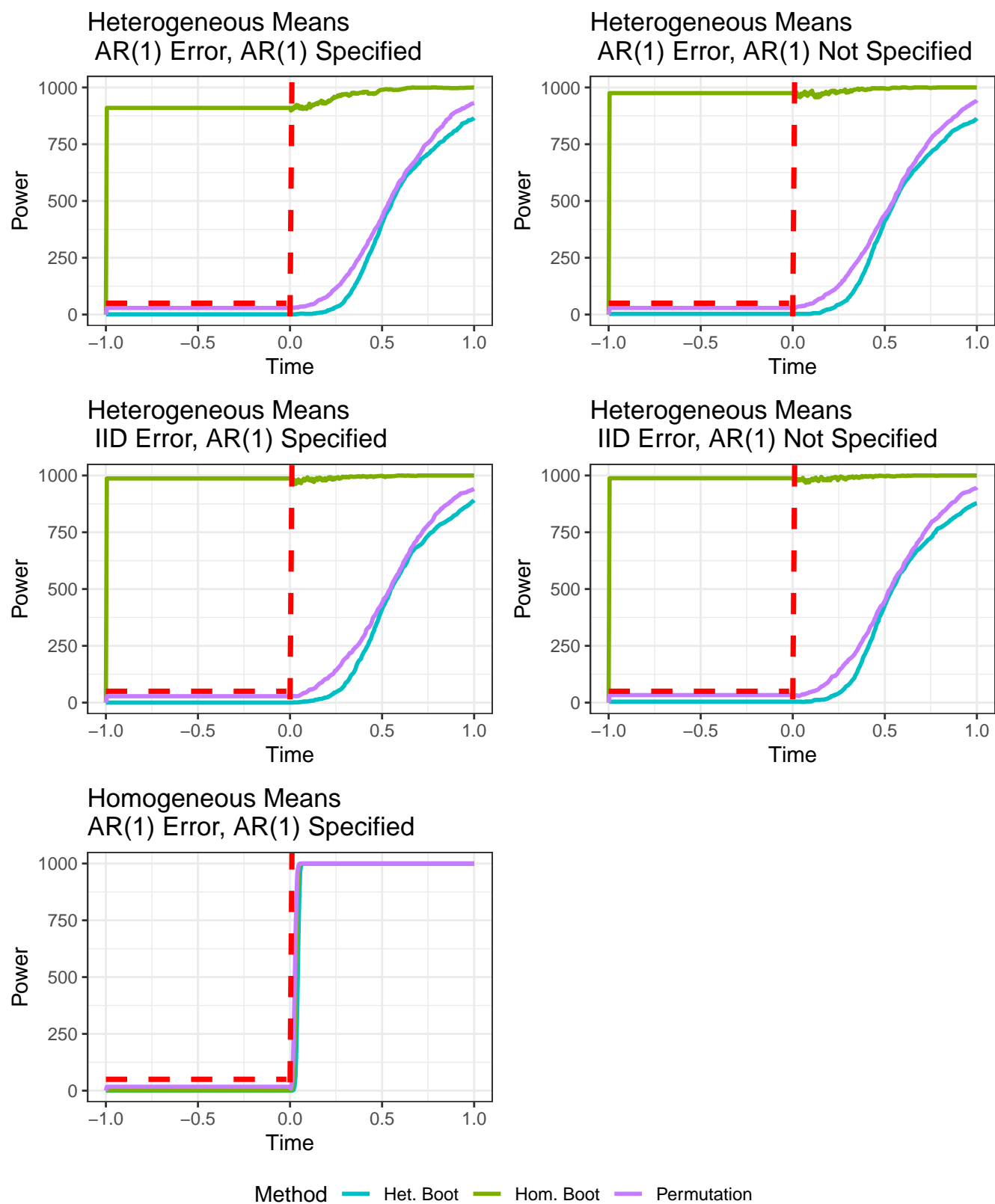
Figure 5: Power plots in time for each of the simulation settings. Note that in the heterogeneous means case, there is little difference when AR(1) is incorrectly specified

Method	Heterogeneity	AR(1) Error	AR(1) Specified	α	β	$1 - \alpha - \beta$	1st Qu.	Median	3rd Qu.
Hom. Boot	No	Yes	Yes	0.00	0.00	1.00	0.025	0.030	0.035
Het. Boot	No	Yes	Yes	0.00	0.00	1.00	0.035	0.040	0.045
Perm	No	Yes	Yes	0.03	0.00	0.97	0.020	0.025	0.030
Hom. Boot	Yes	No	No	0.95	0.00	0.05	0.005	0.008	0.010
Het. Boot	Yes	No	No	0.00	0.01	0.98	0.260	0.330	0.480
Perm	Yes	No	No	0.04	0.00	0.95	0.245	0.325	0.452
Hom. Boot	Yes	No	Yes	0.98	0.00	0.02	0.005	0.008	0.010
Het. Boot	Yes	No	Yes	0.00	0.01	0.99	0.261	0.350	0.475
Perm	Yes	No	Yes	0.04	0.00	0.96	0.225	0.335	0.440
Hom. Boot	Yes	Yes	No	0.94	0.00	0.06	0.005	0.013	0.015
Het. Boot	Yes	Yes	No	0.01	0.01	0.98	0.270	0.370	0.465
Perm	Yes	Yes	No	0.04	0.00	0.96	0.245	0.365	0.440
Hom. Boot	Yes	Yes	Yes	0.83	0.00	0.17	0.021	0.032	0.040
Het. Boot	Yes	Yes	Yes	0.00	0.01	0.98	0.250	0.330	0.450
Perm	Yes	Yes	Yes	0.03	0.00	0.97	0.223	0.335	0.428

Table 9: Power for full piecewise linear simulation