
Article:
Embleton, Jonathan, Knight, Marina Iuliana orcid.org/0000-0001-9926-6092 and Ombao, Hernando (2022) Wavelet testing for a replicate-effect within an ordered multiple-trial experiment. Computational Statistics & Data Analysis. 107456. ISSN 0167-9473

https://doi.org/10.1016/j.csda.2022.107456

Wavelet testing for a replicate-effect within an ordered multiple-trial experiment

Jonathan Embleton[a], Marina I. Knight[a,*], Hernando Ombao[b]

[a]*Department of Mathematics, University of York, UK*
[b]*Statistics Program, King Abdullah University of Science and Technology (KAUST), Saudi Arabia*

Abstract

Experimental time series data collected across a sequence of ordered trials (replicates) often crop up in many fields, from neuroscience to circadian biology. In order to decide when to appropriately evade the simplifying assumption that all replicates stem from the same process, an assumption often untrue even when identical stimuli are applied, two novel tests are proposed that assess whether a significant trial-effect is manifest along the experiment. The modelling framework uses wavelet multiscale constructions that mitigate against the potential nonstationarities often present in experimental data, both across times and across replicates. The proposed tests are evaluated in thorough simulation studies and illustrated on neuroscience data, proving to be flexible tools with great promise in dealing with complex multiple-trials time series data and allowing the analyst to accordingly tune their subsequent analysis.

Keywords: nonstationarity; replicate time series; bootstrap; neuroscience

1. Introduction

Experimental settings that consist of time series measurements obtained during ordered trials (or, interchangeably, replicates) crop up in many scientific fields, such as neuroscience or biology. To aid inference, analysts often naively choose to examine the average process dynamics across all trials, thus arriving at conclusions that completely ignore the timeline of the experiment. Within the neurosciences, Gorrostieta *et al.* (2012); Fiecas and Ombao (2016); Embleton *et al.* (2020) suggest that this is a fraught approach that fails to account for the existence of a meta-process evolution that may exist even for identical stimuli, over the course of the experiment. For clarity, throughout this paper we use the terminology 'meta-process' to refer to the process obtained over the course of the experiment, hence not only over the time within a trial, but also across the *ordered* trials. The data example which we will address in our work here focusses on the hippocampus (Hc) and the nucleus accumbens (NAc), associated with memory recall and the processing of reward, respectively and therefore a key component in cognitive processing. Recordings of electrical activity (at approximately 1000Hz) using local field potentials (LFPs) were obtained from the Hc and NAc of an awake behaving macaque during an associative learning experiment. For each trial, the macaque was presented with one of four pictures and was then tasked with associating this picture with one of the four doors appearing on the screen. Upon making a correct association, the macaque was rewarded with a small quantity of juice. Further experimental details are given in Appendix A. Plots of the LFPs obtained from replicates in which a correct association was made are shown in Figures 1 and 2, where the experimental timeline was ensured to match from trial to trial. Experimental learning studies have acknowledged evolutionary neuronal activity across (both) brain regions (Seger and Cincotta, 2006; Abela *et al.*, 2015; Granados-Garcia *et al.*, 2021) amounting to traces that display a *nonstationary* behaviour across both time and trials.

In this context, we aim to provide an answer to a practically important question: is there a meta-process evolution along the ordered trials (or replicates) of the experiment? The ability to answer this question will enable the analyst to appropriately choose a sensible 'tool' from the analysis toolbox and to subsequently strengthen their inference.

*Corresponding author's address: M. Knight, Department of Mathematics, University of York, YO10 5DD, UK tel.: +441904324166, email: marina.knight@york.ac.uk

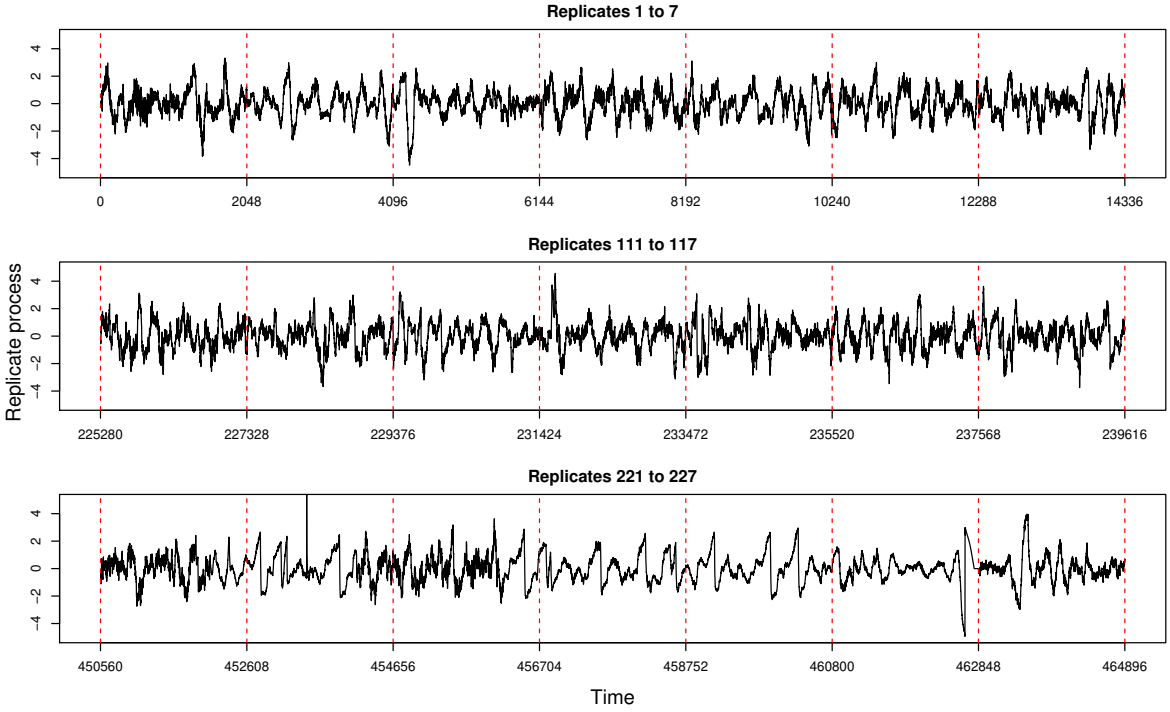


Figure 1: Concatenated series of the hippocampus (Hc) data in the correct response trials (replicates). (Vertical lines denote breaks between trials. Concatenation is only used for meta-process visualisation.)

Time-scale decompositions that are typical of wavelet constructions are appealing for dealing with nonstationarities present within time series, and are already an established strategy for the analysis of brain signals (Sanderson *et al.*, 2010; Park *et al.*, 2014; Embleton *et al.*, 2020). Our aim is to develop a *wavelet*-based testing tool that will assess the existence of a *replicate-effect* along the course of a multiple-trial experiment. In order to work within a setup under which the notion of replicate-effect can be mathematically quantified, we propose to use the novel methodology of Embleton *et al.* (2020) that directly gives a stochastic representation to the meta-process of ordered trials using (i) the experimental timeline across replicates, as well as (ii) the time within each trial. This framework allows for process evolution as quantified by its associated spectral structure viewed as a function of both (rescaled) trial and within-trial time. (The precise mathematical meaning of 'replicate-effect' will be defined later.)

Thus, here we propose two tests that establish and identify the existence of a replicate-effect in a meta-process collected over ordered trials: a *location-specific* (time and trial-specific) test and a *global* test.

The idea of testing for the presence of time-varying dynamics of a process within the spectral domain has existed for some time and is embodied by testing for stationarity (or lack of it). One of the earliest tests for the overall second order nonstationarity of a time series, is the test proposed by Priestley and Subba Rao (1969), which evaluates the log of their evolutionary spectral estimates through a two-factor analysis of variance at different moments of time. von Sachs and Neumann (2000) constructed a stationarity test which estimates the evolutionary (Fourier) spectrum through localised periodograms. They then make use of the capability of Haar wavelets to capture discontinuities across the spectral density, thus identifying any deviation from covariance stationarity. Expanding on the work of Priestley and Subba Rao (1969), Ahamada and Boutahar (2002) attain an alternative test statistic and demonstrate their test on an application to exchange rates. Paparoditis (2009, 2010) compute sample spectral densities (local peridograms) on a moving window of observations and compare these estimates via an L_2 measure of deviation to a spectral density estimate obtained for the overall time series. A testing procedure to identify whether two locally stationary time series are costationary was proposed by Cardinali and Nason (2010). They utilise a parametric bootstrap (see Davison and Hinkley (1997)) resampling procedure to obtained pseudo-test statistic values from processes under the null hypothesis that the estimated

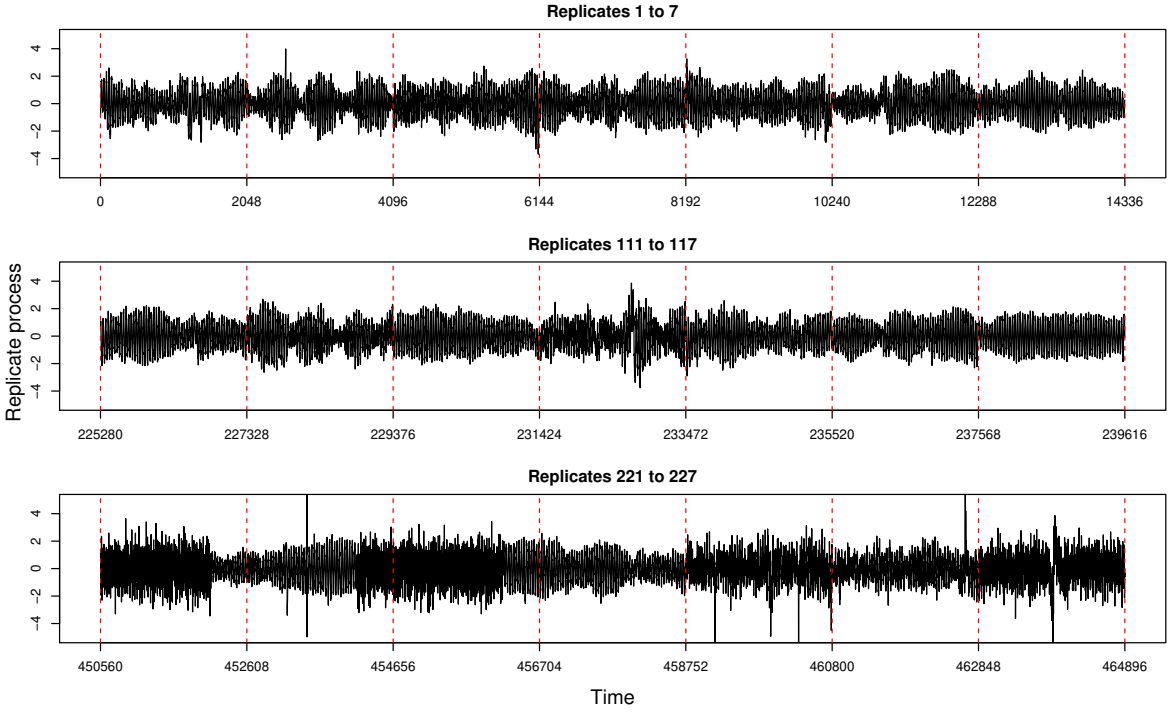


Figure 2: Concatenated series of the nucleus accumbens (NAc) data in the correct response trials (replicates). (Note each trial is a time series and concatenation is only used to aid meta-process visualisation.)

spectrum is constant as a function of time. Several of the Fourier tests of stationarity are based on selecting segments of the time series for comparison, and Dwivedi and Subba Rao (2011) draw attention to the sensitivity of the test to the choice of segment length and to how the number of segments can impact the rate of convergence to the asymptotic distribution. Whilst many of the above mentioned tests were constructed with Fourier spectra in mind, Nason (2013) proposed a test of similar nature to von Sachs and Neumann (2000) that utilises Haar wavelet coefficients in order to investigate the constancy over scales of the wavelet spectrum instead of Fourier frequencies.

Alternatively, approaches that extract between-series structure have been considered for discriminant analysis and clustering (Hargreaves *et al.*, 2018). Krafty (2016) propose to model transfer functions as stochastic variables in order to account for within-group spectral variability. In the presence of within-group variability, their methodology leads to more accurate classification than under the typical assumption that the series within each group are independent and identically distributed. Euán *et al.* (2018) introduce the hierarchical spectral merger (HSM) algorithm which aims to cluster time series with similar spectral behaviour via a measure of total variation distance. The HSM method is developed for univariate stationary time series, however the authors suggest that accounting for nonstationarity can be achieved through use of the evolutionary wavelet spectrum as seen in the multivariate discriminant analysis of Maharaj and Alonso (2014). Perhaps the closest methods that could be framed to compare the nonstationary characteristics between multiple time series lie within spectral comparison for classification. Fryzlewicz and Ombao (2009) provide a general procedure to classify processes into groups through comparing the estimated wavelet spectrum with the spectral characteristics associated to each group. Hargreaves et al. (2019) were first to develop a hypothesis test which establishes whether the associated evolutionary wavelet spectra of two groups are significantly different. However, the setting of these analyses does not feature a natural ordering of the multiple time series.

Thus, to the best of our knowledge, within the current time series literature there exist very limited testing procedures for the existence of evolutionary characteristics across *ordered* multiple-trial processes. Of the little literature that exists specifically for the testing of spectral evolution over the replicate domain, Fiecas and Ombao (2016) proposed a Fourier-based resampling procedure using log-linear models. Their procedure echoes the functional regression problem

of testing the equality of the curves in the time domain (for example see Dette and Neumeyer (2001)).

Functional regression approaches often deal with replicate time series data by projecting it into the Fourier or wavelet domain where the spectral representations become subject to modelling. For an introduction to functional regression and more specifically functional linear models, see Ramsay and Silverman (2005), while Morris (2015) provides a review on current methods in the field. When testing for differences between curves in the spectral domain, techniques are centered around functional ANOVA (see e.g., Ramsay and Silverman (2005)). Shumway (1988) computed individual test statistics for each given frequency (in the Fourier domain) to detect differences in the Fourier spectra of the mean curves. The individual test statistics were combined in Fan and Lin (1998) to form an overall test based on the adaptive Neyman test to identify differences between two groups of curves, which they then generalise to compare multiple groups of curves through an adaptive high-dimensional ANOVA. Developments of functional (F)ANOVA methods have been established in the wavelet domain (WANOVA) by Raz and Turetsky (1999); Vidakovic (2001) and McKay *et al.* (2013), with the second highlighting that the decorrelation properties of wavelet transformations (Chang and Stein (2013)) are beneficial for regularisation. Atkinson et al. (2017) note the favourable properties of WANOVA to smooth, reduce dimensionality and decorrelate time series data and thus develop a WANOVA-based model validation process. In the setting of locally stationary processes (Dahlhaus, 1997), Guo et al. (2003) propose a smoothing spline (SS-)ANOVA model fitted to time-varying log-periodograms constructed using SLEX basis functions (see Ombao *et al.* (2002)).

This article proceeds as follows. Section 2.1 reviews the relevant modelling framework underpinning this work, and Sections 2.2 and 2.3 introduce the proposed location-specific and global tests to identify a replicate-effect, respectively. In Section 3, the performance of both tests is assessed through thorough simulation studies. Section 4 addresses the neuroscience macaque LFPs study and we apply our proposed tests of replicate-effect to both hippocampus and nucleus accumbens data. We conclude this work in Section 5.

2. Methodology

2.1. Brief introduction of replicate locally stationary wavelet processes

Before we describe the framework for the multiple-trial (or replicate) locally stationary wavelet model of Embleton *et al.* (2020) that will underpin the methodology we develop here, we recall some of the defining features of the locally stationary wavelet (LSW) framework of Nason *et al.* (2000). The LSW model provides a *time-scale* representation of nonstationary time series with time-varying second order structure, where the building blocks are the discrete non-decimated wavelets (see e.g. Vidakovic (1999) for an extensive introduction to wavelets). For $T = 2^J$, a sequence of stochastic processes $\{X_{t;T}\}_{t=0,\ldots,T-1}$ is a LSW process if it admits the representation

$$X_{t;T} = \sum_{j=1}^{\infty} \sum_{k \in \mathbb{Z}} \omega_{j,k;T} \psi_{j,k}(t) \xi_{j,k},$$

where for scale j and location k, $\omega_{j,k;T}$ is the amplitude corresponding to the discrete non-decimated wavelet $\psi_{j,k}(t)$ and $\{\xi_{j,k}\}$ are a set of orthonormal random variables. Modelling under the concept of local stationarity means that the amplitudes $\{\omega_{j,k;T}\}_k$, change slowly over time and their variation is controlled by a smoothly varying continuous Lipschitz function $W_j(k/T)$, that can be thought of as a scale (j) and time (k) dependent transfer function (Fryzlewicz and Nason, 2006). Here, scales $j = 1, 2, \ldots$ are understood as finest, second finest and so on.

The original LSW model does not capture the dynamics of time series data recorded for several trials over the course of an entire experiment. Our setting here presents additional challenges, notably the fact that these signals behave in a way that is nonstationary at multiple temporal scales, (i) locally (within each trial) and (ii) globally (across trials). The novel framework of Embleton *et al.* (2022), reviewed below, deals with these challenges.

A sequence of stochastic processes $\{X_{t;T}^{r;R}\}$, with time $t = 0, \ldots, T-1$ where $T = 2^J$ and replicate $r = 0, \ldots, R-1$ where $R = 2^{J'}$ is a *multiple-trials locally stationary wavelet (MULT-LSW)* process if it admits the following representation

$$X_{t;T}^{r;R} = \sum_{j=1}^{\infty} \sum_{k \in \mathbb{Z}} \omega_{j,k;T}^{r;R} \psi_{j,k}(t) \xi_{j,k}^{r}, \tag{1}$$

where for each replicate (or trial) r and within-trial time k, $\omega_{j,k;T}^{r;R}$ are the amplitudes for the non-decimated *discrete* wavelets $\psi_{j,k}(t)$ at scale $j \geq 1$. The family of discrete wavelets can be conceptualised as consisting of scale-specific vectors whose entries are wavelet filters and whose length is intimately connected to the corresponding number of vanishing moments (for further details the reader is directed to Nason *et al.* (2000)).

The innovations $\{\xi_{j,k}^r\}$ are a set of orthonormal random variables (uncorrelated, each with zero mean and variance equal to one) with properties as detailed below. Letting $\nu = r/R$ denote rescaled replicate and $z = k/T$ denote rescaled within-trial time, the quantities in (1) possess the following properties:

1. For all j, k and r, $\mathbf{E}[\xi_{j,k}^r] = 0$ ($\Rightarrow \mathbf{E}[X_{t;T}^{r;R}] = 0$).
2. $\mathbf{E}[\xi_{j,k}^r \xi_{j',k'}^{r'}] = \delta_{j,j'}\delta_{k,k'}\delta_{r,r'}$ ($= \mathrm{cov}(\xi_{j,k}^r, \xi_{j',k'}^{r'})$). This amounts to assuming uncorrelated replicates.
3. For each scale $j \geq 1$, there exists a Lipschitz continuous transfer function in both rescaled time (z) and rescaled replicate (ν), denoted by $\widetilde{W}_j(z,\nu)$ with the following properties

   (a)
   $$\sum_{j=1}^{\infty} \left|\widetilde{W}_j(z,\nu)\right|^2 < \infty \text{ uniformly in } z \in (0,1), \nu \in (0,1).$$

   (b) Let L_j^ν denote the bounded Lipschitz constant corresponding to the time dimension at a particular (rescaled) replicate (ν) and scale j. Similarly, denote by N_j^z the bounded Lipschitz constant corresponding to the replicate dimension at a particular (rescaled) time (z) and scale j. Denote $L_j = \sup_{\nu\in(0,1)} L_j^\nu$ and $N_j = \sup_{z\in(0,1)} N_j^z$, and assume they are uniformly bounded in j. Further assume that
   $$\sum_{j=1}^{\infty} 2^j L_j < \infty \text{ and } \sum_{j=1}^{\infty} 2^j N_j < \infty.$$

   (c) There exist sequences of bounded replicate-specific constants $\{C_j^r\}_r$ and location-specific constants $\{D_j^k\}_k$, such that for each T and R respectively, the amplitudes are forced to vary slowly, in the sense that
   $$\sup_{k=0:T-1} \left|\omega_{j,k;T}^{r;R} - \widetilde{W}_j\left(\frac{k}{T}, \frac{r}{R}\right)\right| \leq \frac{C_j^r}{T}, \quad \forall j, r, \tag{2}$$
   $$\sup_{r=0:R-1} \left|\omega_{j,k;T}^{r;R} - \widetilde{W}_j\left(\frac{k}{T}, \frac{r}{R}\right)\right| \leq \frac{D_j^k}{R}, \quad \forall j, k. \tag{3}$$

   Denote $C_j = \sup_r C_j^r$ and $D_j = \sup_k D_j^k$ and assume the sequences $\{C_j\}$, $\{D_j\}$ fulfill $\sum_{j=1}^{\infty} 2^j C_j < \infty$ and $\sum_{j=1}^{\infty} 2^j D_j < \infty$.

While the original LSW formulation is for a single-replicate, its multivariate extension, the MV-LSW process of Park *et al.* (2014) can deal with a finite number of replicates but does not require, nor can it account for, the ordering of replicates. In contrast, the MULT-LSW framework captures the dynamics of time series data recorded for several *ordered* trials over the course of an entire experiment, with the *multiscale* insight afforded by the wavelet decomposition offering an additional layer of information and further contributing to the understanding of the *within-trial* evolution.

As is common in spectral domain analysis (both Fourier and wavelet-based), the time- and replicate-specific multiscale transfer functions $\{\widetilde{W}_j(\cdot,\cdot)\}_j$ are not examined directly but through the replicate evolutionary wavelet spectrum that quantifies the contribution to the process variance attributed across scales (j) to each time *and* replicate. Using equations (2) and (3), we define the within-trial evolutionary wavelet spectrum at scale j, rescaled replicate ν and rescaled within-trial time z by

$$S_j(z,\nu) = \left|\widetilde{W}_j(z,\nu)\right|^2 = \lim_{\substack{T\to\infty \\ R\to\infty}} \left(\left|\omega_{j,\lfloor zT\rfloor;T}^{\lfloor \nu R\rfloor;R}\right|^2\right),$$

where $\lfloor zT \rfloor$ and $\lfloor \nu R \rfloor$ denote the largest integer less than or equal to zT and νR, respectively.

A useful transformed spectral quantity is the β-spectrum, defined at scale j, rescaled replicate ν and rescaled within-trial time z as

$$\beta_j(z,\nu) = \sum_{l=1}^{\infty} A_{j,l} S_l(z,\nu),$$

where $A_{j,l} = \langle \Psi_j, \Psi_l \rangle = \sum_{\tau \in \mathbb{Z}} \Psi_j(\tau)\Psi_l(\tau)$ is the inner product matrix of the autocorrelation wavelets (Nason *et al.*, 2000). The β-spectrum was introduced by Fryzlewicz and Nason (2006) and it is often easier to work with theoretically than the spectrum (Nason, 2013) since it can be directly estimated by the raw periodogram, as we shall see next.

Embleton *et al.* (2020) remark that the MULT-LSW theoretical development could be extended to encompass bounded variation jumps across the trial-dimension, which is the setting of e.g., the test of stationarity of Nason (2013) across the time-dimension. Under a bounded variation assumption for the amplitudes, we note that the testing methodology that will be introduced next still holds.

2.2. Testing for a replicate-effect: a location-specific approach

The MULT-LSW process is defined to allow for a controlled evolution of the spectral quantities over the ordered trials. Intuitively, in order to test whether this evolution is indeed manifest, we formulate the null hypothesis that the wavelet spectrum is constant *over trials*. We thus construct next a pointwise test that will identify those trials that depart from replicate-domain spectral constancy, at each time k. However, we note here that should a similar spectral quantity be available as derived from e.g. a Fourier as opposed to a wavelet multiple-trial process (Fiecas and Ombao, 2016), the following setup would still notionally hold (von Sachs and Neumann, 2000).

Formally, we propose to test that

$$\begin{aligned} H_0^k &: S_j\left(\frac{k}{T}, \nu\right) \text{ is constant as a function of } \nu, \forall j \quad \text{versus} \\ H_A^k &: \exists j^\star \text{ such that } S_{j^\star}\left(\frac{k}{T}, \nu\right) \text{ is non-constant over } \nu \in (0,1). \end{aligned}$$

For a given time k, it can be immediately verified that our problem above can be re-framed as the pointwise testing

$$\begin{aligned} H_0^k &: \beta_j\left(\frac{k}{T}, \nu\right) \text{ is constant as a function of } \nu, \forall j \quad \text{versus} \\ H_A^k &: \exists j^\star \text{ such that } \beta_{j^\star}\left(\frac{k}{T}, \nu\right) \text{ is non-constant over } \nu \in (0,1). \end{aligned}$$

Tasks such as process representation and spectral estimation often result in a preference towards the use of smoother wavelets due to their superior convergence rates. However, when one is concerned with detecting departures from constancy, Haar wavelets become the choice tool (von Sachs and Neumann, 2000; Nason, 2013), hence here we propose to assess the constancy of $\beta_j(\frac{k}{T}, \cdot)$ by assessing the departures of its Haar wavelet coefficients from 0.

We therefore proceed to define the Haar wavelet coefficients of the transformed spectral function $\beta_j(\frac{k}{T}, \cdot)$ at scale $j = 1, \ldots, J$ and time $k = 0, \ldots, T-1$ as

$$\eta_{i,p}^{(j,k)} = \int_0^1 \beta_j\left(\frac{k}{T}, \nu\right) \psi_{i,p}^H(\nu) d\nu, \tag{4}$$

where $\psi_{i,p}^H(\nu)$ denote the Haar wavelet at (Haar) scale $i = 1, ..., J' = \log_2(R)$ and (Haar) location $p = 1, \ldots, 2^i - 1$. Crucially, note that the Haar wavelet is operating over the (rescaled) *replicate* dimension, as opposed to the time dimension. From the wavelet property $\int_0^1 \psi^H(\nu) d\nu = 0$, it directly follows that the Haar wavelet coefficients in (4) are zero under the null hypothesis H_0^k.

Our null hypothesis now amounts to $H_0^k : \eta_{i,p}^{(j,k)} = 0$ for all j, i, and p. If for any k there exists a scale $j^\star$ and Haar scale and location $(i^\star, p^\star)$ such that $\eta_{i^\star,p^\star}^{(j^\star,k)} \neq 0$, then we conclude that $\beta_j\left(\frac{k}{T}, \cdot\right)$ is not constant and the meta-process

$\{X_{t;T}^{r;R}\}$ displays a replicate-effect, i.e., the spectral quantities are indeed evolving over replicates and a MULT-LSW approach to modelling the data is to be preferred over a blanket-approach across all trials.

In order to estimate the Haar wavelet coefficients in equation (4), we replace the spectral quantity $\beta_j\left(\frac{k}{T},\nu\right)$ by means of its corresponding raw periodogram, denoted $I_{j,k;T}^{\lfloor \nu R\rfloor;R}$. For a given time $k=\lfloor zT\rfloor$, we proceed to use the associated wavelet periodogram at all scales j and over all replicates $r=\lfloor \nu R\rfloor$, and for notational simplicity we denote this quantity as $I_{j,k}^{r}$ and define it as

$$I_{j,\lfloor zT\rfloor;T}^{\lfloor \nu R\rfloor;R}=I_{j,k}^{r}=\left|d_{j,k;T}^{r;R}\right|^2, \tag{5}$$

where for scale j, trial r and within-trial time k, $d_{j,k;T}^{r;R}=\sum_{t=0}^{T-1}X_{t;T}^{r;R}\psi_{j,k}(t)$ are the process empirical wavelet coefficients constructed using a family of discrete non-decimated wavelets, $\{\psi_{j,k}(t)\}_{j,k}$. Under the assumption that within a fixed trial r the process innovations $\{\xi_{j,k}^{r}\}_{j,k}$ are normally distributed, Embleton *et al.* (2022) have shown that (5) is an asymptotically unbiased estimator for the β-spectrum, specifically

$$\mathbf{E}(I_{j,\lfloor zT\rfloor;T}^{\lfloor \nu R\rfloor;R})=\beta_j(z,\nu)+\mathcal{O}(2^jT^{-1})+\mathcal{O}(R^{-1}).$$

We note that this is a commonplace assumption in time series analysis in general and in LSW modelling in particular, e.g. Van Bellegem and von Sachs (2008). Nason (2013) illustrate the non-limiting character of this assumption for practical applications, which holds for data arising from various fields, e.g. for experimental circadian data (Hargreaves et al., 2019). In Appendix B we show that indeed the normality assumption is tenable for our neurological experimental data.

The result above, coupled with equation (4), suggests that at time k and scale j we estimate the Haar wavelet coefficients by

$$\hat{\eta}_{i,p}^{(j,k)}=2^{-i/2}\left(\sum_{l=0}^{2^{i-1}-1}I_{j,k}^{2^ip-l}-\sum_{q=2^{i-1}}^{2^i-1}I_{j,k}^{2^ip-q}\right)$$

where $i=1,...,J'$ and $p=1,...,2^i-1$ and we point out that the Haar wavelet transform is taken over replicates.

For coarse (Haar) scales $i=\mathcal{O}(\log_2(R))$ (and under the technical assumptions of von Sachs and Neumann (2000, Proposition 3.1) or Nason (2013, Proposition 1) on spectral behaviour, here in the replicate argument ν), the estimator $\hat{\eta}_{i,p}^{(j,k)}$ is asymptotically normal and has the following properties for all j, i and p and for a fixed k

1. $\mathbf{E}\left[\hat{\eta}_{i,p}^{(j,k)}\right]=\eta_{i,p}^{(j,k)}+\mathcal{O}(2^jT^{-1}R^{1/2})+\mathcal{O}(R^{-1/2})$,
2. $\text{var}\left(\hat{\eta}_{i,p}^{(j,k)}\right)=2R^{-1}\int_0^1\beta_j^2\left(\frac{k}{T},\nu\right)\left(\psi_{i,p}^{H}(\nu)\right)^2d\nu+\mathcal{O}(2^{2j}(RT)^{-1})+\mathcal{O}(2^jR^{-2})$.

The coarse scale constraint ensures we achieve the asymptotic rates above and the proof follows the steps presented in Lemma 1 of Nason (2013) and uses the results in Proposition 2 of Embleton *et al.* (2020), hence it is omitted here.

When replacing $(\sigma_{i,p}^{(j,k)})^2=\text{var}\left(\hat{\eta}_{i,p}^{(j,k)}\right)$ by an estimator $(\hat{\sigma}_{i,p}^{(j,k)})^2$, von Sachs and Neumann (2000, Theorem 3.1) guarantees interval coverage rates asymptotically equivalent to those attained by means of a normal distribution under the null hypothesis H_0^k, a result that Nason (2013) also use in their development.

For a fixed time k, under the null hypothesis H_0^k we have that $\beta_j\left(\frac{k}{T},\cdot\right)$ is constant over the rescaled replicate dimension, hence $\eta_{i,p}^{(j,k)}=0$ for all j,i, and p. This leads us to propose testing for H_0^k via the test statistics

$$\boldsymbol{T}_{i,p}^{(j,k)}=\frac{\hat{\eta}_{i,p}^{(j,k)}}{\hat{\sigma}_{i,p}^{(j,k)}},\ \text{at all } j,\ i,\ p, \tag{6}$$

which for all original scales j, Haar scales i and Haar locations p are then compared with a critical value obtained from the normal distribution.

One way to estimate $(\sigma_{i,p}^{(j,k)})^2$ is by taking

$$(\hat{\sigma}_{i,p}^{(j,k)})^2=2^{-i}\left[\sum_{l=0}^{2^{i-1}-1}\text{var}(I_{j,k}^{2^ip-l})+\sum_{q=2^{i-1}}^{2^i-1}\text{var}(I_{j,k}^{2^ip-q})\right], \tag{7}$$

where we make use of the fact that the replicates are uncorrelated, hence $\mathrm{cov}(I^r_{j,k}, I^{r'}_{j,k}) = 0$ for $r \neq r'$.

Under the null hypothesis, we could also estimate $(\sigma^{(j,k)}_{i,p})^2$ by replacing the unknown $\beta^2_j(k/T, \nu)$ in property 2 by an average of the squared raw wavelet periodograms across all replicates r.

This gives

$$(\hat{\sigma}^{(j,k)}_{i,p})^2 = 2R^{-1}\overline{I}^{(2)}_{j,k}\int_0^1 \{\psi^H_{i,p}(\nu)\}^2 d\nu = 2R^{-1}\overline{I}^{(2)}_{j,k}, \tag{8}$$

where we have used the unit norm property of Haar wavelets at all i and p, and we denoted $\overline{I}^{(2)}_{j,k} = R^{-1}\sum_{r=0}^{R-1}\left(I^r_{j,k}\right)^2$.

As we formulated a pointwise testing problem, we deal with the multiple hypothesis testing by taking a false discovery rate approach (Benjamini and Hochberg, 1995) or, for a stricter procedure, the Bonferonni correction (Nason, 2013). Additionally we note that the number of hypothesis tests carried out per location is dependent on how many original wavelet scales j and Haar wavelet scales j' we choose to test, namely

$$\#\{tests\ per\ location\} = \#\{original\ j\ scales\} * \#\{Haar\ j'\ scales\}. \tag{9}$$

2.3. Testing for a replicate-effect: a global approach

So far, we aimed to identify those times k for which the null hypothesis, $H^k_0 : S_j\left(\frac{k}{T}, \nu\right)$ is constant as a function of $\nu, \forall j$, holds. This approach identifies which, if any, within-replicate times and trials display sufficient evidence to assert that significant spectral changes occurred across the experiment. However, depending on the context, one might find it more useful to test the overall null hypothesis

$$H_0 : S_j\left(\frac{k}{T}, \nu\right) \text{ is constant as a function of } \nu,\ \forall j,\ k.$$

In order to propose an appropriate test statistic for this overarching hypothesis, we first consider the following measure for the spectral departure from constancy through the replicate-dimension for a particular rescaled within-trial time $z = k/T$ and for each scale j,

$$\mathbf{T}\left(S_j(z,\cdot)\right) = \int_0^1 \left(S_j(z,\nu) - \bar{S}_j(z)\right)^2 d\nu,$$

where we denote by $\bar{S}_j(z) = \int_0^1 S_j(z,\nu)\, d\nu$ (see also Cardinali and Nason (2010) for an alternative context where such a measure proved successful).

The *average spectral departure from constancy across all scales* may then be measured by

$$\mathbf{T}_{ave}\left(\{S_j(z,\cdot)\}_j\right) = J^{-1}\sum_{j=1}^{J}\mathbf{T}\left(S_j(z,\cdot)\right).$$

Furthermore, we quantify the *overall spectral departure* not only across all scales but also through time, which for brevity we denote $\mathbf{T}_{ave}(S)$, as follows

$$\begin{aligned}\mathbf{T}_{ave}(S) &= \int_0^1 \mathbf{T}_{ave}\left(\{S_j(z,\cdot)\}_j\right) dz, \\ &= J^{-1}\sum_{j=1}^{J}\int_0^1 \mathbf{T}\left(S_j(z,\cdot)\right) dz.\end{aligned} \tag{10}$$

Under the null hypothesis (H_0) of spectral constancy across trials, note that all measures above are 0. Conversely, since the spectral quantities are positive, it is straightforward to show that if the aggregated measure $\mathbf{T}_{ave}(S)$ in (10) is 0, then the null hypothesis of spectral constancy through replicates at all times also holds. Hence we shall treat significant departures from 0 as indicative of departures from the null hypothesis.

As the true replicate wavelet spectra are unknown, we estimate the measures of spectral departure from constancy by means of their *sample* equivalents, built upon a well-behaved spectral estimator. Hence we obtain the sample quantities

$$\begin{aligned}
\mathbf{T}(\hat{S}_{j,k}) &= \text{var}_\nu\left(\hat{S}_j\left(\frac{k}{T},\nu\right)\right), \text{ for } \mathbf{T}\left(S_j\left(\frac{k}{T},\cdot\right)\right), \\
\mathbf{T}_{ave}\left(\left\{\hat{S}_{j,k}\right\}_j\right) &= J^{-1}\sum_{j=1}^{J}\text{var}_\nu\left(\hat{S}_j\left(\frac{k}{T},\nu\right)\right), \text{ for } \mathbf{T}_{ave}\left(\left\{S_j\left(\frac{k}{T},\cdot\right)\right\}_j\right), \\
\mathbf{T}_{ave}\left(\left\{\hat{S}_{j,k}\right\}_{j,k}\right) &= (JT)^{-1}\sum_{k=0}^{T-1}\sum_{j=1}^{J}\text{var}_\nu\left(\hat{S}_j\left(\frac{k}{T},\nu\right)\right), \text{ for } \mathbf{T}_{ave}(S).
\end{aligned}$$

In the above, var_ν denotes the usual empirical variance, here taken over replicates within each scale j and at each time k. We propose to use for $\hat{S}_j\left(\frac{k}{T},\nu\right)$ the corrected, replicate- and time-smoothed wavelet periodogram, an asymptotically consistent and unbiased estimator for the unknown wavelet spectrum (Embleton *et al.*, 2022). For completeness, the spectral estimator is given by

$$\hat{S}_j(\frac{k}{T},\nu) = \sum_{l=1}^{J} A_{j,l}^{-1}\tilde{\tilde{I}}_{l,k;T}^{\lfloor\nu R\rfloor;R}, \text{ with the smoothed periodogram defined as} \tag{11}$$

$$\tilde{\tilde{I}}_{j,k;T}^{r;R} = (2M+1)^{-1}(2M_T+1)^{-1}\sum_{s=-M}^{M}\sum_{t=-M_T}^{M_T} I_{j,k+t;T}^{r+s;R}, \tag{12}$$

where $(2M+1)$ and $(2M_T+1)$ denote the smoothing windows in the replicate- and time-dimension, respectively.

The time-smoothing window parameter M_T is chosen automatically using the method proposed by Nason (2013). The trial-smoothing window $(2M+1)$ with a choice of $M = \frac{3}{4}\sqrt{R}$ appears to work well across all our investigations (Embleton *et al.*, 2022). The robustness of LSW estimation to window width choices and form are illustrated in Killick *et al.* (2020), while we note the discussion in Cryer and Chan (2008, §14.2) and suggest that for a deeper understanding of the spectral characteristics, but outside the testing framework, a user might wish to obtain estimates over a range of M, e.g. $M = \frac{1}{2}\sqrt{R}, \frac{3}{4}\sqrt{R}, \sqrt{R}$.

Although under the null hypothesis the distribution of the test statistic $\mathbf{T}_{ave}(\hat{S})$ is unknown, since $\left\{\hat{S}_j\left(\frac{k}{T},\nu\right)\right\}_{j,k}$ is a consistent estimator for the true spectrum $\left\{S_j\left(\frac{k}{T},\nu\right)\right\}_{j,k}$ for all rescaled trials ν, the bootstrap approach of Davison and Hinkley (1997) is a valid alternative which we propose to use here (see also Cardinali and Nason (2010)). In order to allow for parametric resampling, we carry out the bootstrap simulations under a Gaussian innovations assumption, in agreement with the location-specific testing development in Section 2.2 and with the assumptions under which the asymptotic consistency results of Embleton *et al.* (2022) were obtained. The statistical significance of the observed test statistic is then established by means of a Monte Carlo approach used to generate pseudo-test statistics values from a process with properties akin to those of the original process under the null hypothesis. The p-value of the test is then simply approximated by the proportion of these pseudo-test statistics (generated under the null hypothesis of spectral constancy across replicates) that exceed the value of the observed test statistic.

The proposed algorithm for testing the presence of a replicate-effect is described in Algorithm 1.

3. Simulation study

In this section we thoroughly investigate through simulation the capability of our proposed (location-specific and global) tests to identify whether a replicate-effect is present in simulated data that may or may not feature a replicate-dimension evolution. Test performance for both tests is quantified by means of empirical power and size. However, since the location-specific test provides T decisions, we explore its performance by further employing binary classification measures that allow us to meaningfully evaluate the *overall* test behaviour.

`Proposed BootReplicateTest algorithm:`
Assume we observe time series across several trials (replicates), and we model the data as a MULT-LSW process (see Section 2.1). We want to assess whether a replicate-effect exists over the course of the experiment.

0. Estimate the meta-process spectral structure, and denote the estimator $\hat{S}_j(z,\nu)$ at scale j, rescaled replicate $\nu = r/R$ and rescaled within-trial time $z = k/T$. Use the estimator constructed as shown in (11), by means of equations (5) and (12).
1. For each scale j and time k, under the null hypothesis of constancy through replicates, compute the average scale- and time- specific periodogram
$\bar{\hat{S}}_j(k/T) = R^{-1}\sum_{r=0}^{R-1}\hat{S}_j(k/T, r/R)$.
2. Compute the test statistic $\mathbf{T}_{ave}\left(\left\{\hat{S}_j(k/T,\cdot)\right\}_j\right)\left(=\mathbf{T}_{ave}^{obs;k}\right)$ across all times k, and aggregate these into
$\mathbf{T}_{ave}^{obs} = ((R-1)JT)^{-1}\sum_{j=1}^{J}\sum_{k=0}^{T-1}\sum_{r=0}^{R-1}\left(\hat{S}_j(k/T,r/R) - \bar{\hat{S}}_j(\frac{k}{T})\right)^2$.
3. *Iterate* for $b = 1$ to B bootstraps:
   - for each scale j, simulate a MULT-LSW process $\{X_{t;T}^{r;R}\}^{(b)}$ with squared amplitudes given by $\bar{\hat{S}}_j(\frac{k}{T})$ for all r,
   - compute the corresponding test statistic $\mathbf{T}_{ave}^{(b)}$ corresponding to the simulated data.
4. Compute the test $p - value = (1 + \#\{\mathbf{T}_{ave}^{(b)} \geq \mathbf{T}_{ave}^{obs}\})/(B+1)$.

Algorithm 1: Proposed global bootstrap test for assessing the existence of a replicate-effect across the meta-process modelled using the MULT-LSW framework.

3.1. Simulations for the location-specific testing methodology

We first turn to exploring the capacity of our location-specific test to identify which, if any, locations along a simulated MULT-LSW process exhibit sufficient evidence to deem that a replicate-effect exists across the meta-process. We investigate the test performance for various values for R and T, and we employ the false discovery rate (FDR) with 5% nominal size to control the number of false rejections (we perform $(J-3)\times(J' - \lceil J'/2\rceil + 1)$ tests, recalling that $J = \log_2(T)$ and $J' = \log_2(R)$). For each time location, we obtain empirical power/ size estimates through counting the number of times the test correctly/ incorrectly identifies a breach in spectral constancy. In practice, we may only be concerned with whether the test successfully identifies an evolution in the spectral quantities over the replicates of a MULT-LSW process. As such, to quantify the overall test performance, we provide statistical measures for a binary classification test, averaged over 100 simulations. These measures are given in Table 1, where a 'positive' is understood to be a location that is identified by the test as rejecting the null hypothesis of spectral constancy across replicates, and a 'negative' is understood to be a location that fails to reject the null. The true positive and true negative rates should be viewed as a collective assessment of the empirical test power and size, and are also evaluated in the context of varying sample sizes. The reported results were obtained through estimating the variance using equation (7).

The simulations that follow are designed to encapsulate a wide range of scenarios that challenge the proposed test, from the lack of replicate-effect across time (simulation S1A), to featuring 'bursts' of spectral activity along time and replicate segments that are progressively shrinking in both length and intensity (S1B-C, S2), to finally exhibiting a complex replicate-effect behaviour along the entire duration of the experiment (S3), thus also exploring the inference robustness to data departing from the modelling assumptions.

Simulation 1 consists of variations (S1A–C) of the same underlying squared sine spectral characteristics overlaid with a 'burst' value ranging from 0 (the absence of replicate evolution in the true spectrum, thus measuring test size in S1A) to 1, 2 and 5, as well as a change in the ranges of replicates and times exhibiting these bursts (S1B-C, see Figure 3 and Figure 15 in Appendix C for $R = 256$ and $T = 512$). For their precise mathematical definitions, please refer to Appendix C. Here we also provide visualisations of the empirical power and size estimates across all locations, see Figure 16 where the horizontal lines at 1 and 0 depict locations expected to be rejected and not rejected, respectively. These illustrate the empirical rates are well calibrated to the nominal ones, see also Nason (2013).

For various R and T, binary classification measures (and additional 'sd' values for their corresponding standard deviations across the 100 simulations), as defined in Table 1, are reported in Table 2 for simulation S1A. These

Binary Classification rates	
Rate	Formula
True Positive Rate (TPR)	$\frac{tp}{tp+fn}$
True Negative Rate (TNR)	$\frac{tn}{tn+fp}$
False Discovery Rate (FDR)	$\frac{fp}{tp+fp}$
False Omission Rate (FOR)	$\frac{fn}{tn+fn}$
Fowlkes-Mallows Index (F-M)	$\sqrt{\frac{tp}{tp+fp} \cdot \text{TPR}}$
Accuracy (ACC)	$\frac{tp+tn}{tp+tn+fp+fn}$
Prevalence Threshold (PT)	$\frac{\sqrt{\text{TPR}(-\text{TNR}+1)}+\text{TNR}-1}{\text{TPR}+\text{TNR}-1}$

Table 1: List of binary classification rate formulas where 'tp' and 'fp' denote the true and false positives, and 'tn' and 'fn' denote the true and false negatives.

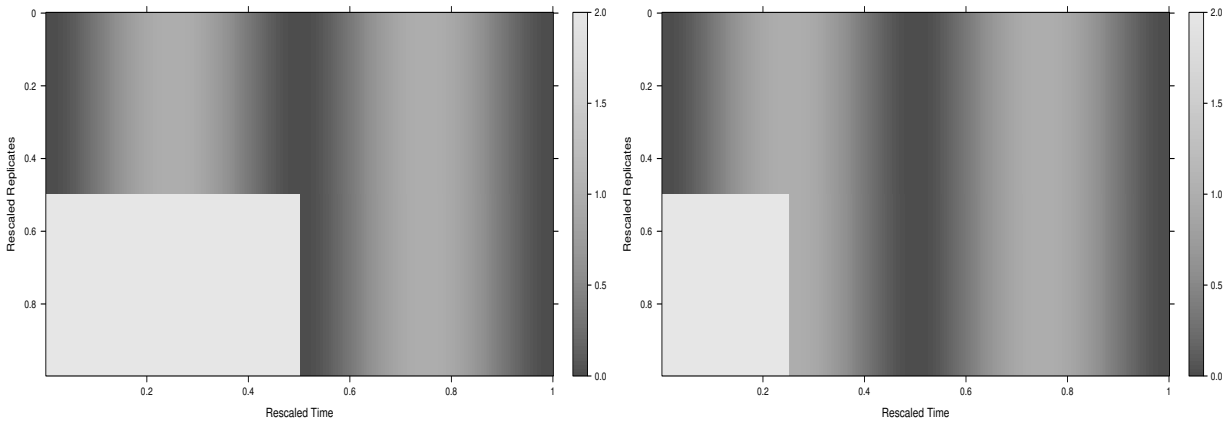


Figure 3: Time-replicate plot of the true spectra in level 5 with a burst value of 2, for simulation 1B (*left*) and simulation 1C (*right*).

demonstrate that the test is extremely capable to correctly identify all locations that do not breach the null hypothesis of no replicate-effect, when indeed none exists. The true negative rates (TNR) measure the proportion of actual negatives correctly identified by the test and the accuracy (ACC) gives the proportion of the locations correctly identified as either positive or negative. Here we report fewer measures due to the absence of true positives (locations with significantly differing spectra across replicates) and for the same reason we observe that the TNR $=$ ACC. Note however that in general the ACC alone can be misleading since it is not explicit on which of the positives or negatives are contributing more to the accuracy.

Binary classification measures are reported for simulation S1B in Table 3 for different bursts, R and T values. Considering each burst individually, it is apparent in each case that the test performance improves as R and T increases, which comes as no surprise given the improved spectral estimation as $R, T \to \infty$. This improvement is seen in the true positive rates (TPR), and furthermore the false discovery rates (FDR) and false omission rates (FOR) which respectively measure the proportion of identified positives and identified negatives that are false, are decreasing. The Fowlkes-Mallows index (F-M) measures the similarity between the TPR and the proportion of identified positives that are true (the positive predictive rate), where greater similarity is indicated by a higher index value. When comparing across the bursts, the measures show that the test gets better at identifying the true positives as the burst increases

(greater difference between spectra) at a cost of a slight worsening of the TNR. This is probably due to the power leakage across locations during estimation. Also, the measures give evidence to the test struggling to identify small differences between a defined burst and the squared sine spectral value, while there is significant improvement as the burst increases. Finally, in Figure 4 we visually supplement the above results with a receiver operator characteristic (ROC) plot, which plots the TPR against the false positive rate (FPR), with the latter measuring the proportion of actual negatives incorrectly identified as positive, i.e., $(1 - \text{TNR})$. This plot displays how an increase in R and T, and an increase in the difference in value of the spectra between replicates (stronger burst) lead to an improvement in the performance of the test correctly identifying those locations associated to evolving spectral characteristics over the replicates.

In Table 4 we report statistical measures on the overall test for simulation S1C. The results for this test are very favourable, with the majority of true positives and negatives being identified.

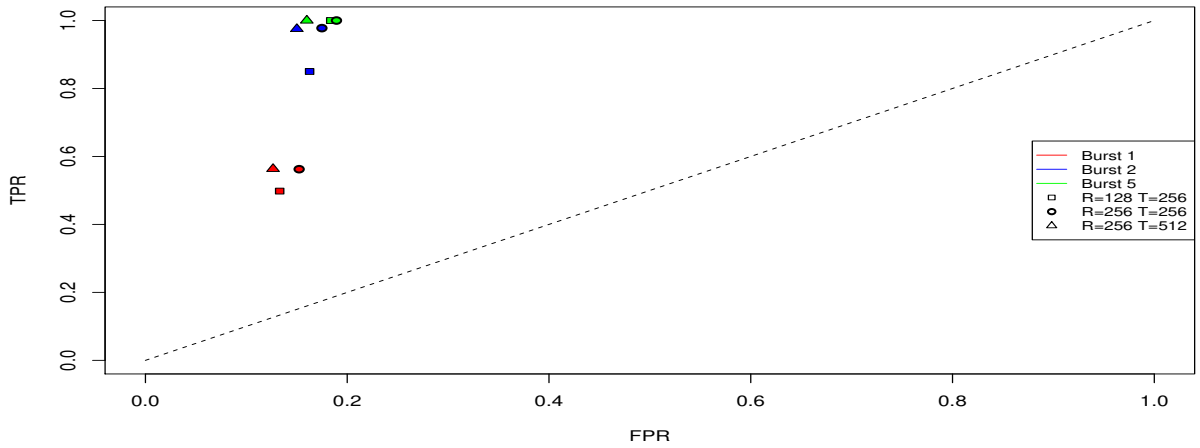


Figure 4: ROC plot for simulation 1B.

Binary classification measures

R	T	TNR	sd_{TNR}	ACC
128	256	0.9918	0.0078	0.9918
256	256	0.9941	0.0058	0.9941
256	512	0.9921	0.0049	0.9921

Table 2: Binary classification measures obtained over 100 runs for simulation 1A.

Simulation 2 places a squared cosine behaviour in all spectral replicates except for the middle $R/4$, which have a burst (with values of 1, 2 or 5) across the middle $T/4$ locations (S2A), and then we further dramatically reduce the number of replicates and locations defined to have changing spectral characteristics (S2B). It is for these locations that we would expect to see rejections of the null hypothesis of constancy across the replicate dimension. Examples of their true spectra and realisations appear in Figures 5, and 17 and 18 (Appendix C) for $R = 256$ and $T = 512$. Further mathematical details are given in Appendix C, which also illustrates the test behaviour across bursts of various sizes by means of the location-specific empirical power and size estimates, see Figures 19 and 20 and associated discussion. These highlight that locations marking the transition into and out of the burst are associated to incorrect test outcomes, as indicated by their empirical sizes and powers.

The statistical measures in Table 5 for simulation 2A further confirm the impact of increasing the burst, with the TPR increasing and TNR decreasing. Whilst other measures remain favourable, we do note the increase in the FDR

Binary classification measures

R	T	TPR	sd_{TPR}	TNR	sd_{TNR}	FDR	FOR	F-M	ACC	PT
128	256	0.4980	0.0327	0.8591	0.0185	0.2107	0.3685	0.6267	0.6786	0.3467
256	256	0.5625	0.0282	0.8398	0.0167	0.2130	0.3422	0.6652	0.7012	0.3476
256	512	0.5634	0.0223	0.8696	0.0139	0.1833	0.3341	0.6782	0.7165	0.3244

Binary classification measures

R	T	TPR	sd_{TPR}	TNR	sd_{TNR}	FDR	FOR	F-M	ACC	PT
128	256	0.85	0.0571	0.8295	0.0177	0.1610	0.1503	0.8441	0.8397	0.3092
256	256	0.9777	0.0218	0.8173	0.0102	0.1518	0.0260	0.9106	0.8975	0.3017
256	512	0.9752	0.0197	0.8460	0.0095	0.1333	0.0280	0.9193	0.9106	0.2842

Binary classification measures

R	T	TPR	sd_{TPR}	TNR	sd_{TNR}	FDR	FOR	F-M	ACC	PT
128	256	1	0	0.8085	0.0117	0.1551	0	0.9192	0.9043	0.3042
256	256	1	0	0.8028	0.0088	0.1592	0	0.9170	0.9014	0.3074
256	512	1	0	0.8361	0.0070	0.1379	0	0.9285	0.9181	0.2881

Table 3: Binary classification measures obtained over 100 runs for simulation 1B. *Top*: 'burst'= 1; *middle*: 'burst'= 2; *bottom*: 'burst'= 5. Prevalence = 0.5.

Binary classification measures

R	T	TPR	sd_{TPR}	TNR	sd_{TNR}	FDR	FOR	F-M	ACC	PT
128	256	0.9013	0.0540	0.9775	0.0140	0.0522	0.0322	0.9235	0.9584	0.1297
256	256	0.9842	0.0222	0.9688	0.0124	0.0722	0.0053	0.9553	0.9726	0.1477
256	512	0.9769	0.0230	0.9799	0.0078	0.0505	0.0077	0.9629	0.9791	0.1227

Table 4: Binary classification measures obtained over 100 runs for simulation 1C for 'burst'= 2. Prevalence = 0.25.

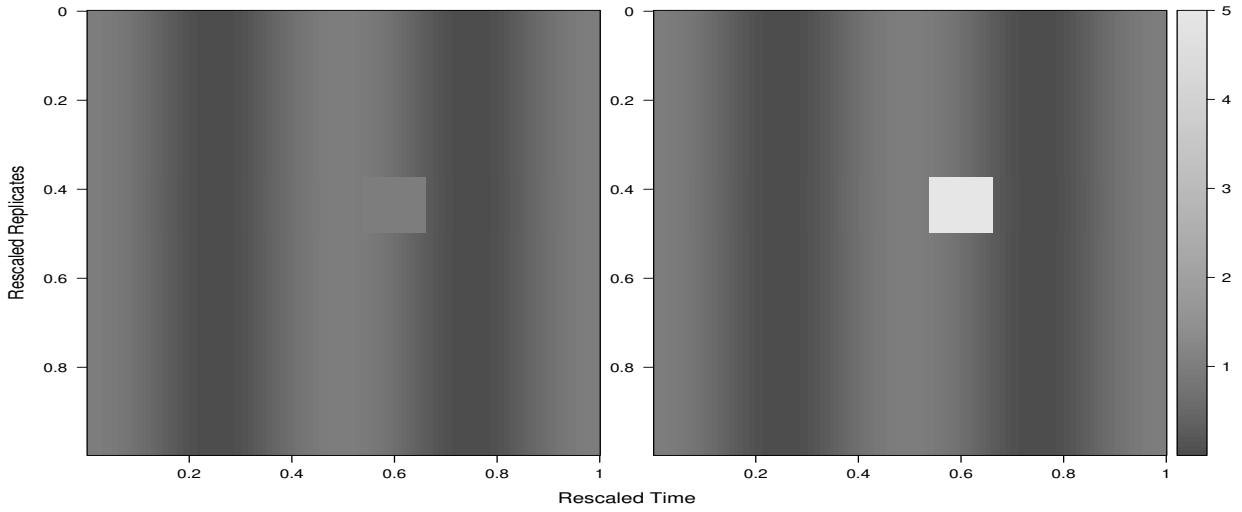


Figure 5: Time-replicate plots of the true spectra in level 5 for simulation 2B. *Left*: 'burst'=1; *right*: 'burst'=5.

corresponding to a burst jump from 2 to 5, such that approximately 30% of identified positives are false. This increase moves with the decrease in TNR. However, keeping in mind the goal of the test, the increase in TPR suggests all true

positives are identified at a burst of 5 and despite the increase in FDR, the higher F-M index values tell us that there is greater similarity between the identified positives and true positives clusters. Note that the presence of 'na' values is a result of the test failing to make any positive identifications on some runs of the simulation (see how the formulas in Table 1 break down if $tp = fp = 0$). Figure 6 (left) displays the corresponding ROC plot, which gives a visual comparison of the TPR and FPR for simulations with differing R, T and bursts. The measures shown in Table 6 for simulation 2B also improve for higher values of R and T, and furthermore as the burst increases. Recalling that we have defined an evolution in the spectra over a much smaller window of replicates and locations, it is understandable that the measures for bursts 2 and 5 are slightly lower than in previous simulations, however still indicate that the test performs well despite the more challenging structure. The ROC plot for this simulation appears in Figure 6 (right), leading to similar conclusions.

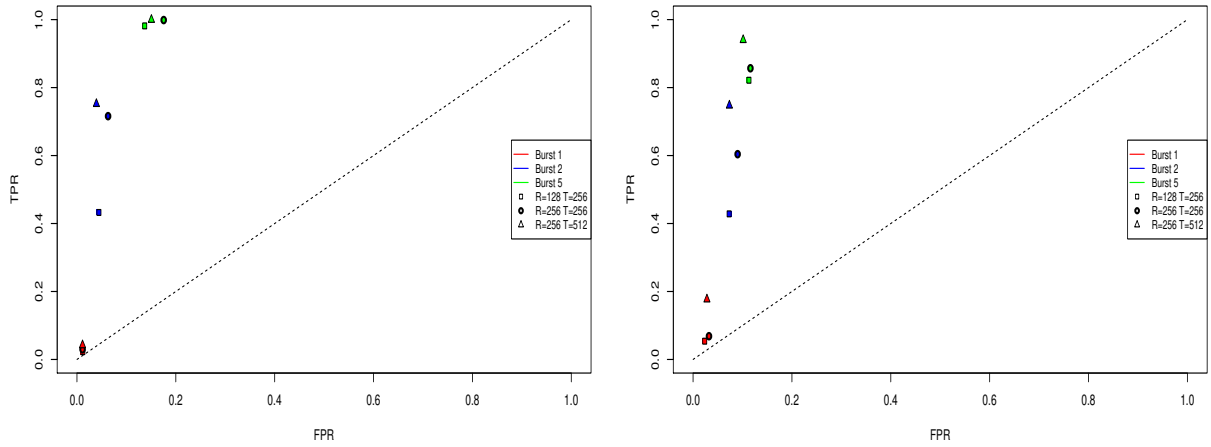


Figure 6: ROC plots for simulation 2A (left) and 2B (right).

Simulation 3 evaluates the power of the test by means of a setup that displays slowly evolving spectral characteristics over both (rescaled) time and replicates of a MULT-LSW process, with structure present at one (S3A) or two (S3B) scales. Their precise mathematical definitions appear in Appendix C, and Figures 21 (Appendix C) and 7 (below) respectively, represent their true spectra when $R = 256$ and $T = 512$. For this simulation all locations have a degree of varying spectra over the replicates and therefore we would like the test to reject all locations. This is the opposite scenario to Simulation 1A. The location-specific empirical test powers can be visualised in Figures 22 and 23, showing that the tests scale well with increasing sample size as the spectral estimates improve asymptotically with R, $T \to \infty$ (see Appendix C for a detailed discussion).

Furthermore, in Table 7 we report the statistical measures on the test as a whole for simulation 3A, and in Table 8 for simulation 3B, while further results appear in Appendix C. Note that there are fewer rates reported due to the absence of true negatives (locations with constant spectra across replicates). The results show how the increase in replicates much improves the TPR and the F-M index, which we recall gives a measure of the similarity between the identified positives and true positives. Both the TPR and F-M demonstrate the test performs well.

Concluding remarks for the location-specific testing methodology. Overall, the simulation study has shown that the location-specific test correctly identifies whether a replicate-effect exists in most settings, except when the spectra evolves too slowly over the trials. We recommend the use of this test if the analysis aim is to gain an indication on whether there is any evolution in the spectral characteristics over the replicate-dimension. However, this test goes further and is also able to indicate the times and trials at which an evolutionary behaviour over the spectra exists. Whilst the location-specific test is most informative, the next section employs the global test for replicate-effect, offering the user a direct diagnostic as opposed to assessing the individual locations.

Binary classification measures

R	T	TPR	sd_{TPR}	TNR	sd_{TNR}	FDR	FOR	F-M	ACC	PT
128	256	0.0230	0.0277	0.9828	0.0145	na	0.2489	na	0.7429	na
256	256	0.03	0.0315	0.9835	0.0127	na	0.2474	na	0.7451	na
256	512	0.0420	0.0238	0.9861	0.0075	0.4443	0.2446	0.1492	0.7500	na
512	512	0.0946	0.0382	0.9817	0.0093	0.3279	0.2351	0.2477	0.7599	na

Binary classification measures

R	T	TPR	sd_{TPR}	TNR	sd_{TNR}	FDR	FOR	F-M	ACC	PT
128	256	0.4325	0.1160	0.9503	0.0258	0.2318	0.1651	0.5721	0.8209	0.2485
256	256	0.7159	0.0979	0.9318	0.0233	0.2071	0.0914	0.7519	0.8779	0.2334
256	512	0.7520	0.0667	0.9581	0.0139	0.1341	0.0790	0.8060	0.9066	0.1884
512	512	0.9627	0.0222	0.9228	0.0233	0.1857	0.0132	0.8849	0.9328	0.2177

Binary classification measures

R	T	TPR	sd_{TPR}	TNR	sd_{TNR}	FDR	FOR	F-M	ACC	PT
128	256	0.9816	0.0233	0.8577	0.0322	0.2919	0.0071	0.8331	0.8887	0.2736
256	256	0.9988	0.0057	0.8193	0.0117	0.3448	0.0005	0.8089	0.8642	0.2982
256	512	0.9995	0.0019	0.8468	0.0176	0.3104	0.0002	0.8301	0.8850	0.2808

Table 5: Binary classification measures obtained over 100 runs for simulation 2A. *Top*: 'burst'= 1; *middle*: 'burst'= 2; *bottom*: 'burst'= 5. Prevalence = 0.25. The 'na' values are a result of the test failing to make any positive identifications (such that $tp = fp = 0$ in Table 1) on some runs of the simulation.

Binary classification measures

R	T	TPR	sd_{TPR}	TNR	sd_{TNR}	FDR	FOR	F-M	ACC	PT
128	256	0.0538	0.0665	0.972	0.0185	na	0.1220	na	0.8572	na
256	256	0.0684	0.0632	0.9633	0.0177	0.7636	0.1213	0.1229	0.8515	na
256	512	0.1769	0.0664	0.9697	0.0130	0.5180	0.1081	0.2879	0.8706	0.2943

Binary classification measures

R	T	TPR	sd_{TPR}	TNR	sd_{TNR}	FDR	FOR	F-M	ACC	PT
128	256	0.4284	0.1293	0.9226	0.0235	0.5398	0.0810	0.4408	0.8609	0.2997
256	256	0.6041	0.0901	0.9054	0.0186	0.5077	0.0587	0.5442	0.8677	0.2833
256	512	0.7473	0.0703	0.9243	0.0152	0.4045	0.0375	0.6662	0.9022	0.2405

Binary classification measures

R	T	TPR	sd_{TPR}	TNR	sd_{TNR}	FDR	FOR	F-M	ACC	PT
128	256	0.8216	0.0629	0.8825	0.0148	0.4894	0.0280	0.6472	0.8748	0.2742
256	256	0.8569	0.0543	0.8795	0.0093	0.4864	0.0227	0.6632	0.8767	0.2727
256	512	0.9402	0.0294	0.8964	0.0086	0.4295	0.0094	0.7322	0.9019	0.2490

Table 6: Binary classification measures obtained over 100 runs for simulation 2B. *Top*: 'burst'= 1; *middle*: 'burst'= 2; *bottom*: 'burst'= 5. Prevalence = 0.125. The 'na' values are a result of the test failing to make any positive identifications (such that $tp = fp = 0$ in Table 1) on some runs of the simulation.

3.2. Simulations for the global replicate-effect testing methodology

Our attention now turns to investigating the performance of our global testing approach. For various R and T, we carry out the proposed testing procedure with smoothing performed over both the replicate and time domains, shown to yield the most accurate estimator of the true, unknown replicate evolutionary wavelet spectrum (Embleton *et al.*, 2022).

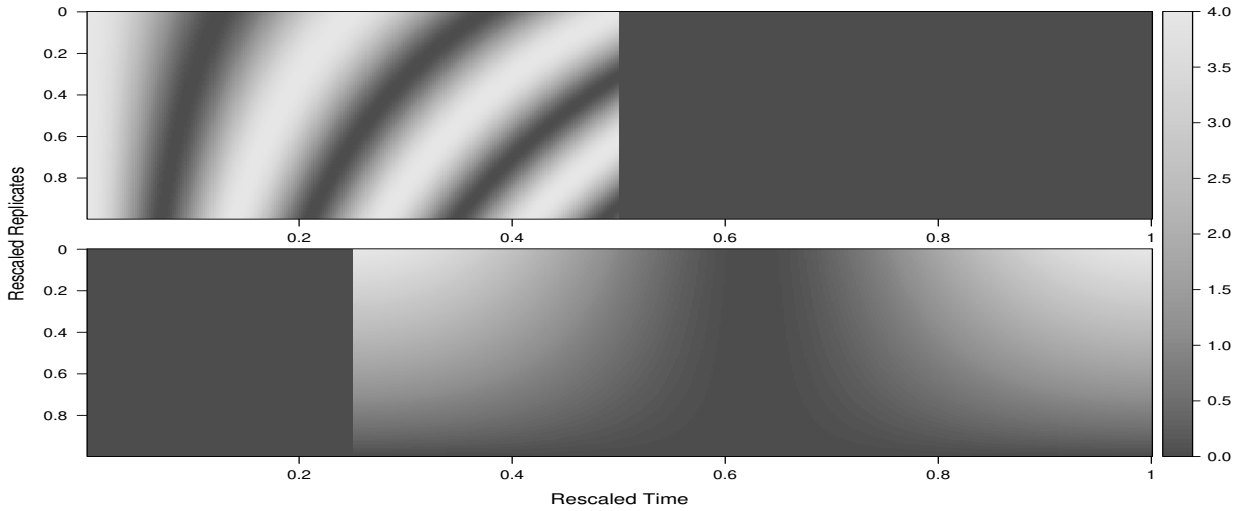


Figure 7: Time-replicate plots of the true spectra in levels 7 (*top*) and 6 (*bottom*) for simulation 3B.

Binary classification measures

R	T	TPR	sd_{TPR}	F-M	ACC
128	256	0.4944	0.0716	0.7012	0.4944
256	256	0.8470	0.0375	0.9201	0.8470
256	512	0.8954	0.0226	0.9462	0.8954

Table 7: Binary classification measures obtained over 100 runs for simulation 3A.

Binary classification measures

R	T	TPR	sd_{TPR}	F-M	ACC
128	256	0.4916	0.0568	0.6999	0.4916
256	256	0.7980	0.0207	0.8933	0.7980
256	512	0.8042	0.0162	0.8967	0.8042

Table 8: Binary classification measures obtained over 100 runs for simulation 3B.

In what follows, we use $B = 100$ and a blanket value of $M = 4$ but note that the quality of the spectral estimation, and thus our test performance, would improve with a choice for M that would grow with R, at approximately $M = \frac{3}{4}\sqrt{R}$. For each simulation setup we report empirical size and power estimates obtained through counting the number of times the test rejects the null hypothesis of constancy (over replicates) at a significance level of 5%.

Simulations to investigate size. We first aim to assess how the test performs in the absence of any spectral evolution across the replicates. To do so we consider the following setups that purposely include processes that depart from the MULT-LSW assumptions, in order to probe the test robustness to model misspecification.

s1. Simulated processes as detailed in Simulation 1A of Section 3.1.

s2. (a) An autoregressive AR(1) process, $X_t^{r;R} = \gamma X_{t-1}^{r;R} + \epsilon_t^r$, with AR parameter $\gamma = 0.9$ and independent standard normal innovations.

(b) As above but with AR parameter $\gamma = -0.3$.

s3. (a) A time-varying autoregressive AR(1) process, $X_t^{r;R} = \gamma_t X_{t-1}^{r;R} + \epsilon_t^r$, with AR parameter γ_t evolving linearly within a trial from 0.9 to -0.9, and independent standard normal innovations.

(b) As above but with AR parameter γ_t evolving constantly from 0.3 to -0.3.

Results. We compute our spectral estimates using the discrete non-decimated wavelets built by means of Daubechies Least Asymmetric wavelet family with 10 vanishing moments (Daubechies, 1992) for s1, whilst Haar wavelets were chosen for s2(a,b) and s3(a,b). Unlike for the location-specific test where Haar wavelets alone were suitable for estimation, here a choice of wavelet family most reflective of the process smoothness might be more appropriate. However, Embleton *et al.* (2022) comment on the robustness of their estimation results to the choice of wavelet, hence the choices in our simulations here. Table 9 reports the resulting empirical size estimates. For the majority of the simulations, the size estimates are less than the 5% nominal size and thus indicate that the test does a good job of not incorrectly rejecting the hypothesis that spectra are constant across the replicate-dimension. The one exception is simulation s3(a), for which the size values range from 27% to 60%.

		Size				
R	T	s1	$s2_a$	$s2_b$	$s3_a$	$s3_b$
64	128	0	0.01	0.03	0.27	0.05
128	128	0	0.03	0.04	0.33	0.08
128	256	0	0	0.02	0.4	0
256	256	0	0.01	0.01	0.6	0.01
256	512	0	0.01	0	0.6	0

Table 9: Empirical size values computed for simulations s1 - s3(a,b). Spectral estimation was via the MULT-LSW model with smoothing over replicates and time.

It is not obviously clear why the global test is struggling with simulation s3(a), however a similar scenario occurs for the simulations using the Haar wavelet based test of stationarity proposed by Nason (2013) on a single AR(1) process with an AR parameter of -0.9. Nason (2012) investigated this scenario and drew attention to the possible 'volatility clustering' exhibited by the process (see Figure 24 in Appendix D) and comparable to typical financial time series behaviour (see e.g. Tsay (2010)). For a series of length $T = 512$, Nason (2012) found the empirical size decreasing approximately 20% with decreasing the AR parameter to -0.8, down to an empirical size around 4%.

An improvement in the empirical size values can also be obtained in our setup through decreasing the negativeness of the AR parameter. When the AR parameter γ_t is defined to evolve linearly from 0.9 to -0.9, for $R = 128$ and $T = 256$ the empirical size in Table 9 is shown to be 40%. Setting the AR parameter to evolve linearly from 0.9 to -0.8 yields an empirical size value of 18%, further still reducing if the parameter is set to evolve linearly from 0.9 to -0.7, down to an empirical size of 12%. The additional estimated size values have not fallen below 5%, however they do demonstrate that the test dramatically improves when we set the AR parameter to evolve to a lesser extreme than -0.9, analogous to the results found in Nason (2012).

Simulations to investigate power. To examine how well our global test identifies a breach in spectral constancy over the trials, or in other words whether there exists spectral evolution over the trials, we consider the following setups which include processes that (i) adhere to (p1–p3) and (ii) fall outside of (p4-p5) MULT-LSW assumptions.

p1. Simulated processes as detailed in Simulation 1B of Section 3.1 with burst values of 1 (b1) and 2 (b2) .

p2. Simulated processes as detailed in Simulation 3A of Section 3.1.

p3. Simulated processes as detailed in Simulation 3B of Section 3.1.

p4. (a) A 'replicate-varying' autoregressive AR(1) process, $X_t^{r;R} = \gamma^r X_{t-1}^{r;R} + \epsilon_t^r$, with AR parameter γ^r evolving constantly from 0.9 to -0.9, and independent standard normal innovations.

(b) As above but with AR parameter γ^r evolving linearly from 0.3 to -0.3.

p5. (a) A 'time–replicate-varying' autoregressive AR(1) process, $X_t^{r;R} = \gamma_t^r X_{t-1}^{r;R} + \epsilon_t^r$, with independent standard normal innovations and the AR parameter γ_t^r evolving linearly across both time- and replicate-dimensions, as described in Appendix D.

(b) As above but with values 0.9 and -0.9 being replaced with 0.3 and -0.3, respectively.

		Power							
R	T	$p1_{b1}$	$p1_{b2}$	p2	p3	$p4_a$	$p4_b$	$p5_a$	$p5_b$
64	128	0.05	0.38	0.97	1	1	0.97	1	0.3
128	128	0.08	0.97	1	1	1	1	1	0.76
128	256	0.23	1	1	1	1	1	1	0.97
256	256	0.79	1	1	1	1	1	1	1
256	512	0.99	1	1	1	1	1	1	1

Table 10: Empirical power values computed for simulations p1 - p5(a,b). Spectral estimation was via the MULT-LSW model with smoothing over replicates and time.

Results. In the same vein as for the size simulations, we compute our spectral estimates using the discrete non-decimated wavelets built by means of Daubechies Least Asymmetric wavelet family with 10 vanishing moments for p2, and with 6 vanishing moments for p3, whilst Haar wavelets were chosen for p1(b1,b2), p4(a,b) and p5(a,b). In Table 10 we report the empirical power estimates, which suggest the test performs well for most of our setups. The setup p1 is the only concern when R is low, however this was not unexpected considering the previous results of the location-specific test for the same setup (see Simulation 1B in Section 3.1). We recall that there the test broke down when the difference between spectra across the replicates assumed values less than 1. However, the power values do improve as our sample size increases ($R, T \to \infty$).

Concluding remarks for the global testing methodology. The simulation study has demonstrated that the global test, through encapsulating the spectral characteristics of the whole MULT-LSW process, is capable of answering the general question of 'does a replicate-effect exist'. This is to be understood in the sense of assessing whether there is enough evidence to establish the existence of any departures from the overall spectral constancy across trials, and we identified a test sensitivity when the process exhibits very slowly varying amplitudes. Akin to Fryzlewicz and Ombao (2009), an option might be to perform the test on a suitably selected subset of the wavelet scales, times and replicates, hence even though the global test performs well already, there is room for improvement.

Both the global and location-specific tests have proven to be successful within the simulation studies and the next section will illustrate the tests when applied to real data.

4. Is there a trial-effect along the macaque local field potentials?

Our simulation studies have given a thorough demonstration on the performance of both the location-specific and global tests aiming to identify the existence of replicate spectral evolution. We now perform both tests on the macaque brain processes data introduced in Section 1 and further described in Appendix A. We will carry out our analysis over the set of trials grouped chronologically for correct responses. Each trial was of length $T = 2048$ (hence $J = 11$) and the correct responses consisted of $R = 256$ (hence $J' = 8$) trials, thus forming our replicate experimental time series that we model using the MULT-LSW framework of Embleton *et al.* (2022).

4.1. Testing for a replicate-effect along the macaque hippocampal (Hc) response

We estimate the spectral characteristics of the hippocampus (Hc) correct trials by means of non-decimated discrete wavelets from Daubechies Least Asymmetric family with 10 vanishing moments. The spectral estimates as functions of (rescaled) time and replicates are captured visually in Figure 8 (bottom row), where for comparison we show the resulting estimators obtained when averaging the hippocampal response across the entire experiment (top row). The question we are aiming to answer is whether or not there is sufficient evidence to deem that spectral evolution across

the trials (replicates) is indeed manifest and thus the macaque's memory recall evolves through the experiment. We will apply both the location-specific and global tests, and if appropriate, identify the times and trials for which the memory recall becomes manifest.

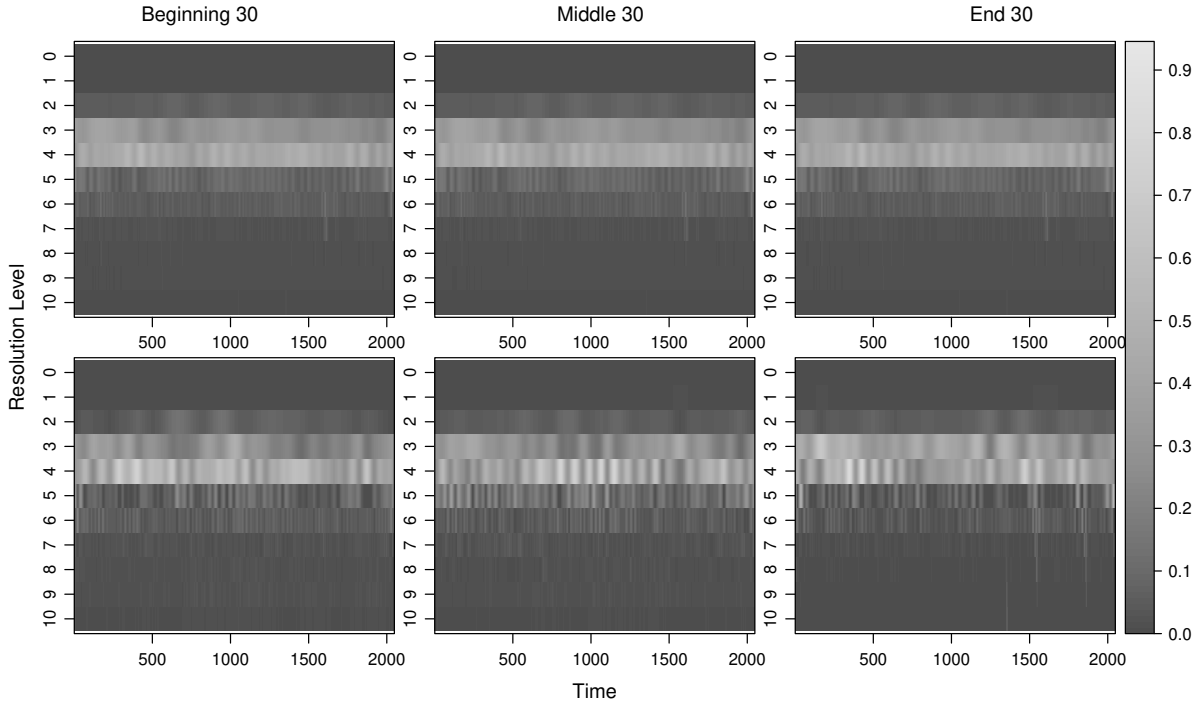


Figure 8: Time-scale hippocampus (Hc) plots for the correct trials. Spectral estimates are shown for the average over 30 replicates in the beginning, middle and end of the experiment. *Top*: estimates from the LSW method averaged over all replicates; *Bottom*: MULT-LSW method with smoothing over time and replicates. Note the process evolution through trials is highlighted by the MULT-LSW representation, rather than averaged out. The resolution level notation here is $j := J - j$ for the jth finest scale.

Location-specific test

For each location, using the settings in Section 3.1, there are a total of 40 hypothesis tests (also recall equation (9)), over 8 wavelet scales and 5 Haar scales. We will deal with this multiple-testing aspect by reporting results that use the false discovery rate (FDR) procedure.

In plot (a) of Figure 25 in Appendix E we present a visualisation of the number of locations identified by the test as rejecting the null hypothesis of spectral constancy over the trials. The total number of times rejected was 1153. Here, a location is classified as rejected if any one of the hypothesis tests for that location is rejected. This approach is a little naive and at most tells us that our test 'successfully' identifies numerous locations across the experiment that exhibit spectral evolution over the trials. What qualifies as a 'successful' identification is open to interpretation, for instance whether 1 rejection out of 40 is satisfactory to classify a location as having a significant replicate-effect. Thus, plot (c) of Figure 25 gives a more informative visualisation through displaying the percentage of the 40 hypothesis tests that were rejected for each location, while in plot (e) we threshold the percentage values at 25%. By considering the percentage of tests rejected we can identify the locations that give a strong indication of a replicate-effect, which could then be the focus for further analysis. Plot (c) indeed corroborates the evolutionary patterns we see in the spectral estimates in Figure 8. Interestingly, the test also highlights locations around the timepoint 512, which we recall corresponds to the visual exposure time block, and in the final quarter time block corresponding to the macaque making correct associations. Thus the test appears to confirm the experimental design. However, we reinforce our previous point on the flexibility of choosing a threshold that would decide whether a location should be deemed as 'successfully' rejecting the null hypothesis and currently this is left to the discretion of the user.

It is not unreasonable to assume that if a significant replicate-effect exists for one location, then the spectral characteristics of neighbouring locations may echo a similar replicate-effect. As such, in plots (b), (d) and (f) of Figure 25 we choose to only display the locations rejected that are in the vicinity of the two rejected location on each side. This process of eliminating the weakly rejected locations gives a way to identify the locations with the strongest evidence for the existence of a replicate-effect. As well as identifying the locations where a replicate-effect exists, we can also observe the window of trials for which the replicate-effect was detected. In Figure 9, for specific time locations 560 (left) and 1610 (right), indicative of the experimental blocks corresponding to picture exposure and exercised choice respectively, we plot the observed data as a time series across the replicate domain. The double headed arrows indicate the replicates for which evolutionary spectra were detected, with the width determined by the support of the underlying Haar wavelet. The right-hand axis gives the wavelet scale j of the spectral estimates tested and the vertical position of the arrow within each wavelet scale indicates the Haar scale i with the topmost arrow corresponding to the finest of the Haar scales tested. Both locations 560 and 1610 have over 50% of hypothesis tests rejected under the FDR control, with the majority of rejections occurring over the final 128 replicates, indicating memory recall activation towards the end of the experiment. Many rejections occur for a narrower window of 16 replicates ($256/2^4$, where 4 is the finest Haar scale tested) at the end of the replicates for location 560 and around replicate 135 for location 1610.

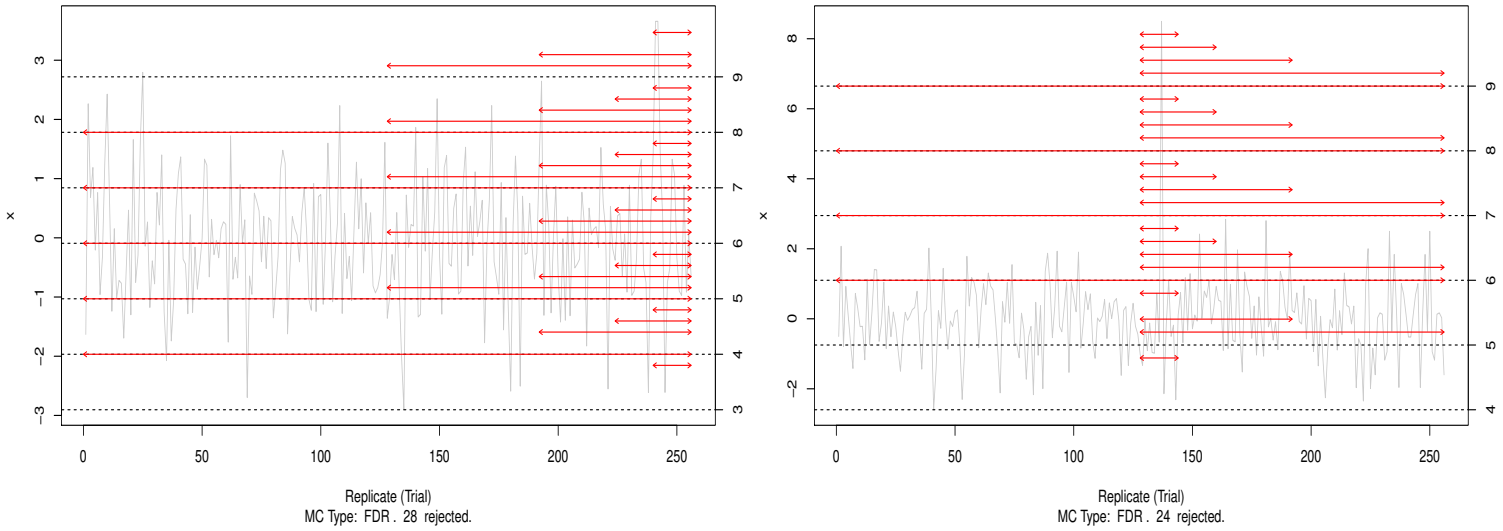


Figure 9: Replicate domain plots for correct hippocampus trials at times 560 (*left*) and 1610 (*right*). Trials where replicate-effect detections were made are indicated by the double headed arrows, and the corresponding wavelet scales $j = 3, ..., 9$ are indicated on the right axis. Within each wavelet scale, the vertical position of the arrow indicates in ascending order the Haar scales $i = 0, ..., 4$.

Global test

To test for a global replicate-effect, we bootstrap the process under the null hypothesis that assumes spectral constancy over the replicates, $\bar{\hat{S}}_j(\frac{k}{T})$, and estimate the corresponding process spectra. The histogram of the bootstrap test statistics appears in Figure 10. We compute the test p-value by means of step 4 in the `BootReplicateTest` algorithm 1 and obtain a p-value < 0.01, giving a strong indication that a replicate-effect exists, in agreement to the findings of the location-specific tests.

4.2. Testing for a replicate-effect along the nucleus accumbens (NAc) response

We estimate the spectral characteristics of the NAc correct trials by means of non-decimated discrete wavelets from Daubechies Least Asymmetric family with 6 vanishing moments, captured visually in Figure 13 (top row). We are aiming to assess whether or not the macaque has learned to process the reward as the experiment progresses. Moreover,

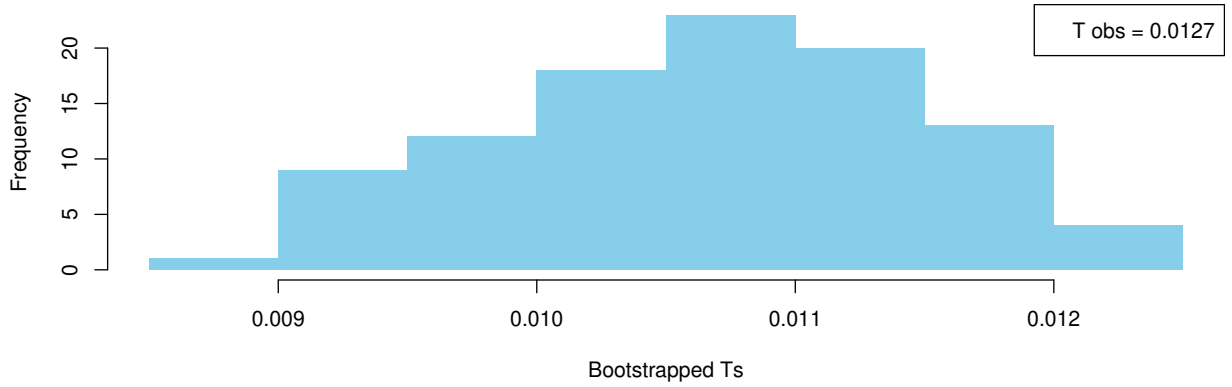


Figure 10: Histogram of the bootstrap test statistics for the global replicate-effect test carried out on the correct trials of the hippocampus (Hc) dataset.

if the macaque has indeed learned, we would like to identify the specific trial(s) over which the evidence of learning is manifest.

Location-specific test

Similar to the location-specific tests for the hippocampus, for each location there are a total of 40 hypothesis tests and we tackle the multiple-hypothesis testing by using the FDR procedure. The test identified 344 locations as breaking the null hypothesis of spectral constancy over the replicates and we plot these locations in plot (a) of Figure 26, Appendix E. Once again, we learn that several locations were identified by the test and whilst this is important, further information is gained by considering the percentage of the 40 hypothesis tests per location visualised in plot (c). Evidently, there are far fewer locations with a higher number of rejections than 1 in 40. It appears that for the NAc, whilst the test gives evidence of a replicate-effect for some locations, the potential evolution of the spectra across the replicates is not as substantial as was identified for the Hc. We observe that locations around timepoint 512, the start of the second experimental time block, are also identified by the test for the NAc, which is interesting as this observation is not as clear in the spectral estimates of Figure 13, top row. This demonstrates how our location-specific test can be utilised as an essential tool to screen the locations that are 'successfully' (decided through interpretation of the percentage of hypothesis tests rejected) rejected by the test as displaying a potential replicate-effect. We can then carry out further analysis on the spectral characteristics at these locations, instead of simply plotting and comparing the within-trial evolutionary wavelet spectra. To gain further clarity on which rejected locations give the strongest evidence for the existence of spectral evolution over the replicates, we eliminate weakly rejected locations and display the results through plots (a)-(f) in Figure 26 in Appendix E. Quite clearly, the impact of thresholding the rejected locations drastically narrows down the locations of most interest. In fact, plot (f) indicates that there was no local window of (5) neighbouring locations with 25% or more significant Haar wavelet coefficients (rejected hypothesis tests). Plots in Figure 11 give an indication of the replicates for which an evolution in the wavelet spectra was detected for specific time locations 565 (left) and 1865 (right), indicative of the visual exposure and task times of the experiment, respectively. A description for the plots' construction was previously given in the discussion for the hippocampus. Both locations 565 and 1865 have under 25% of hypothesis tests rejected under the FDR control. Of these, most rejections occur over the final 128 replicates with few rejections occurring for 16 replicates ($256/2^4$, where 4 is the finest Haar scale tested) at the end of the replicates for location 565 and around replicate 220 for location 1865.

Global test

We carry out the global test on the NAc dataset and following algorithm 1 we obtain a p-value of 1, hence not enough evidence is present to deem a significant replicate-effect exists along the macaque's NAc response across the

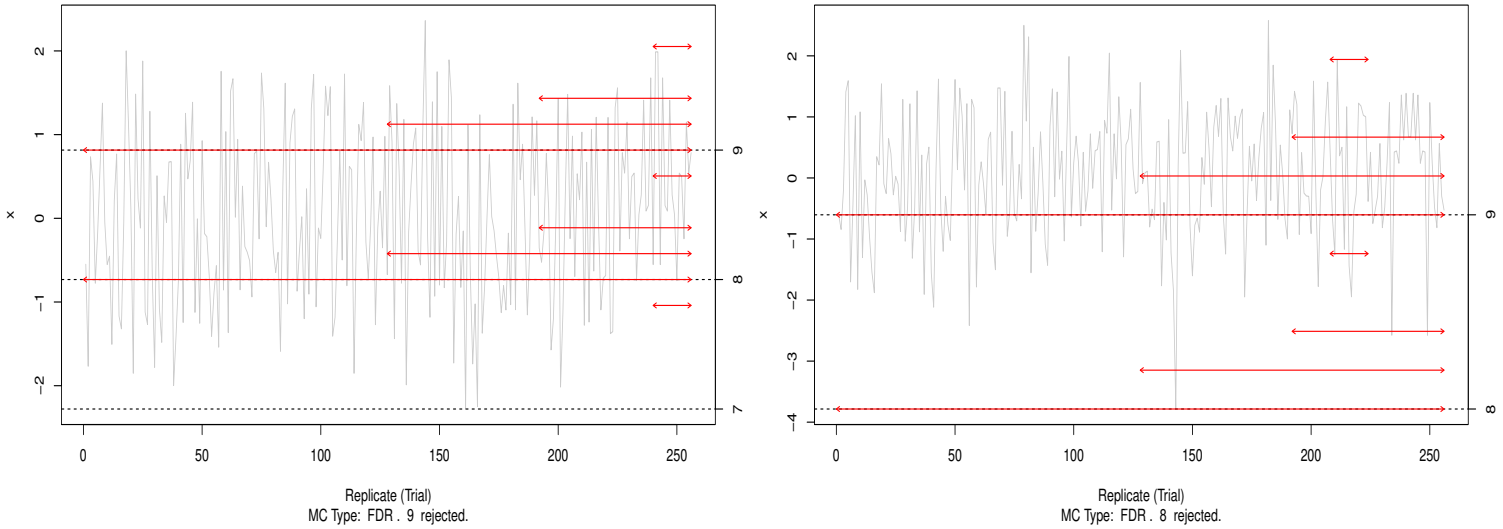


Figure 11: Replicate domain plots for correct nucleus accumbens trials at times 565 (*left*) and 1865 (*right*). Trials where replicate-effect detections were made are indicated by the double headed arrows, and the corresponding wavelet scales $j = 7, 8, 9$ are indicated on the right axis. Within each wavelet scale, the vertical position of the arrow indicates in ascending order the Haar scales $i = 0, ..., 4$.

experiment. We plot the histogram of the bootstrap test statistics in Figure 12, highlighting these are larger than the observed test statistics. This indicates that we do not reject the null hypothesis of no replicate-effect since the real data displays even less variability than realistic data simulated under the null hypothesis.

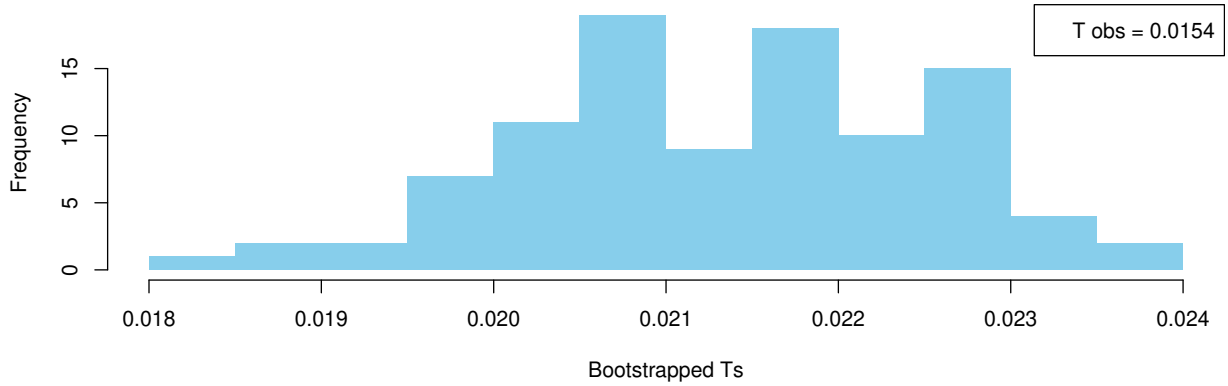


Figure 12: Histogram of the bootstrap test statistics for the global replicate-effect test carried out on the correct trials of the nucleus accumbens (NAc) dataset.

To attempt to understand why the test does not identify a global replicate-effect, in Figure 13 we examine the spectral estimates of the correct NAc trials, the averaged (over replicates) NAc spectral estimates $\bar{\hat{S}}_j(\frac{k}{T})$, and the spectral estimates computed for one bootstrapped process. The test statistics that we compute give a measure for the difference between the averaged spectral estimates (middle row) and; (for $\mathbf{T}^{obs}_{ave}$) the spectral estimates (top row); (for

$\mathbf{T}_{ave}^{(b)}$) the bootstrap spectral estimates (bottom row). Visually, it is clear how $\mathbf{T}_{ave}^{(b)}$ computed for the estimated spectra of one bootstrap process (displayed in Figure 13) is larger than $\mathbf{T}_{ave}^{obs}$. This observation thus supports the outcome of not enough evidence being present to suggest a global replicate-effect exists for the correct trials of the NAc. Note the averaged bootstrap spectral estimates (simulated under the null hypothesis) are similar both to their 'truth' and original NAc estimates (see Figure 27 in Appendix E, bottom row versus middle and top rows), thus leading us towards two points to discuss.

Firstly, recall that the simulation study for both local and global tests highlighted a weaker test performance when the spectral evolution across the replicates is very slow (e.g. simulation 1B in Section 3.1 and simulation p1 (b1) in Section 3.2). Similarly, the Haar wavelet based test of stationarity of Nason (2012, 2013) struggled to identify nonstationarity for their models P2 and P3 which both contained wavelet spectra defined to evolve very slowly over time. This is likely the situation here, as displayed in Figure 13 (top row). Secondly, recall that both the localised and global tests, are developed under the assumption of uncorrelated replicates. However, the NAc analysis of Embleton *et al.* (2022) uncovered the existence of a potential coherence along the replicate dimension, thus this dataset is likely evading this underpinning assumption.

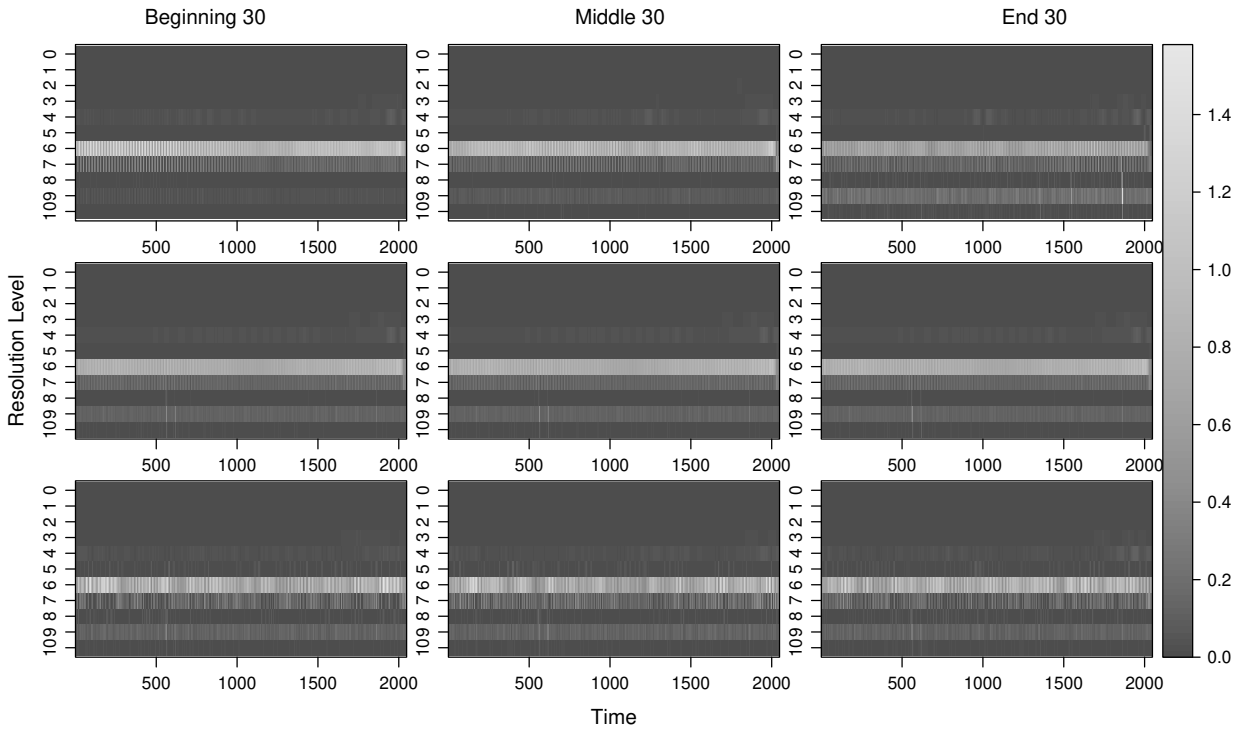


Figure 13: Time-scale nucleus accumbens (NAc) plots for the correct trials computed for the global test of replicate-effect. Spectral estimates are shown for the average over 30 replicates in the beginning, middle and end of the experiment. *Top*: spectral estimates of the correct NAc trials; *Middle*: NAc spectral estimates averaged across all trials; *Bottom*: spectral estimates corresponding to a simulated bootstrap process.

Concluding remarks on the analysis of local field potentials. We have applied both the location-specific and global tests on the correct trials of the macaque hippocampus (Hc) and nucleus accumbens (NAc) datasets. We note that both our tests were constructed under the assumption of uncorrelated trials, an assumption which appears reasonable for the Hc data but perhaps less so for the NAc data (Embleton *et al.*, 2022). For the global test, inclusion of an estimated coherence structure between trials could improve the bootstrap simulation and estimation, therefore an important next step in the methodology would be to develop novel tests that incorporate between-trial correlation.

5. Concluding remarks

In this work we have proposed two wavelet-based tests for the existence of a replicate-effect along a multiple-trial experiment consisting of time series that feature nonstationarities across both time and replicate-dimensions. The replicate-effect is to be understood as a measure of the departure from constancy of the process spectral characteristics across the trials, framed here by means of a MULT-LSW meta-process representation. Through simulation studies and an application to real data from the neuroscience, both tests have been shown to perform successfully, with weaker performance identified for processes with very slowly evolving amplitudes.

So, which test should we use? From the results a natural order presents itself, with the global test offering a way to determine whether over all wavelet scales and times, a significant replicate-effect exists. Hence our recommendation is that the global test could be used as a first step in the analysis of the characteristics of multiple-trial time series modelled as a MULT-LSW process. Its purpose could simply be a clarification tool. The location-specific test would then provide a more in-depth assessment through carrying out multiple hypothesis tests of replicate-effect for each location. Informed by the application context, the user can determine which locations are significantly rejected by the test through assessing the percentage of the multiple hypothesis test per location that reject the null hypothesis of spectral constancy over the replicates.

Crucial to the proposed tests of replicate-effect is the assumption of uncorrelated replicates, hence naturally a next step would be to develop tests for replicate-effect that incorporate the potential for dependence between replicates. Different future directions would be to gain a further understanding into how many significant Haar wavelet coefficients (multiple hypothesis test rejections) would be sufficient to deem a location as significantly rejected by the location-specific test, or to explore testing encompassing one-sided alternatives.

Acknowledgements

J. Embleton's research was supported by EPSRC studentship EP/N509802/1.

References

Abela, A., Duan, Y. and Chudasama, Y. (2015). Hippocampal interplay with the nucleus accumbens is critical for decisions about time. *Eur. J. Neuroscience* , 42, 2224–2233.

Ahamada, I. and Boutahar, M. (2002). Tests for covariance stationarity and white noise, with an applicationto Euro/US dollar exchange rate: an approach based on the evolutionary spectral density. *Econ. Lett.*, 77, 177–186.

Atkinson, A. D., Hill, R. R., Pignatiello, J. J. Jr., Vining, G. G., White, E. D., and Chicken, E. (2017). Wavelet ANOVA approach to model validation. *Simul. Model. Pract. Theory*, 78, 18–27.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57, 289–300.

Cardinali, A. and Nason, G. P. (2010) Costationarity of locally stationary time series. *J. Time Ser. Econom.*, 2, doi:10.2202/1941-1928.1074.

Chang, X. and Stein, M. L. (2013). Decorrelation property of discrete wavelet transform under fixed-domain asymptotics. *IEEE Trans. Inf. Theory*, 59, 8001–8013.

Cryer, J. D. and Chan, K-S. (2008). *Time Series Analysis with Applications in R*. Springer.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.*, 25, 1–37.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

Dette, H. and Neumeyer, N. (2001). Nonparametric analysis of covariance. *Ann. Appl. Stat.*, 29, 1361–1400.

Dwivedi, Y. and Subba Rao, S. (2011). A test for second-order stationarity of a time series based on the discrete Fourier transform. *J. Time Ser. Anal.*, 32, 68–91.

Embleton, J., Knight, M. I., and Ombao, H. (2020). Multiscale modelling of replicated nonstationary time series. *arXiv:2005.09440*.

Embleton, J., Knight, M.I. and Ombao, H. (2022). Multiscale spectral modelling for nonstationary time series within an ordered multiple-trial experiment. *Ann. Appl. Stat.*, to appear.

Euán, C., Ombao, H. and Ortega, J. (2018). The hierarchical spectral merger algorithm: A new time series clustering procedure. *J. Classif.*, 35, 71–99.

Fan, J. and Lin, S-K. (1998). Test of significance when data are curves. *J. Am. Statist. Ass.*, 93, 1007–1021.

Fiecas, M. and Ombao, H. (2016). Modelling the evolution of dynamic brain processes during an associative learning experiment. *J. Am. Statist. Ass.*, 111, 1440–1453.

Fryzlewicz, P. (2005). Modelling and forecasting financial log-returns as locally stationary wavelet processes. *J. Appl. Stat.*, 32, 503–528.

Fryzlewicz, P. and Nason, G. (2006). Haar - Fisz estimation of evolutionary wavelet spectra. *J. R. Statist. Soc. B*, 68, 611–634.

Fryzlewicz, P. and Ombao, H. (2009). Consistent classification of nonstationary time series using stochastic wavelet representations. *J. Am. Statist. Ass.*, 104, 299–312.
Gorrostieta, C., Ombao, H., Prado, R., Patel, S. and Eskandar, E. (2012). Exploring dependence between brain signals in a monkey during learning. *J. Time Ser. Anal.*, 33, 771–778.
Granados-Garcia, G., Fiecas, M., Babak, S., Fortin, N.J. and Ombao, H. (2021). Brain waves analysis via a non-parametric Bayesian mixture of autoregressive kernels, *Comput. Stat. Data Anal.*, https://doi.org/10.1016/j.csda.2021.107409.
Guo, W., Dai, M. Ombao, H. and von Sachs, R. (2003). Smoothing spline ANOVA for time-dependent spectral analysis. *J. Am. Statist. Ass.*, 98, 643–652.
Hargreaves, J.K., Knight, M.I., Pitchford, J.W., Oakenfull, R. and Davis, S.J. (2018). Clustering nonstationary circadian plant rhythms using locally stationary wavelet representations. *Multiscale Model. Simul.*, 16, 184—214.
Hargreaves, J. K., Knight, M. I., Pitchford, J. W., Oakenfull, R., Chawla, S., Munns, J. and Davis, S. J. (2019). Wavelet spectral testing: application to nonstationary circadian rhythms. *Ann. Appl. Stat.*, 13, 1817–1846.
Killick, R., Knight, M.I., Nason, G. P. and Eckley, I.A. (2020). The local partial autocorrelation function and some applications. *Electron. J. Statist.*, 14, 3268–3314.
Krafty, R.T. (2016). Discriminant analysis of time series in the presence of within-group spectral variability. *J. Time. Ser. Anal.*, 37, 435–450.
Maharaj, E. A. and Alonso, A. M (2014). Discriminant analysis of multivariate time series: Application to diagnosis based on ECG signals. *Comput. Stat. Data Anal.*, 70, 67–87.
McKay, J. L., Welch, T. D. J., Vidakovic, B. and Ting, L. H. (2013). Statistically significant contrasts between EMG waveforms revealed using wavelet-based functional ANOVA. *J. Neurophysiol.*, 109, 591–602.
Morris, J. S. (2015). Functional Regression. *Ann. Rev. Stat. Appl*, 2, 321–359.
Nason, G. P. (2012). Simulation study comparing two tests of second-order stationarity and confidence intervals for localized autocovariance. *Technical Report 12:02*. Statistics Group, University of Bristol, Bristol. *arXiv:1603.06415*.
Nason, G. P., von Sachs, R. and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. R. Statist. Soc. B*, 62, 271–292.
Nason, G. P. (2013). A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary times series. *J. R. Statist. Soc. B*, 75, 879–904.
Ombao, H,. Raz, J., von Sachs, R. and Guo, W. (2002). The SLEX model of a non-stationary random process. *Ann. Inst. Statist. Math.*, 54, 171–200.
Park, T., Eckley, I. and Ombao, H. (2014). Estimating time-evolving partial coherence between signals via multivariate locally stationary wavelet processes. *IEEE Trans. Signal Process.*, 62, 5240–5250.
Paparoditis, E. (2009). Testing temporal constancy of the spectral structure of a time series. *Bernoulli*, 15, 1190–1221.
Paparoditis, E. (2010). Validating stationarity assumptions in time series analysis by rolling local periodograms. *J. Am. Statist. Ass.*, 105, 839–851.
Priestley, M. B. and Subba Rao, T. (1969). A test for non-stationarity of time-series. *J. R. Statist. Soc. B*, 31, 140–149.
Ramsay, J. O. and Silverman, B. M. (2005). *Functional Data Analysis*. Springer.
Raz, J. and Turetsky, B. (1999). Wavelet ANOVA and fMRI. In: *Proceedings of the SPIE: Wavelet applications in signal and image processing VII*, 3813.
Sanderson, J., Fryzlewicz, P. and Jones, W. (2010). Estimating linear dependence between nonstationary time series using the locally stationary wavelet model. *Biometrika*, 97, 435–446.
Seger, C. and Cincotta, C. (2006). Dynamics of frontal, striatal, and hippocampal systems during rule learning. *Cerebral Cortex*, 16, 1546–1555.
Shumway, R. H. (1988). *Applied Statistical Time Series Analysis*. Prentice-Hall.
Tsay, R. S. (2010). *Analysis of Financial Time Series, Third Edition*. John Wiley & Sons, Inc.
Van Bellegem, S. and von Sachs, R. (2008). Locally adaptive estimation of evolutionary wavelet spectra. *Ann. Statist.*, 36, 1879–1924.
Vidakovic, B. (1999). *Statistical Modelling by Wavelets*. John Wiley & Sons, Inc.
Vidakovic, B. (2001). Wavelet-based functional data analysis: Theory, applications and ramifications. In: *Proceedings of PSFVIP-3*, Maui, Hawaii.
von Sachs, R. and Neumann, M. (2000). A wavelet-based test for stationarity. *J. Time Ser. Anal.*, 21, 597–613.

Appendix A Experimental data description

Each trial (replicate) consists of $T = 2048$ time points, corresponding to approximately 2 seconds of data. The design of the experiment rigorously splits each trial into four time blocks of 512 milliseconds each, ensuring that the timeline matched from trial to trial, as follows. For the first block the macaque fixated on a screen; a picture (one of four) was then presented on the screen for the next time block; this was followed by an empty screen for the next interval; for the last 512 milliseconds the macaque was presented with a picture of four doors, one of which associated with the picture visual from the second time block. The macaque's task was to select the correct door using a joystick. Correct and incorrect choices were signified via a visual cue and a juice reward was given each time a correct selection was made. Our analysis here focusses on the correct trials, with the macaque having to learn the associations through repeated trials.

Appendix B Supporting evidence for normality of macaque data

In what follows we explore the tenability of the Gaussian assumption, as suggested by Fryzlewicz (2005) and also demonstrated by Hargreaves et al. (2019). For each trial, we propose to standardise the (zero-mean) process using a localised estimate of the standard deviation. The estimate was obtained by means of taking the square root of the estimated lag zero localised autocovariances (Nason, 2013). Further, we have found similar results when obtaining the estimate by means of a localised Gaussian kernel with bandwidth chosen using the methods of Fryzlewicz (2005). We report Q–Q plots of standardised series against the normal quantiles. These correspond to trials displaying typical behaviour for Hc and NAc series and the plots show that the normality assumption mostly holds, with some departures for NAc records.

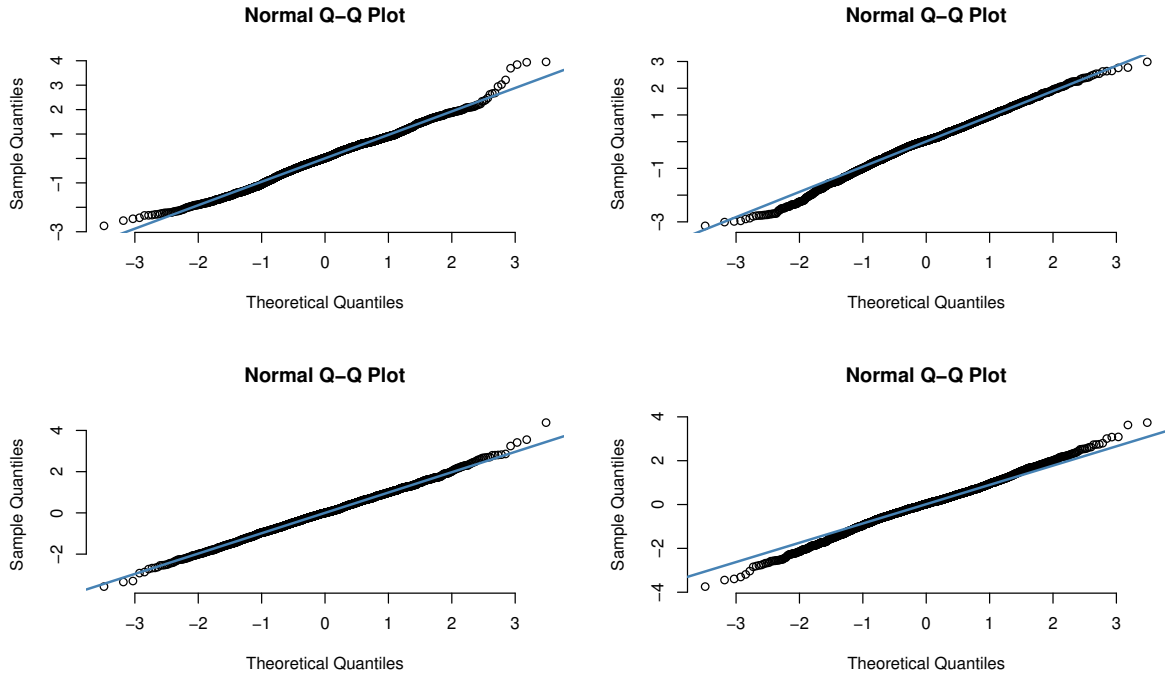


Figure 14: Q-Q plots of the Hc (*top row*) and NAc (*bottom row*) data in the correct trials (replicates). *Left column*: replicate 121; *Right column*: replicate 231.

Appendix C Further simulation details for the location-specific test

We simulate MULT-LSW processes with the spectral structures $\{S_j(z,\nu)\}$ (below) by making use of the equation (1), with amplitudes $\tilde{W}_j(z,\nu) = S_j^{1/2}(z,\nu)$ and with orthonormal Gaussian innovations. We use the resolution level notation $j := J(T) - j$ for the jth finest scale.

C.1 Simulation 1 (S1)

For rescaled time $z = k/T$ and replicates $\nu = r/R$, variations of simulation 1 are defined as follows.

1A. Firstly, we aim to evaluate the test size. Specifically, we want to get a sense of how the test performs in the absence of any evolution in the true spectra across the replicates, hence we define the true spectra to be

$$S_j(z,\nu) = \begin{cases} \sin^2\left(2\pi\frac{k}{T}\right), & \text{for } j = J(T)-4, k \in (0,T), r \in (0,R) \\ 0, & \text{otherwise.} \end{cases}$$

1B. We now introduce a burst value for the spectra which is contained in the first $T/2$ locations and defined for the second half of the replicates. We mathematically define the true spectra as a function of R and T for simulation 1B as follows

$$S_j(z,\nu) = \begin{cases} \sin^2\left(2\pi\frac{k}{T}\right), & \text{for } j = J(T)-4, k \in (0,T), r \in (0,\frac{1}{2}R) \\ & \quad \text{and } k \in (\frac{1}{2}T+1,T), r \in (\frac{1}{2}R+1,R) \\ \text{'burst'}, & \text{for } j = J(T)-4, k \in (0,\frac{1}{2}T), r \in (\frac{1}{2}R+1,R) \\ 0, & \text{otherwise,} \end{cases}$$

where 'burst' will assume values 1, 2 or 5. A visualisation of such a process realisation appears in Figure 15.

1C. For this simulation we reproduce the spectra defined in 1B but reduce the time sequence length containing a burst to the first $T/4$, such that

$$S_j(z,\nu) = \begin{cases} \sin^2\left(2\pi\frac{k}{T}\right), & \text{for } j = J(T)-4, k \in (0,T), r \in (0,\frac{1}{2}R) \\ & \quad \text{and } k \in (\frac{1}{4}T+1,T), r \in (\frac{1}{2}R+1,R) \\ 2, & \text{for } j = J(T)-4, k \in (0,\frac{1}{4}T), r \in (\frac{1}{2}R+1,R) \\ 0, & \text{otherwise.} \end{cases}$$

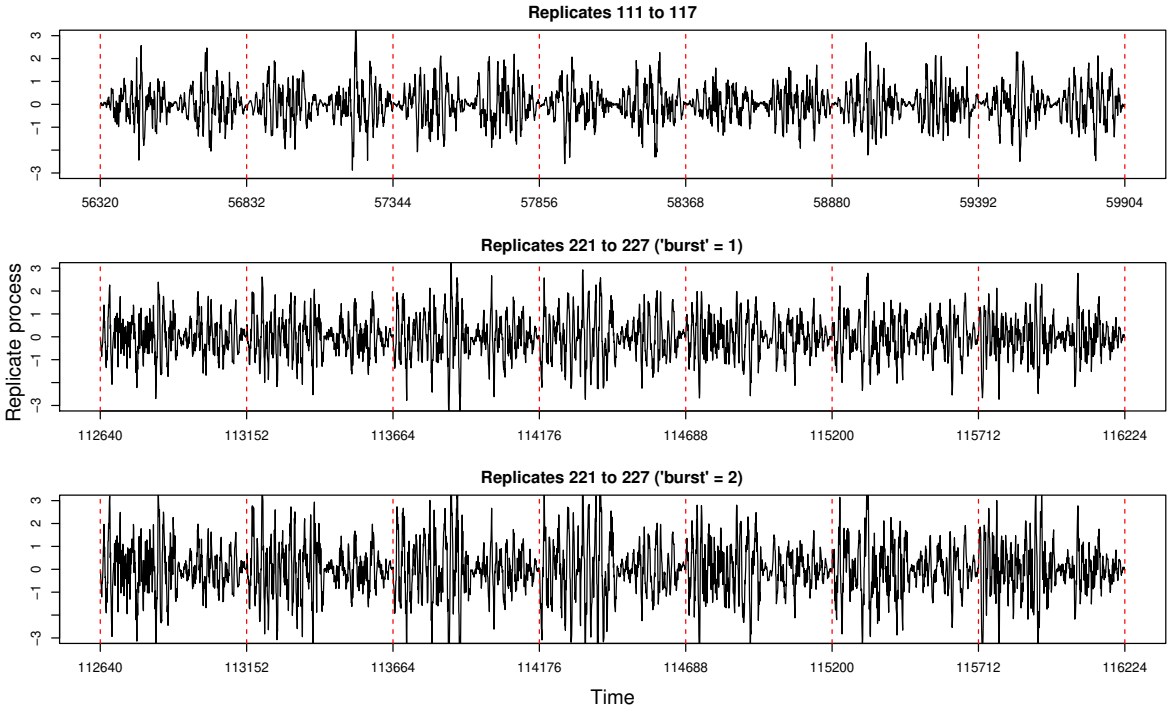
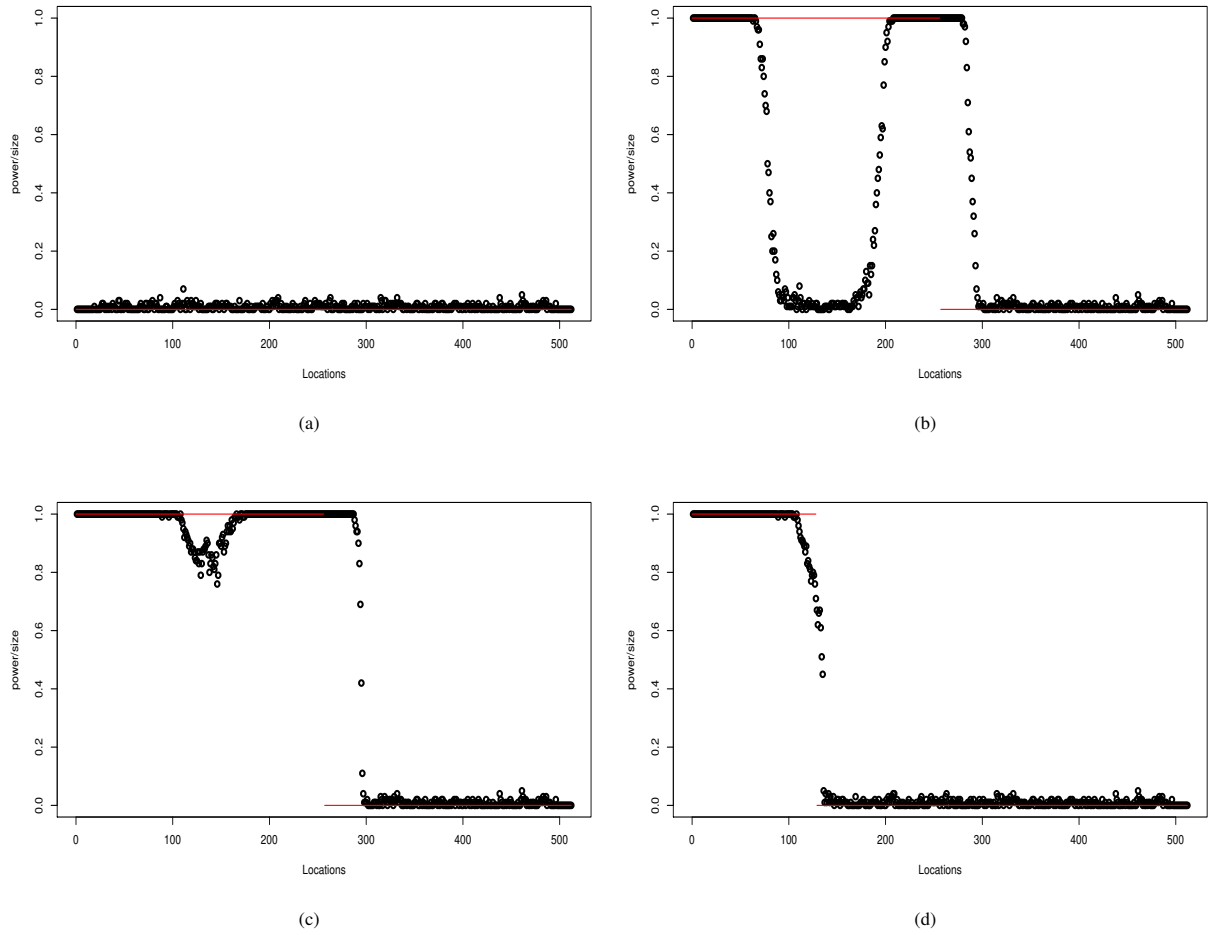


Figure 15: Concatenated simulated series for simulation 1B.

Within this simulated setup, we will investigate the power and size of the proposed test, pictorially shown in Figure 16, where the horizontal lines at 1 and 0 identify the locations expected to be rejected and not rejected, respectively.

C.2 Simulation 2 (S2)

For rescaled time $z = k/T$ and replicates $\nu = r/R$, variations of simulation 2 are defined as follows.



(a) (b) (c) (d)

Figure 16: Simulation 1 empirical power and size estimates for each location over 100 runs. *(a)*: simulation 1A; *(b)*: simulation 1B with 'burst'=1; *(c)*: simulation 1B with 'burst'=2; *(d)*: simulation 1C with 'burst'=2.

2A. The simulated true spectra for our second example are defined by

$$S_j(z,\nu) = \begin{cases} \cos^2\left(2\pi\frac{k}{T}\right), & \text{for } j = J(T)-4, k \in (0,T), r \in (0,\frac{3}{8}R) \cup (\frac{5}{8}R+1, R) \\ & \quad \text{and } k \in (0,\frac{3}{8}T) \cup (\frac{5}{8}T+1, T), r \in (\frac{3}{8}R+1, \frac{5}{8}R) \\ \text{'burst'}, & \text{for } j = J(T)-4, k \in (\frac{3}{8}T+1, \frac{5}{8}T), r \in (\frac{3}{8}R+1, \frac{5}{8}R) \\ 0, & \text{otherwise.} \end{cases}$$

where the burst is now defined for fewer replicates. There exists a squared cosine behaviour in all replicates except for the middle $R/4 = 64$ (here, replicates 97 to 160), which have a burst across the middle $T/4 = 128$ locations (here, locations 193 to 320). See also Figure 17.

2B. We now dramatically reduce the number of replicates and locations that we define to have changing spectral

characteristics. The spectra are defined as follows

$$S_j(z,\nu) = \begin{cases} \cos^2\left(2\pi\frac{k}{T}\right), & \text{for } j = J(T)-4, k \in (0,T), r \in (0,\frac{3}{8}R) \cup (\frac{4}{8}R+1,R) \\ & \quad \text{and } k \in (0,T^*-\frac{1}{16}T) \cup (T^*+\frac{1}{16}T+1,T), r \in (\frac{3}{8}R+1,\frac{4}{8}R) \\ \text{'burst'}, & \text{for } j = J(T)-4, k \in (T^*-\frac{1}{16}T+1, T^*+\frac{1}{16}T), r \in (\frac{3}{8}R+1,\frac{4}{8}R) \\ 0, & \text{otherwise}, \end{cases}$$

where 'burst' will assume values 1, 2 or 5. Here $T^* = \lfloor 0.6T \rfloor$ and in rescaled time $z = k/T$ and $\nu = r/R$. For a simulation with $R = 256$ and $T = 512$ we again have a squared cosine behaviour in all the replicates except for $R/8 = 32$ (here, replicates 97 to 128) where we have defined a burst of 1 in the spectra over $T/8 = 64$ locations (here, 276 to 339). A process realisation appears in Figure 18.

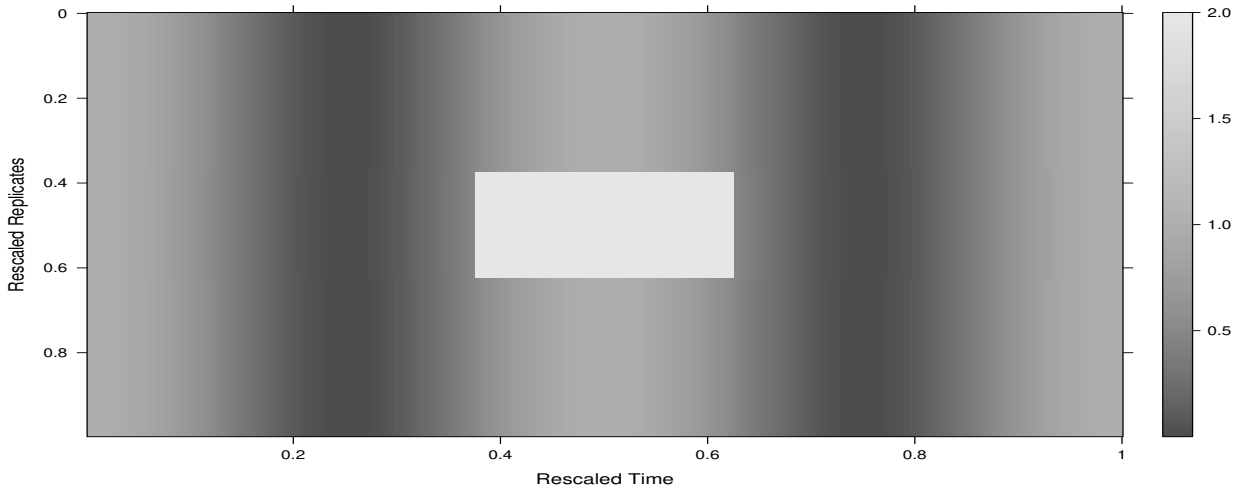


Figure 17: Time-replicate plot of the true spectra in level 5 for simulation 2A.

Visualisations of the empirical test power appear below in Figures 19 and 20.

From Figure 19 for simulation 2A, it can be seen how the test breaks down when the burst is 1 and thus the difference between spectra over replicates is of 1 or less. For a burst of 2 and then 5 we see much improvement in the number of true positives being correctly identified (rejected) but as noticed previously, for a higher burst more negatives are being incorrectly identified by the test at the location rejection boundaries.

From the empirical power and size estimates shown in Figure 20 for simulation 2B, it is evident that the closeness in spectra of the squared cosine and burst of 1 is again causing the test to struggle to reject the null hypothesis at the locations where an evolution across replicates does exist. For bursts of 2 and 5, the test sufficiently rejects the correct locations but present again is the effect at the rejection boundaries.

C.3 Simulation 3 (S3)

Stepping up the challenge posed by the previous setup, for rescaled time $z = k/T$ and replicates $\nu = r/R$, variations of simulation 3 are defined as follows. This is equates to a further investigation into the power of the proposed test.

3A. For evaluating the power of our test, we define a MULT-LSW process with evolving spectra across both the time and replicate arguments, as follows,

$$S_j(z,\nu) = \begin{cases} 4\frac{r}{R}\sin^2\left(2\pi\frac{k}{T}(1+2\frac{r}{R})\right), & \text{for } j = J(T)-4, k \in (0,T), r \in (0,R) \\ 0, & \text{otherwise.} \end{cases}$$

A visualisation appears in Figure 21.

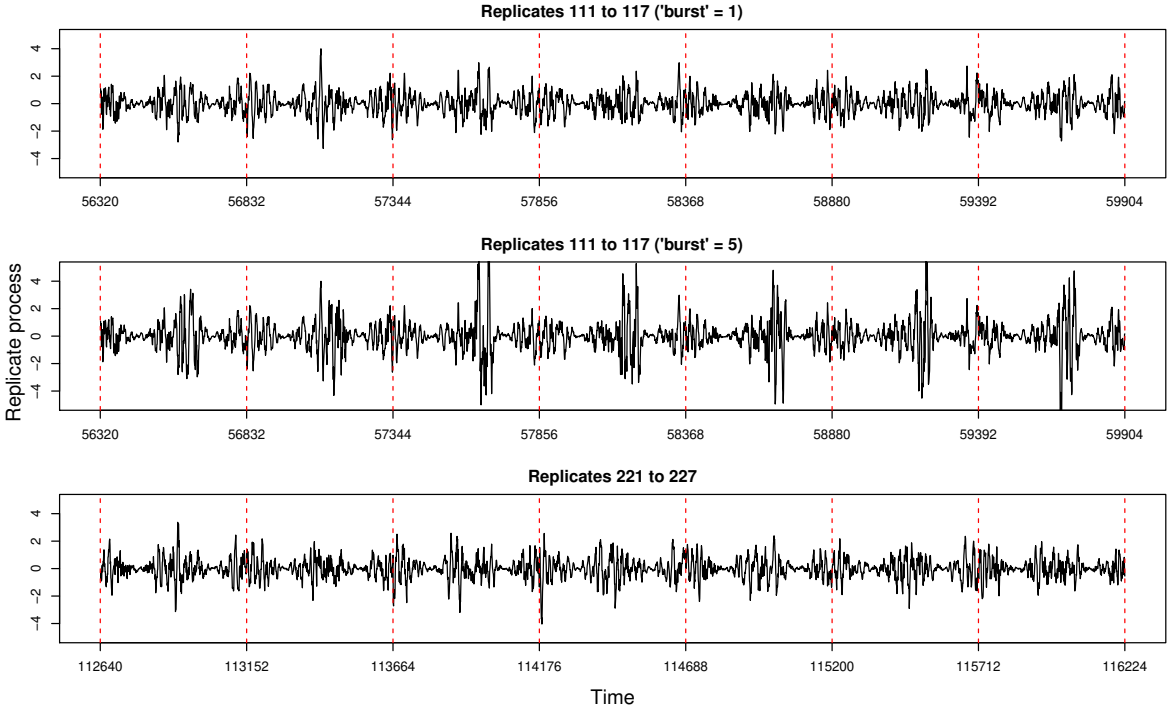


Figure 18: Concatenated simulated series for simulation 2B.

3B. We now introduce an evolving spectral structure for two neighbouring wavelet scales. The spectra are defined as follows

$$S_j(z,\nu) = \begin{cases} 4(1-\frac{r}{R})\cos^2\left(\frac{-1}{3}\pi + \frac{4}{3}\pi\frac{k}{T}\right), & \text{for } j = J(T)-3, k \in (\frac{1}{4}T+1, T), r \in (0,R) \\ 4\cos^2\left((4\pi + 10\frac{r}{R})\frac{k}{T}\right), & \text{for } j = J(T)-2, k \in (0, \frac{1}{2}T), r \in (0,R) \\ 0, & \text{otherwise.} \end{cases}$$

We expect the test to make rejections across all time locations.

Figure 22 displays the empirical power estimates for simulation 3A, with different R and T. It is clear that the test performance increases with R due to improved estimation of the spectra as $R, T \to \infty$. For both $R = 128$ and $R = 256$, note the power estimates at locations preceding location $0.2T$ are poor indicating the test is having difficulty to identify a breach in the null hypothesis of constancy of the spectra over replicates. Comparing these locations where the test struggles to the true spectra across replicates around a rescaled time of 0.15 in Figure 21, the results suggest that the spectra evolve too slowly over the replicates to be picked up by the test. Recall that the test struggled in previous simulations when the difference between the burst and sine/cosine spectra was of 1 or less.

The empirical power estimates for simulation 3B are displayed in Figure 23 for different R and T. We have evidence that the test does not always correctly reject when the evolution of the spectra over the replicates is too slow, specifically between rescaled time points (approximately) 0 to 0.05 and 0.15 to 0.2 (see Figure 23), comparable with rescaled time points (approximately) 0 to 0.05 and 0.15 to 0.2 (see Figure 7). The results here follow in similar vein to those above, showing that as spectral estimation improves asymptotically with $R, T \to \infty$, the performance of the test improves.

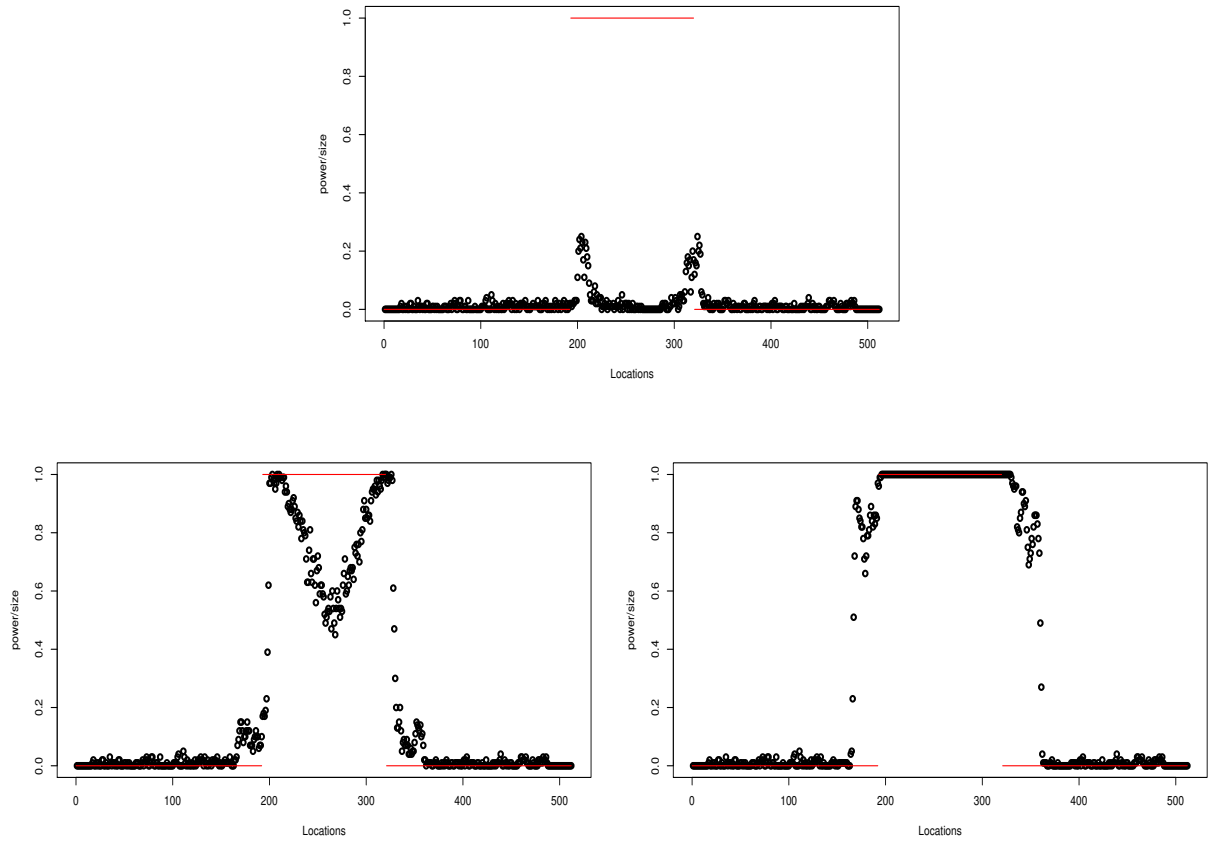


Figure 19: Simulation 2A empirical power and size estimates for each location over 100 runs. *Top*: 'burst'=1; *bottom left*: 'burst'=2; *bottom right*: 'burst'=5.

Appendix D Further details for the global-specific test

Figure 24 provides an illustration of the potential volatility clustering manifest in the process realisation corresponding to simulation setup 's3a'.

The parameter design for simulation 'p5' is as follows

$$\gamma_{t;T}^{r;R} = \begin{bmatrix} 0.9 & \cdots & \gamma_t^1 & \cdots & -0.9 \\ \vdots & & \vdots & & \vdots \\ \gamma_1^r & \cdots & \vdots & \cdots & \gamma_T^r \\ \vdots & & \vdots & & \vdots \\ -0.9 & \cdots & \gamma_t^R & \cdots & 0.9 \end{bmatrix} \text{ with } t = 0, ..., T-1,\ r = 0, ..., R-1.$$

Appendix E Further evidence for the macaque LFPs investigation

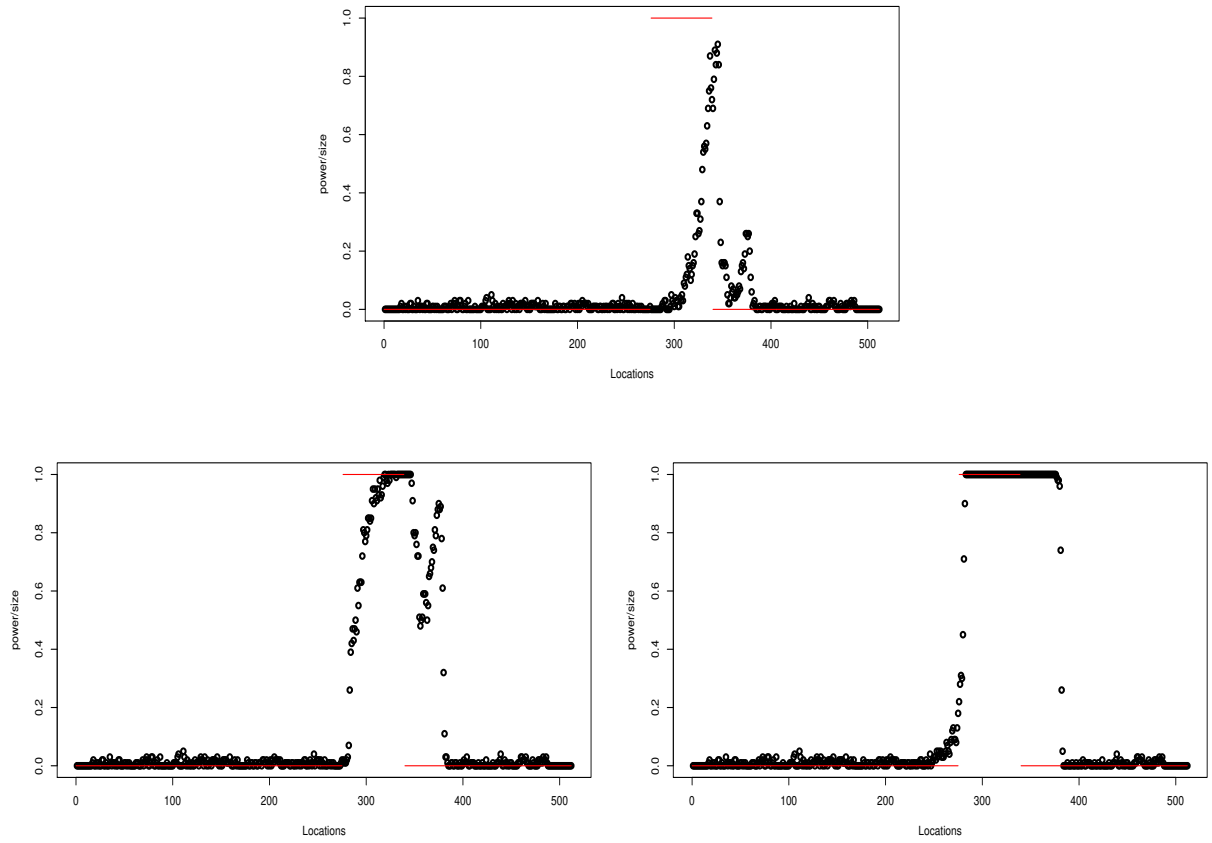


Figure 20: Simulation 2B empirical power and size estimates for each location over 100 runs. *Top*: 'burst'=1; *bottom left*: 'burst'=2; *bottom right*: 'burst'=5.

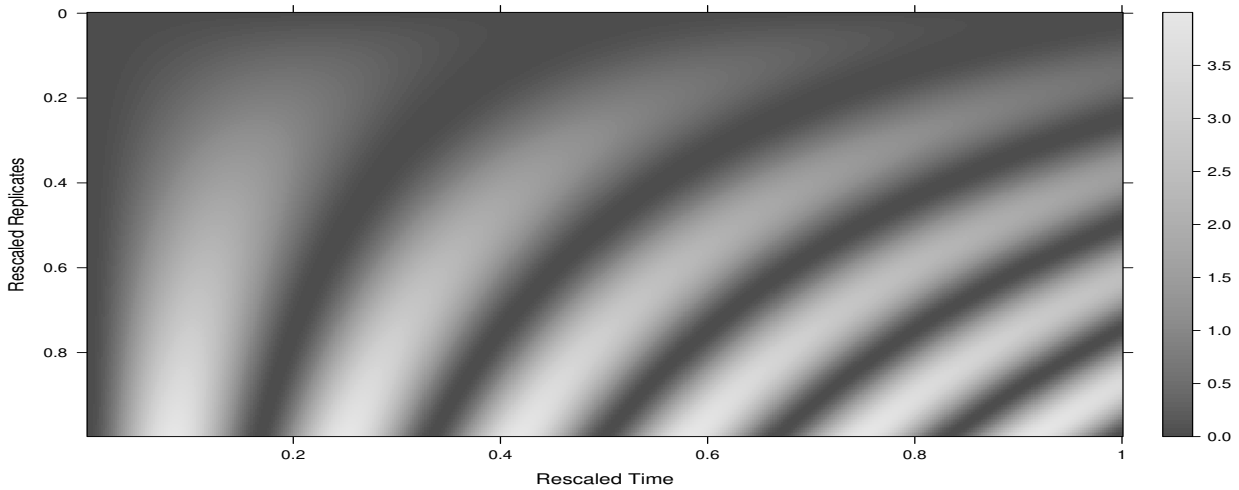


Figure 21: Time-replicate plot of the true spectra in level 5 for simulation 3A.

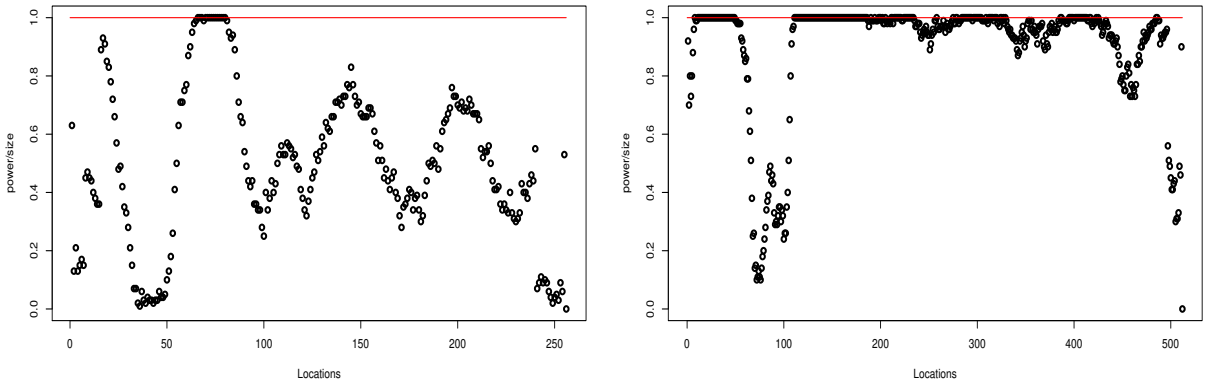


Figure 22: Simulation 3A empirical power estimates for each location over 100 runs. *Left*: R=128, T=256; *right*: R=256, T=512.

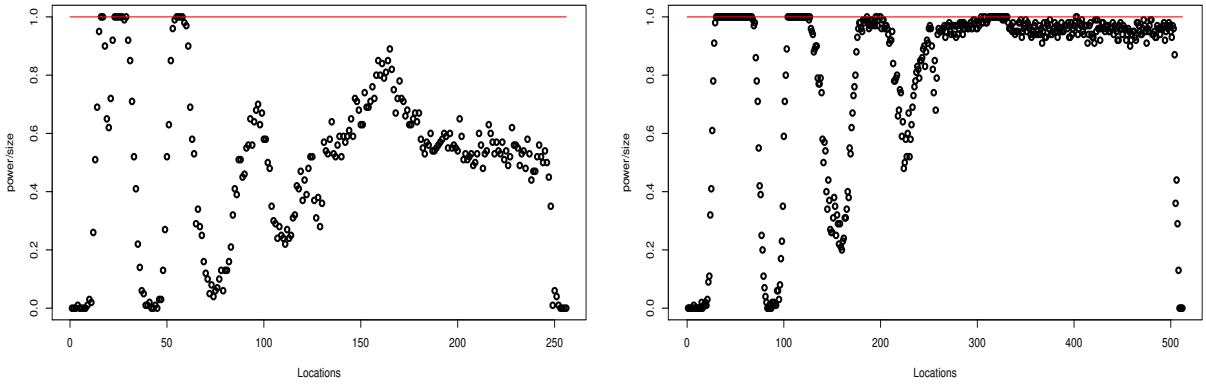


Figure 23: Simulation 3B empirical power estimates for each location over 100 runs. *Left*: R=128, T=256; *right*: R=256, T=512.

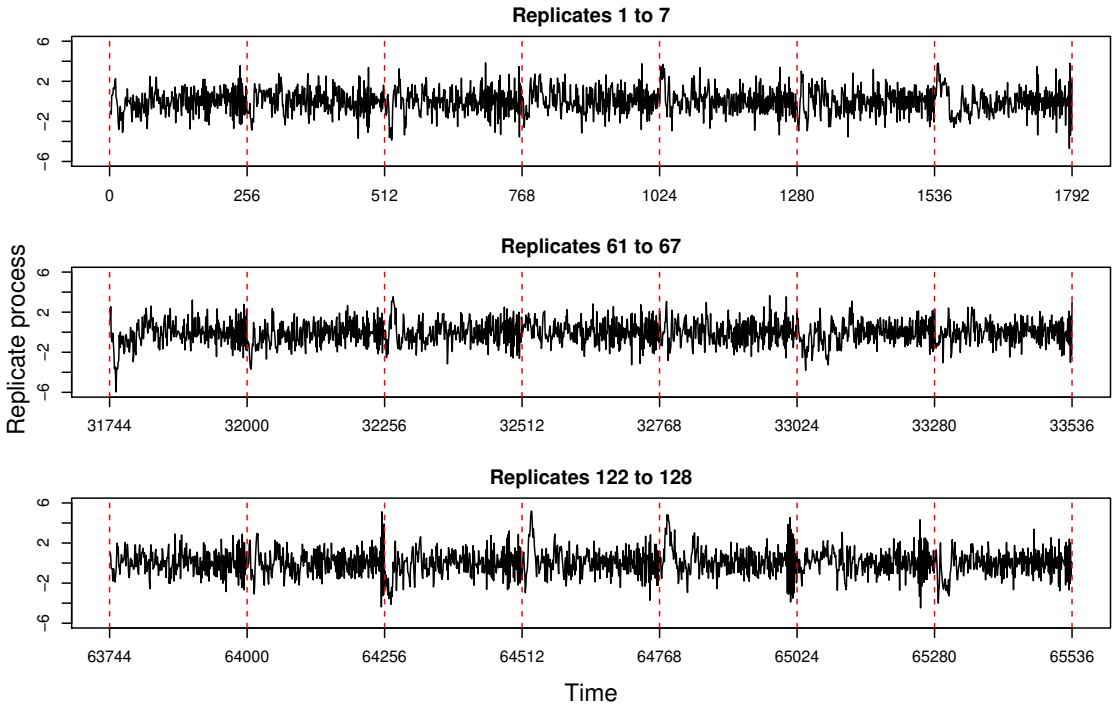
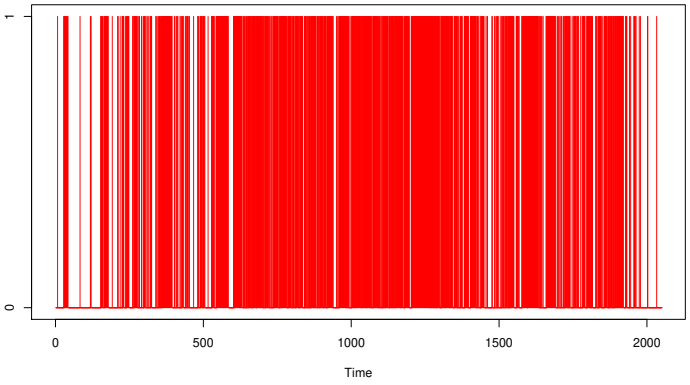
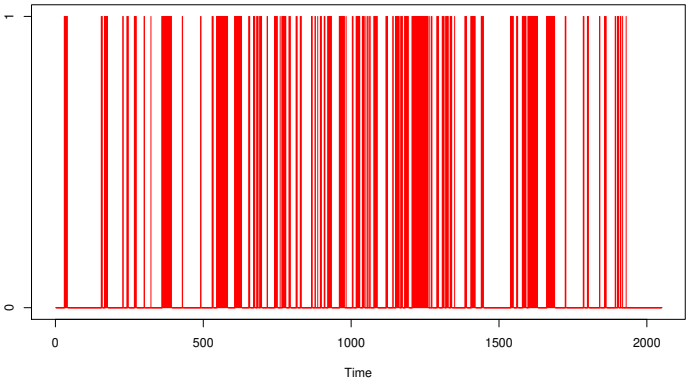


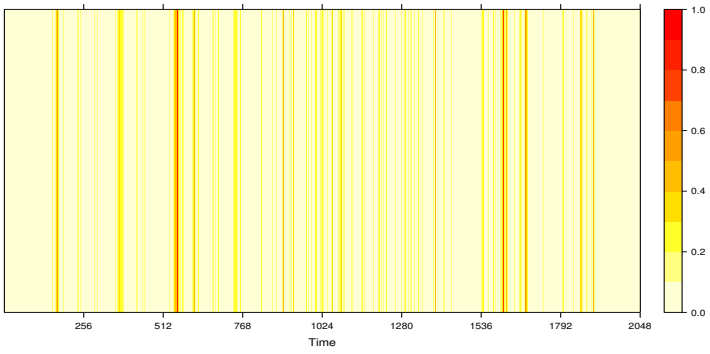
Figure 24: Concatenated series of a simulated ‘s3a’ process.



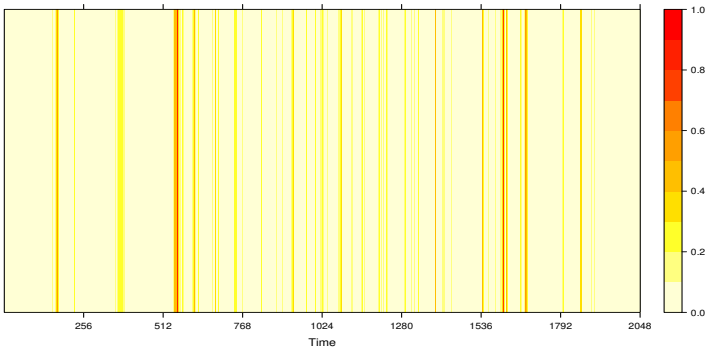
(a)



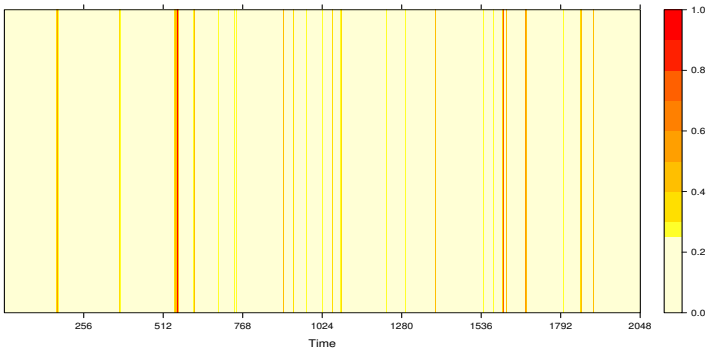
(b)



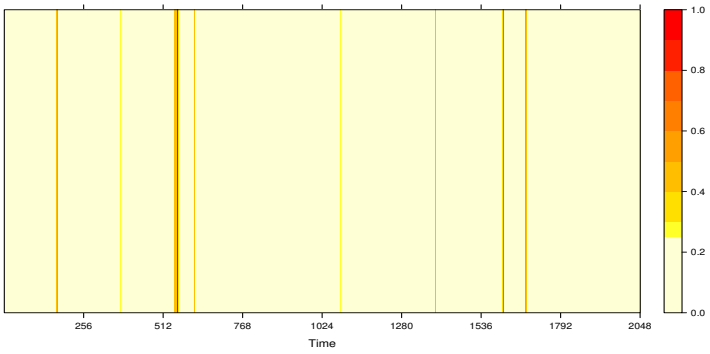
(c)



(d)



(e)



(f)

Figure 25: Rejection plots for the location-specific replicate-effect test carried out on the correct trials of the hippocampus (Hc) dataset. *(a)*: binary plot where a vertical line to 1 indicates that the test identified the location as rejecting the null hypothesis of constancy; *(b)*: as in (a) but only for rejected locations with two rejected neighbour locations each side; *(c)*: percentage of hypothesis tests rejected for each location; *(d)*: as in (c) but only for rejected locations with two rejected neighbour locations each side; *(e)*: percentage of hypothesis tests rejected for each location thresholded at 25%; *(f)*: as in (e) but only for rejected locations with two rejected neighbour locations each side. The scale for plots (c)-(f) indicates the percentage of rejections out of 40 hypothesis tests per location.

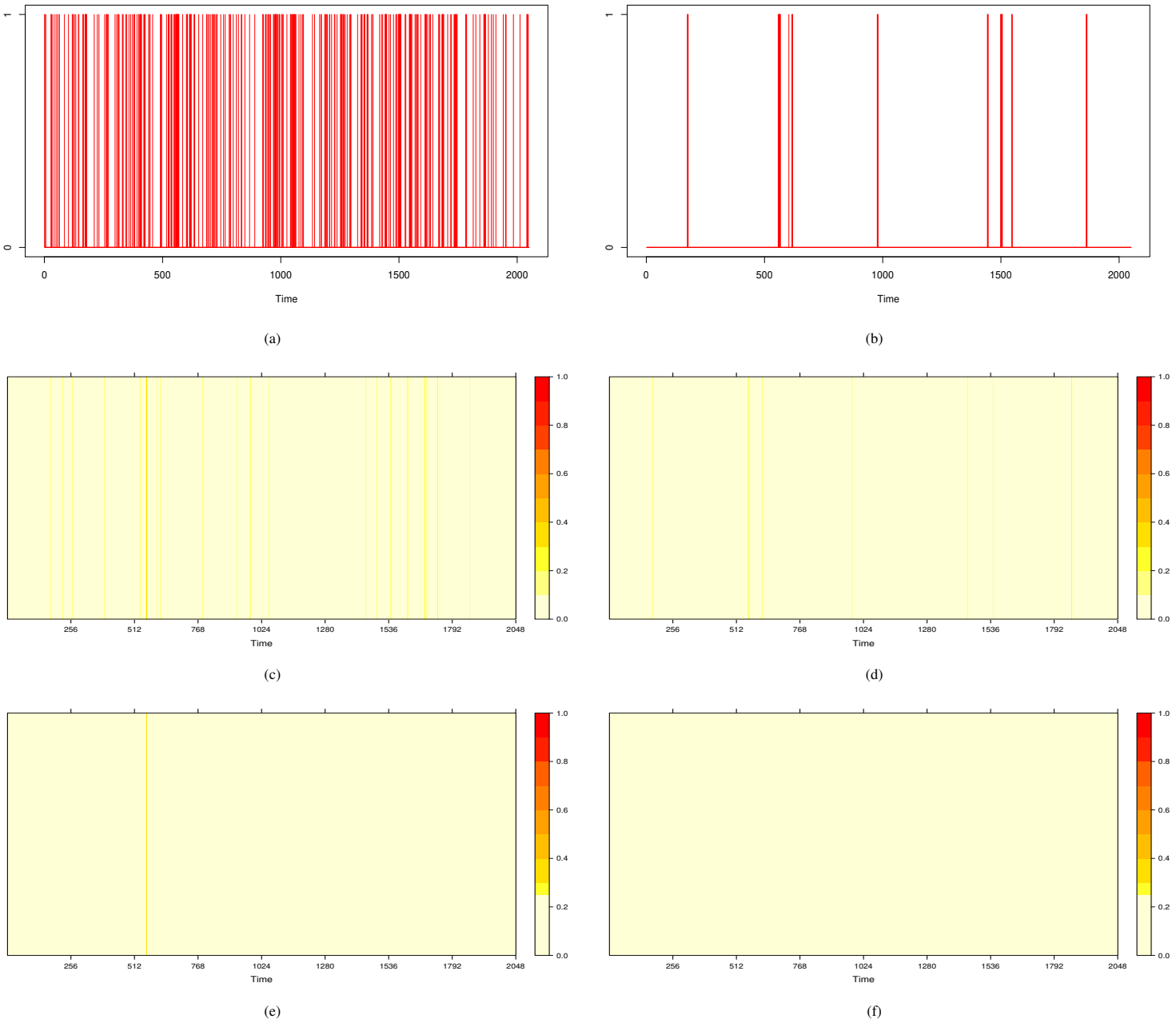


Figure 26: Rejection plots for the location-specific replicate-effect test carried out on the correct trials of the nucleus accumbens (NAc) dataset. *(a)*: binary plot where a vertical line to 1 indicates that the test identified the location as rejecting the null hypothesis of constancy; *(b)*: as in (a) but only for rejected locations with two rejected neighbour locations each side; *(c)*: percentage of hypothesis tests rejected for each location; *(d)*: as in (c) but only for rejected locations with two rejected neighbour locations each side; *(e)*: percentage of hypothesis tests rejected for each location thresholded at 25%; *(f)*: as in (e) but only for rejected locations with two rejected neighbour locations each side. The scale for plots (c)-(f) indicates the percentage of rejections out of 40 hypothesis tests per location.

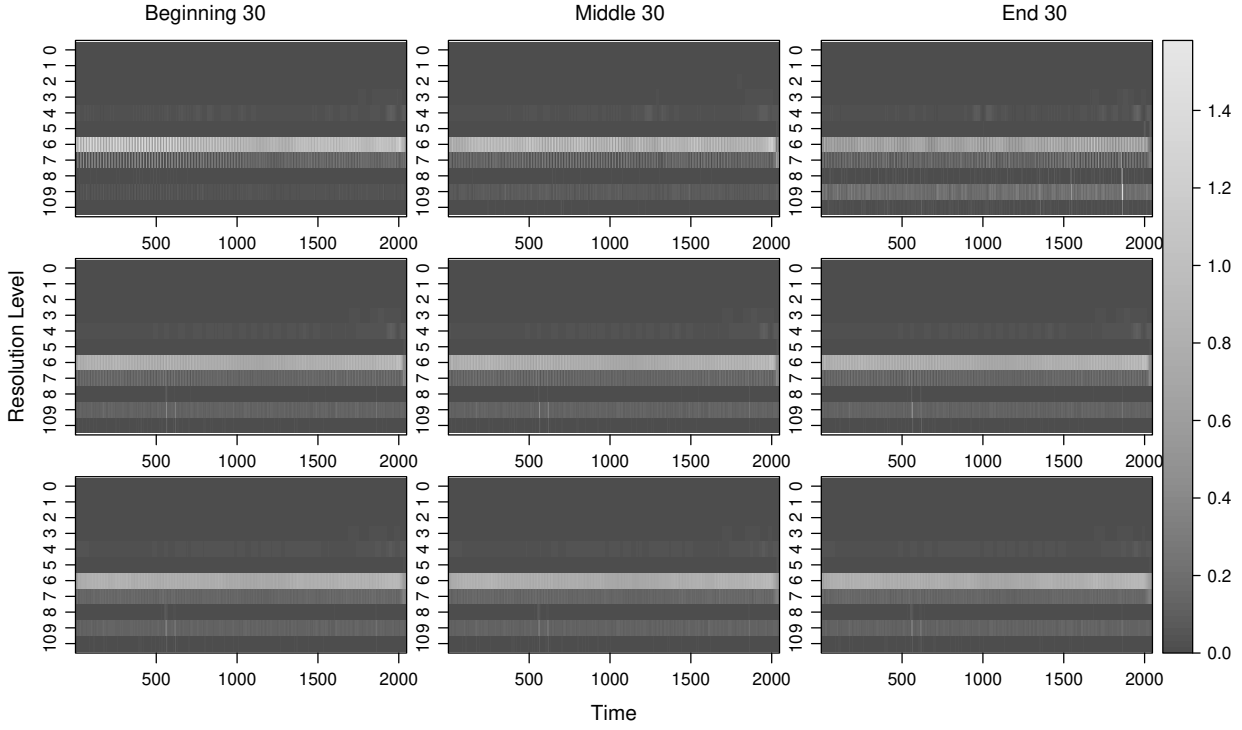


Figure 27: Time-scale nucleus accumbens (NAc) plots for the correct trials computed for the global test of replicate-effect. Spectral estimates are shown for the average over 30 replicates in the beginning, middle and end of the experiment. *Top*: spectral estimates of the correct NAc trials; *Middle*: NAc spectral estimates averaged across all trials; *Bottom*: spectral estimates averaged over 100 bootstrap realisations.