Statistical analysis of periodic data in neuroscience

Daniel H. Baker

2021-01-11

1 Abstract

Many experimental paradigms in neuroscience involve driving the nervous system with periodic sensory stimuli. Neural signals recorded with a variety of techniques will then include phase-locked oscillations at the stimulation frequency. The analysis of such data often involves standard univariate statistics such as T-tests, conducted on the Fourier amplitude components (ignoring phase). However, the assumptions of these tests will often be violated because amplitudes are not normally distributed, and furthermore weak signals might be missed if the phase information is discarded. An alternative approach is to conduct multivariate statistical tests using the real and imaginary Fourier components. Here the performance of two multivariate extensions of the T-test are compared: Hotelling's T^2 and a variant called T^2_{circ}. A novel test of the assumptions of T^2_{circ} is developed, based on the condition index of the data (the square root of the ratio of eigenvalues of a bounding ellipse), and a heuristic for excluding outliers using the Mahalanobis distance is proposed. The T^2_{circ} statistic is then extended to multi-level designs, resulting in a new statistical test termed $ANOVA^2_{circ}$. This has identical assumptions to T^2_{circ}, and is shown to be more sensitive than MANOVA when these assumptions are met. The use of these tests is demonstrated for two publicly available empirical data sets, and practical guidance is suggested for choosing which test to run. Implementations of these novel tools are provided as an R package, in the hope that their wider adoption will improve the sensitivity of statistical inferences involving periodic data.

Keywords: multivariate statistics, Fourier analysis, steady-state, condition index, Mahalanobis distance.

2 Background

A widely used paradigm in many branches of neuroscience is to drive the nervous system using periodic stimuli. This entrains neural responses at the stimulation frequency, resulting in high signal-to-noise ratios relative to single stimulus presentations. These periodic responses, often called the *steady-state* or *frequency following* response, can be recorded using invasive methods from single neurons (Enroth-Cugell and Robson, 1966) and local field potentials (Morrone et al., 1987), or with non-invasive electroencephalography (EEG) and magnetoencephalography (MEG) systems, both in humans (Norcia et al., 2015) and in diverse animal species including insects (Afsari et al., 2014), birds (Porciatti et al., 1990), rodents (Hwang et al., 2019) and primates (Nakayama and Mackeben, 1982). Steady-state methods are used to measure early sensory responses in vision (Regan, 1966), hearing (Rees et al., 1986) and somatosensation (Snyder, 1992), and closely related paradigms have been developed to target specific stimulus features such as orientation (Braddick et al., 1986), and facial expression (Gray et al., 2020) and identity (Liu-Shuang et al., 2014). In fMRI research, *travelling wave* methods (Engel et al., 1994; Sereno et al., 1995) are used to map the retinotopic responses of early visual cortex using stimuli that change periodically in spatial position. Finally, physiological reflexes such as the pupillary response to light can be entrained in a similar way (Spitschan et al., 2014).

A convenient way to analyse the data from periodic stimulation experiments is to take the Fourier transform of the measured signal. The amplitude of the response at the stimulation frequency (and its harmonics -

integer multiples of the stimulation frequency) is a precise and well-defined index of the brain's response (see Figure 1a,b). Fourier spectra comprise both amplitude and phase information that can be expressed in polar coordinates (Figure 1c), or equivalently as complex numbers with real and imaginary components (Figure 1d). In many studies the phase information is routinely discarded, and statistical comparisons are performed on the amplitude data only. However an alternative is to use multivariate statistics, which take into account both the amplitude and phase information (represented as real and imaginary components). Multivariate methods have the advantage that they are more sensitive to weak signals, and therefore offer increased statistical power relative to univariate methods.

For pointwise and pairwise comparisons, Hotelling's T^2 statistic (Hotelling, 1931) is a multivariate extension of the T-test. For the one-sample case, the test statistic is defined as:

$$T^2 = N(\bar{x} - \mu)'C^{-1}(\bar{x} - \mu), \tag{1}$$

where N is the number of observations, $\bar{x}$ is the multivariate sample mean, μ is the point of comparison, C^{-1} is the inverse covariance matrix, and $'$ denotes vector transposition. Conceptually, the T^2 statistic extends the univariate T-statistic by incorporating the covariance between the dependent variables. Two-sample and paired variants are also available, and the test can be applied with an arbitrary number of dependent variables (though here only the bivariate case will be considered).

More recently, Victor and Mast (1991) proposed a simpler version of T^2, called T^2_{circ}. The T^2_{circ} statistic makes the strong assumption that the dependent variables (real and imaginary Fourier components) are uncorrelated and have equal variance. When these conditions are met, the test statistic for the one-sample case is defined as:

$$T^2_{circ} = (N - 1)\frac{|\bar{x} - \mu|^2}{\Sigma|x_j - \bar{x}|^2} \tag{2}$$

where x_j denotes the jth observation of the dependent variables, and all other terms retain their previous meanings. Notice that no covariance term is present in equation (2), because of the independence assumption. This makes the statistic simpler to calculate, but causes problems when the assumption is violated (as will be demonstrated below). Conceptually, this statistic takes the vector difference between the centroid ($\bar{x}$) and a comparison point (μ), and scales by the mean length of the residual vector lines joining each data point to the centroid. Two-sample and repeated measures versions of the T^2_{circ} statistic are also possible.

In the present paper, best practice guidelines are developed for performing statistical tests on multivariate Fourier components derived from periodic stimulation paradigms. It is first demonstrated why parametric univariate statistics are inappropriate for such data, because amplitudes for weak signals are not normally distributed. Then conditions are investigated under which either the T^2 or T^2_{circ} statistic should be used. The range of sample sizes and effect sizes where T^2_{circ} is more sensitive is identified. A novel method for testing the assumptions of the T^2_{circ} statistic is developed, based on calculating the condition index of a multivariate data set. Appropriate methods for identifying outliers using the Mahalanobis distance are discussed, and a heuristic proposed. Next the logic of T^2_{circ} is extended to situations with more than two levels of the independent variable, and the performance of this novel $ANOVA^2_{circ}$ statistic is compared to MANOVA. Finally, the proposed techniques are demonstrated on two example data sets (from mice and humans), and best practice guidelines are recommended for analysis decisions.

All scripts used to generate this manuscript are available at: https://github.com/bakerdh/FourierStats. This includes a *Matlab* toolbox and an *R* package called *FourierStats*, featuring functions to implement Hotelling's (1931) T^2 statistic, Victor & Mast's (1991) T^2_{circ} statistic, and the condition index and $ANOVA^2_{circ}$ statistics proposed in this paper.
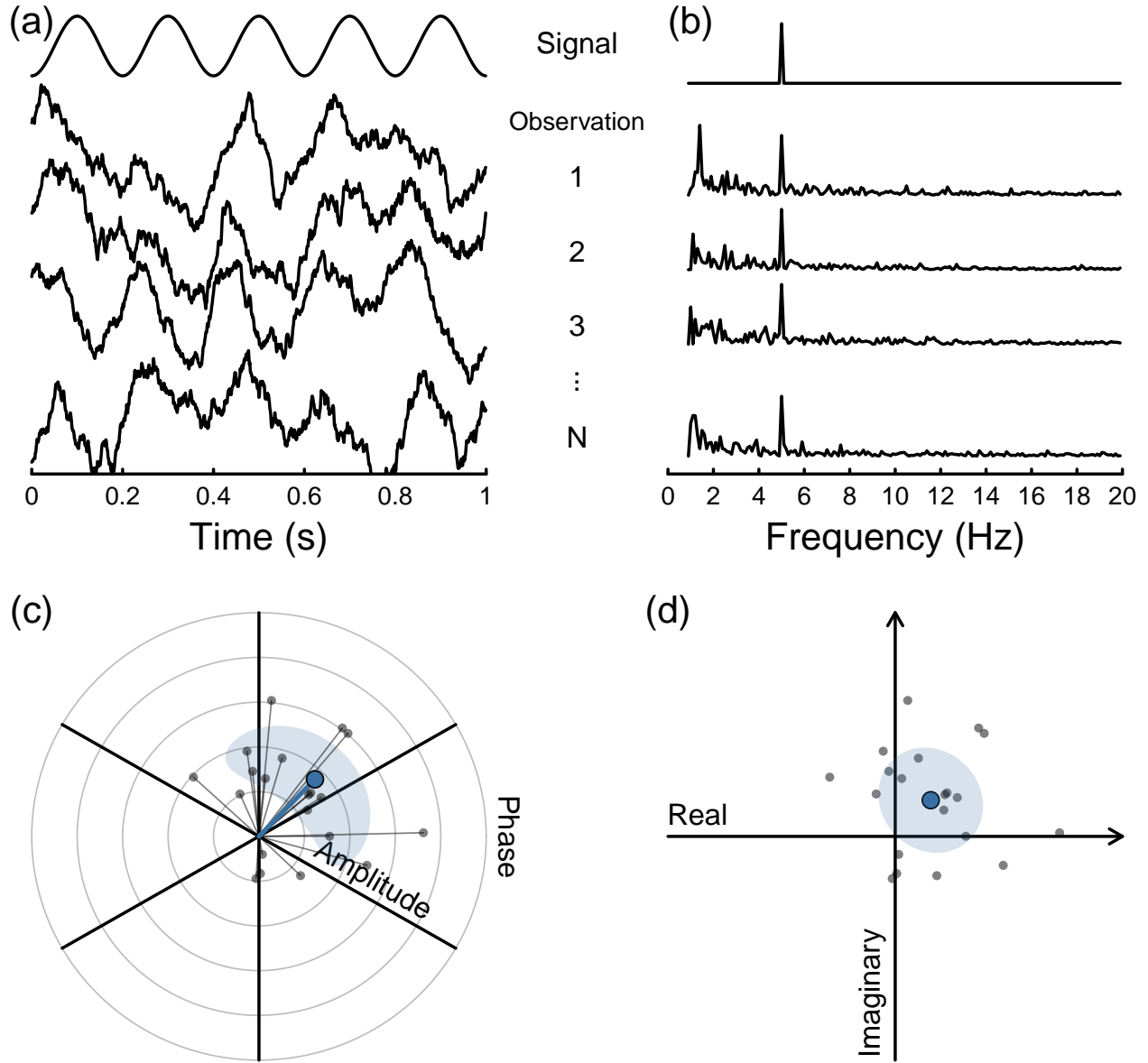
Figure 1: Illustration of the principles of Fourier analysis for time-varying signals. Panel (a) shows a sinusoidal signal waveform, and simulated neural responses for successive observations (repetitions in the same individual, or recorded from multiple individuals). Panel (b) shows the Fourier amplitude spectra of the waveforms in (a), with clear peaks at the signal frequency (5Hz) in each example. The Fourier spectrum also includes a phase term, which can be represented in polar coordinates (amplitude and phase); panel (c) shows this for example responses at the signal frequency (grey points) and their average (blue point), which is computed separately for amplitude and phase terms. Alternatively, the same information is contained in a Cartesian representation of the real and imaginary parts of the complex spectrum (panel (d)). Notice that the individual observations (grey points) are the same in panels (c,d), but the different representations have implications for how the average (blue points) and measures of spread (shaded regions) are calculated.

3 Fourier amplitudes violate parametric assumptions of univariate statistics

Many empirical studies use univariate T-tests or analysis of variance (ANOVA) to analyse periodic data. Specifically, the amplitude component of the Fourier spectrum at the stimulation frequency is used as the dependent variable, discarding the phase information. This is problematic, because the amplitude is an absolute quantity, and can never fall below zero. Distributions of amplitudes for weak signals are therefore positively skewed, and will generally violate the assumption of normality.
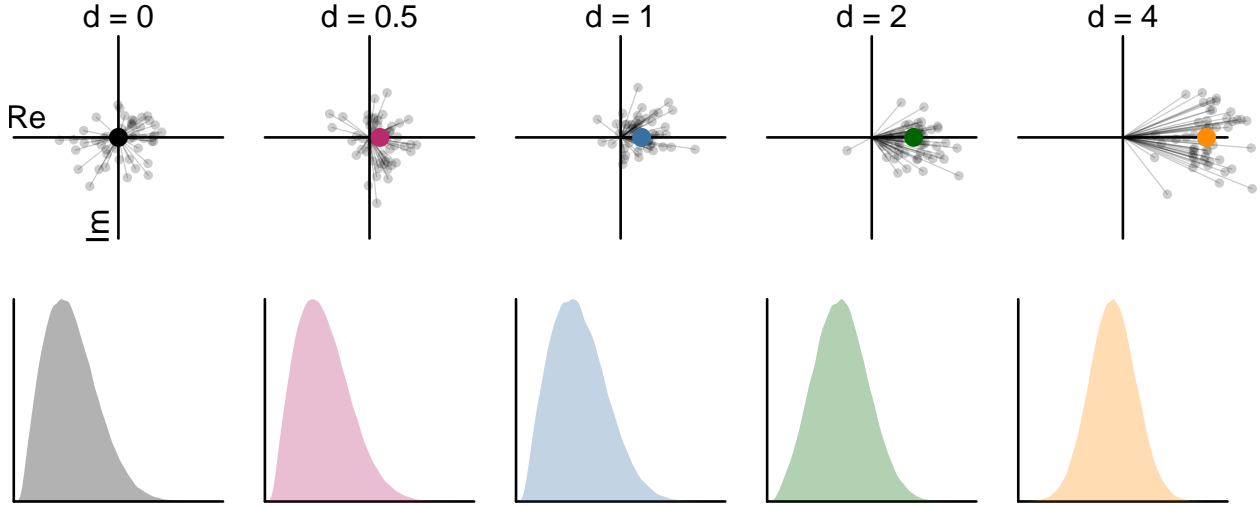

Figure 2: Demonstration of skew in absolute Fourier amplitudes for signals of different strengths. Signal strength is quantified as Cohen's d, defined as the ratio of the mean to the standard deviation of the sample. The upper row shows samples of 50 grey points, and the population mean (coloured points). The lower row shows kernel density functions generated from 100,000 amplitude values. Note that the mean phase of the signal is irrelevant for these simulations, and is shown in the positive x-direction for consistency.

The upper row of Figure 2 shows scatterplots of simulated Fourier components, expressed using real (x) and imaginary (y) components. The amplitudes are the lengths of the lines joining each grey point to the origin. The lower row in Figure 2 shows distributions of amplitudes for the same set of signal strengths. These distributions only approach normality when the signal strength is more than twice the standard deviation (Cohen's $d > 2$; Cohen's d is the mean difference scaled by the standard deviation, see Cohen (1988)). One consequence of this is that T-tests will potentially have an inflated Type I error (false positive) rate for many signals encountered empirically, especially if used to make pointwise comparisons to an amplitude of 0.

Typical solutions, such as log-transforming the data, are unlikely to be equally applicable to all conditions. For example, if one wishes to compare a baseline where no stimulus was presented with a condition involving a strong signal, the former will be skewed and the latter normal. Applying a transform to both conditions is therefore problematic. Non-parametric statistics are a potential option, but these have generally lower statistical power than their parametric equivalents. Instead, the bivariate statistics discussed in the introduction avoid these issues and have greater statistical power, and should be used in preference to univariate methods.

4 Conditions under which T^2_{circ} is more sensitive than T^2

Using the T^2 or T^2_{circ} statistic allows the phase information to be retained, and therefore provides greater power than univariate T-tests, as well as avoiding problems caused by using absolute amplitude values. Victor & Mast (1991) report simulations showing situations where T^2_{circ} has greater power than T^2. This

involved generating random data sets of different sample sizes, and different signal strengths, and comparing the number of such tests where each statistic was significant. Their simulations show the largest advantage for T^2_{circ} for effect sizes around $d = 1$ (where the mean signal strength is equal to the standard deviation of the data). The advantage appeared to be stronger for smaller sample sizes.

Here these simulations are replicated and extended (see Figure 3), and it is shown that the regime where T^2_{circ} has greater power occurs particularly for large effect sizes and small sample sizes (see Figure 3f). However, for effect sizes around $0.5 < d < 1$, T^2_{circ} is more sensitive even with around 16 observations. This advantage is lost for large sample sizes (N > 32) and large effect sizes (when $d > 2$ and N > 8). These simulations suggest a straightforward heuristic - there is no advantage to using the T^2_{circ} statistic for large sample sizes (when N > 32), so its use should be restricted to small sample studies.
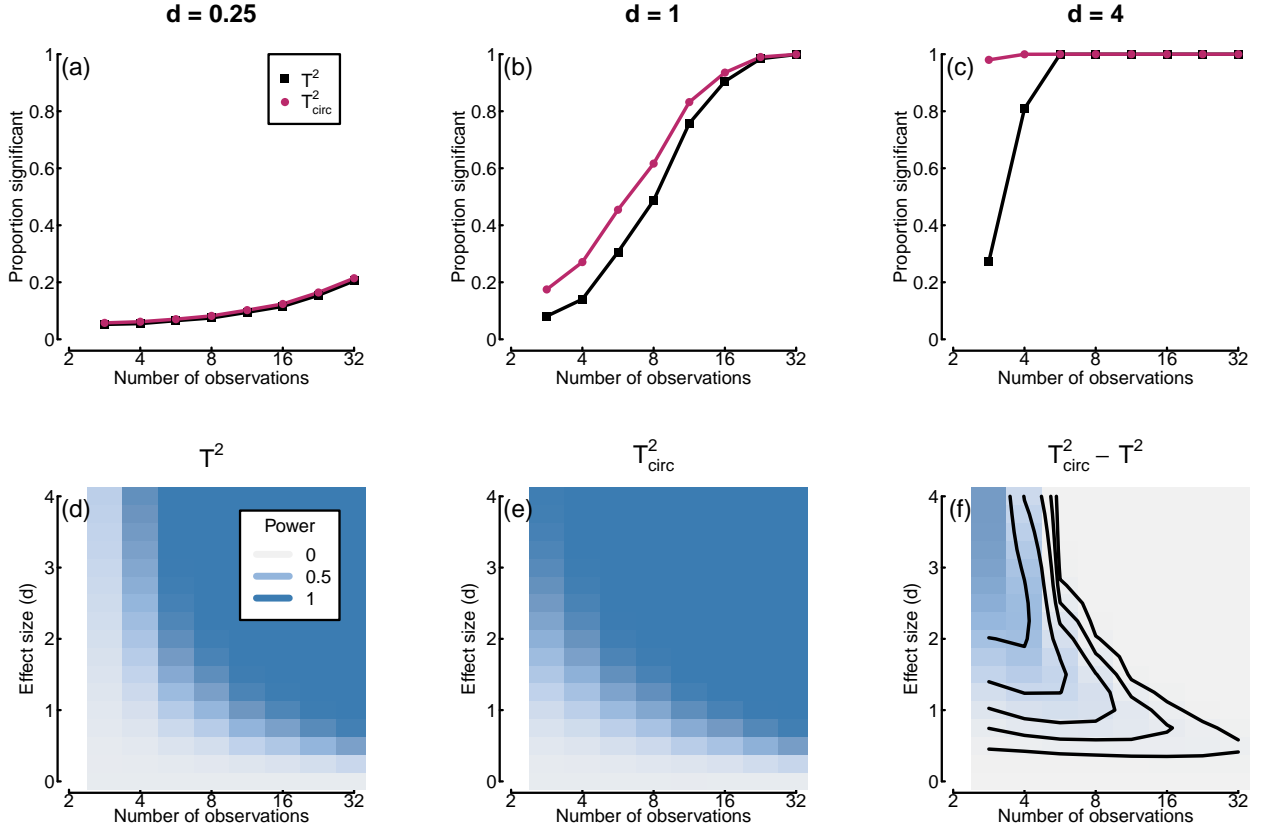

Figure 3: Simulations estimating the proportion of significant tests for simulated data with different sample sizes and effect sizes (100,000 simulated data sets per condition). Panels a-c replicate conditions reported by Victor and Mast (1991). Panels (d) and (e) show a wider range of conditions for each statistic. Panel (f) shows the difference between the two statistics, with contour lines indicating differences of 0.02, 0.05, 0.1, 0.2 and 0.4.

5 Limitations of T^2_{circ} when assumptions are violated

Although T^2_{circ} can be more sensitive than T^2, this greater sensitivity relies on satisfying the T^2_{circ} test's more stringent assumptions. The two variables must be independent (i.e. uncorrelated), and of equal variance. These restrictions may hold for some data sets, but it is instructive to ask what happens when they do not. Figure 4 shows the results of simulations with randomly generated bivariate data in which no signal is present. When the data are uncorrelated and have equal variance (mid-points of the functions in each panel), both tests have the nominal Type I error (false positive) rate of $\alpha = 0.05$ (horizontal dashed lines). However,

as the data become increasingly correlated (Figure 4a), or the variances of the two dependent variables more disparate (Figure 4b), the Type I error rate of the T^2_{circ} statistic (shown in red) increases by almost a factor of 2. In contrast, the T^2 statistic, which explicitly takes account of the covariance matrix (see equation (1)) shows no increase (black curves).

One possible remedy to control the Type I error rate would be to adjust either the alpha level or the degrees of freedom (as is done in repeated measures ANOVA when sphericity assumptions are violated). However, this will reduce the statistical power of the T^2_{circ} test, and its advantage over T^2 is relatively marginal in most situations to begin with (see Figure 3). What is required is a method to objectively assess whether the assumptions of T^2_{circ} hold; this is developed in the following section.

6 A novel method to test the assumptions of T^2_{circ}

Despite the severe consequences of violating the assumptions of the T^2_{circ} statistic (see Figure 4), there is currently no accepted test of those assumptions that could be applied to an empirical data set. Victor and Mast (1991) suggest that their test should be applicable to multiple repetitions of a stimulus condition collected from a single participant, whereas data pooled across multiple participants may be less likely to exhibit independence of the real and imaginary components (see also Pei et al., 2017). However it would be useful to develop a method that can tell us whether the assumptions hold for a given data set.

One convenient way to test the assumptions of T^2_{circ} is to assess the *condition index* of a data set, which describes the ratio of eigenvalues for a cloud of points. The eigenvectors are the major and minor axes of the bounding ellipse (the straight lines in the example icons at the foot of Figure 4a,b). Conventionally, the condition index is calculated as the square root of the longest/shortest eigenvector length. For uncorrelated random numbers the expected distribution of condition indices is positively skewed, with a minimum of 1 (Edelman, 1988). This is because two independent samples of numbers from the same underlying distribution will generally by chance have unequal eigenvectors, and the definition of the condition index ($\sqrt{longest/shortest}$) prevents its values dropping below 1 (values < 1 would imply that the shortest eigenvector is longer than the longest one). For bivariate data, Edelman (1988) provides an equation (his Eq. 14) for the probability density function of condition indices (x) as a function of sample size (N):

$$pdf = (N-1)2^{N-1}\frac{x^2 - 1}{(x^2 + 1)^N}x^{(N-2)}, \qquad (3)$$

Attempts to validate this by simulation suggest that for small sample sizes ($N < 10$) a closer approximation is given by:

$$pdf = (N-2)2^{N-2}\frac{x^2 - 1}{(x^2 + 1)^{(N-1)}}x^{(N-3)}. \qquad (4)$$

Figure 5a shows example distributions derived from both expressions (black and red curves), and simulations from 100,000 random data sets with $N = 4$ (blue shading). The vertical lines show the critical (95%) threshold for the analytic and simulated results. A ratio lying beyond this threshold can be considered to violate the assumption of either independence or equal variance, because it has a condition index larger than expected by chance (assuming $\alpha = 0.05$). Ratios below the threshold imply that the eigenvalues can be considered equal (in a statistical sense). Figure 5b shows how these critical thresholds change as a function of the number of observations (N), and it appears that the modified expression (red) most closely approximates the simulation results (blue).

The eigenvalue ratio can be used as a test of the assumptions of T^2_{circ}. If a condition index is observed that is above the critical threshold for the number of observations, then the data set can be said to significantly violate the assumption of equal eigenvalues. Because the modified equation permits estimation of an inverse density function, this can be used to calculate a *p*-value for the test. If the test is non-significant, one
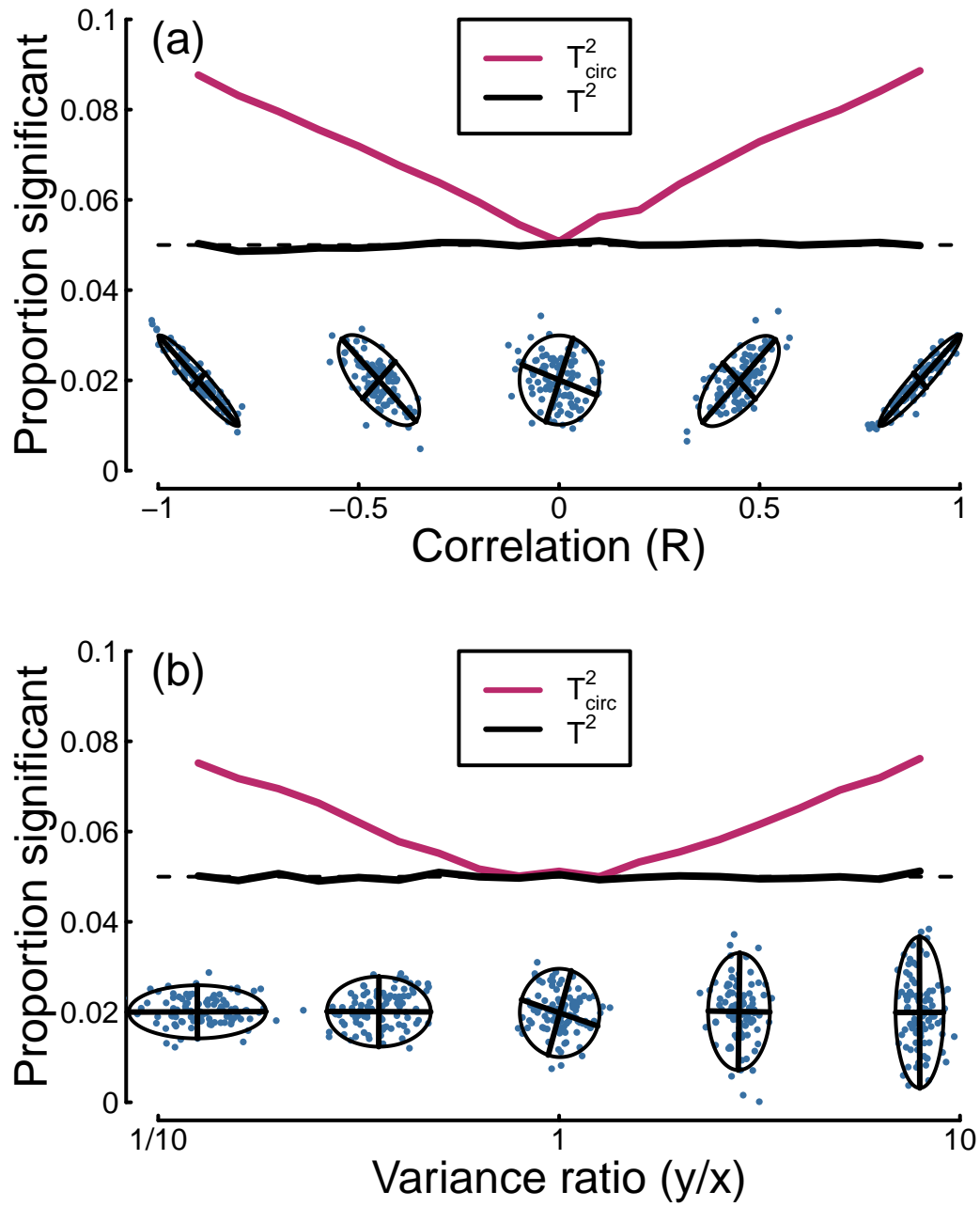
Figure 4: Simulations showing the Type I error rate for both tests as a function of the correlation between two variables (a) and the ratio of variances (b). Estimates are for 100,000 simulated data sets per condition, with N = 10 observations. The icons at the foot of each panel show example scatterplots with bounding ellipses and eigenvectors.
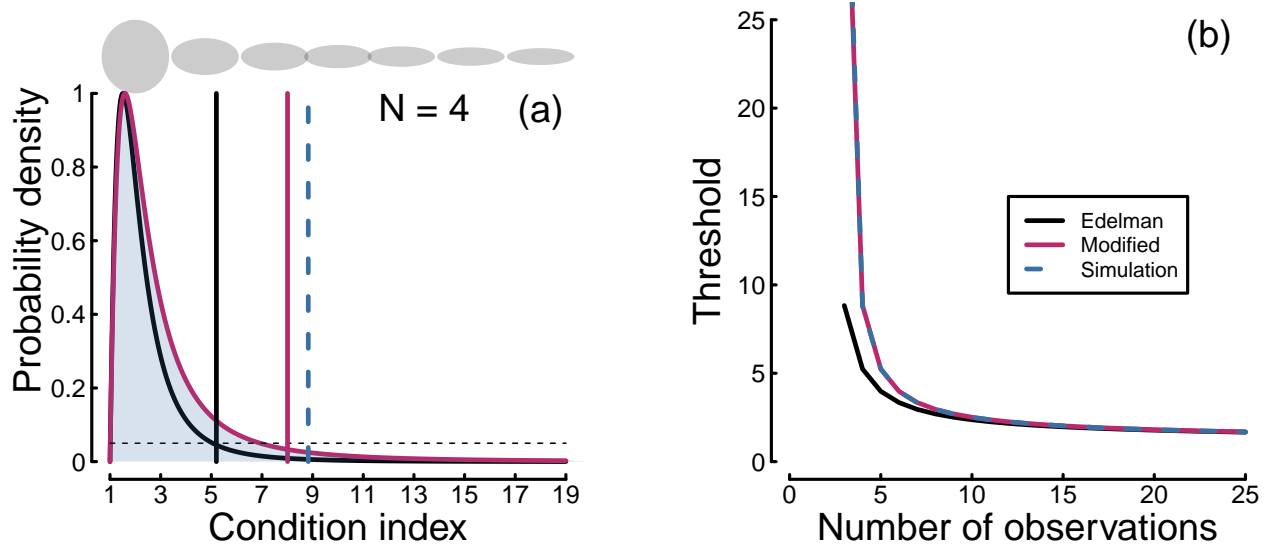
Figure 5: Logic of the condition index test. Panel (a) shows the distribution of condition indices derived from Equations 3 (black curve) and 4 (red curve), and by stochastic simulation (blue shading), for a sample size of N = 4 observations. The vertical lines show the 95 percent thresholds on the distributions (where 95 percent of values lie to the left of the line). Ellipse icons above panel (a) illustrate different condition indices between 1 and 19 (note that rotation does not affect the condition index). Panel (b) shows how 95 percent thresholds change as a function of the number of observations.

can proceed with T^2_{circ}; if it is significant, T^2 should be used instead. A function implementing this test is included in the *FourierStats* package (the *CI.test* function in *R*, and *CI_test* function in *Matlab*).

7 Identifying and removing outliers using the Mahalanobis distance

If a data set produces a significant result using the condition index test, this could be due to the presence of one or more outliers. The Mahalanobis distance (Mahalanobis, 1936) is a useful metric for identifying such multivariate outliers so that they can be excluded. It calculates the Euclidean distance between each data point and the centroid, and scales it by the variance in the direction of the vector that joins the two points. This means that any correlations in the data set are taken into account when calculating the distance metric, *D*.

The effectiveness of this approach to outlier exclusion can be assessed by simulation using the condition index test. Figure 6 shows the proportion of significant condition index tests as a function of the Mahalanobis distance of a single outlier, for a range of sample sizes (curves). In all cases, the functions depart from the Type I error rate ($\alpha = 0.05$; horizontal dashed line in Figure 6) when the outlier's Mahalanobis distance exceeds a value around 3. This seems a reasonable heuristic for outlier exclusion, and is the multivariate equivalent of excluding data points more than 3 standard deviations from the mean (note that many implementations of the Mahalanobis distance statistic, such as the core *mahalanobis* function in *R* (or *mahal* function in *Matlab*), return D^2, which can be converted to D by taking the square root). Following this heuristic should reduce the likelihood that outliers will invalidate the assumptions of the T^2_{circ} test.

A variant of the Mahalanobis distance (the pairwise Mahalanobis distance) can also be used to compute a multivariate measure of effect size, equivalent to Cohen's d statistic (see e.g. Del Giudice, 2009). This is a valuable statistic to include when reporting the results of multivariate tests, and a function (*pairwisemahal*) to calculate it is available as part of the *FourierStats* package.
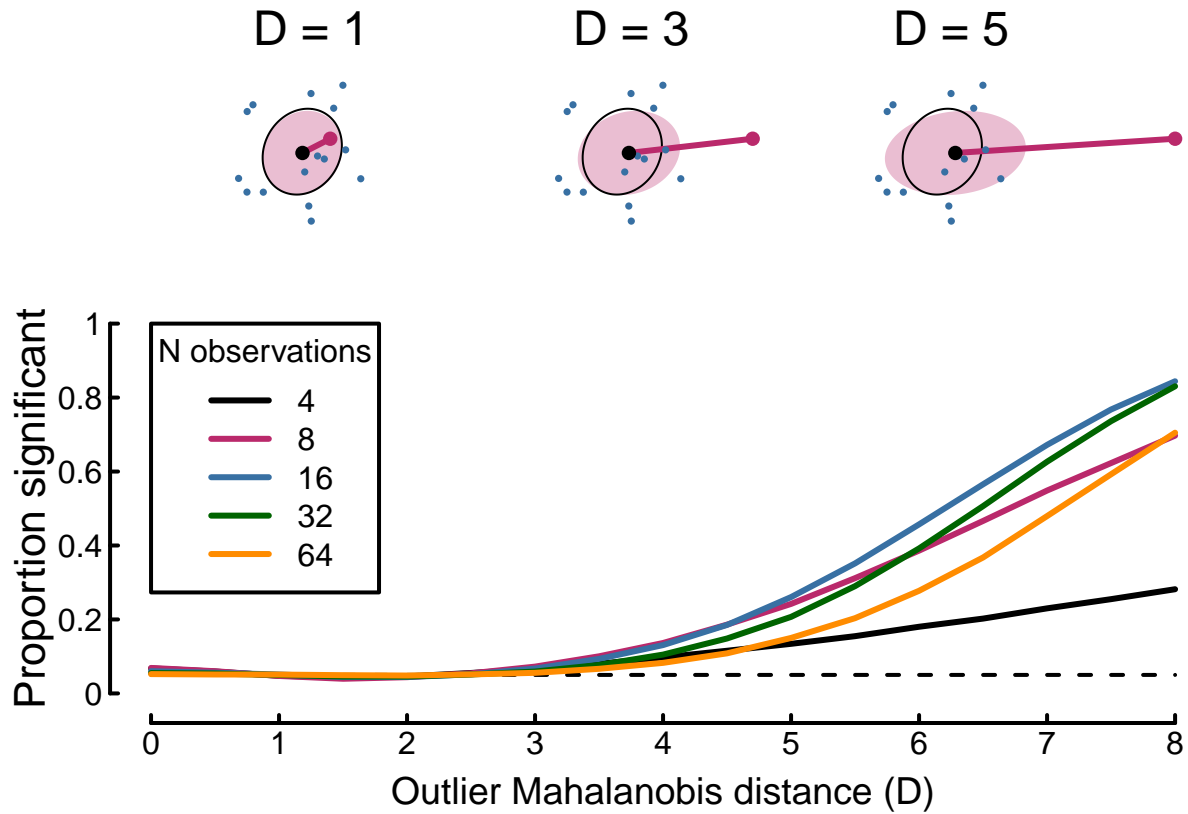
Figure 6: Simulations illustrating the Mahalanobis distance metric, and showing how a single outlier affects the condition index. The upper row shows three example data sets, each with a single outlier shown in red. The outliers have approximate Mahalanobis distances of 1, 3 and 5. The ellipses are calculated with the outlier included (red) and excluded (black), illustrating how the outlier distorts the aspect ratio of the ellipse. The main plot shows how the proportion of significant condition index tests depends on the outlier distance and the sample size.

8 Controlling for multiple comparisons across location and time

In some studies, it is important to compare responses over space and/or time. However with large numbers of sensors, voxels or temporal epochs, the familywise error quickly becomes problematic, inflating the Type I error (false positive) rate. Solutions such as Bonferroni correction, which adjust the α level based on the number of comparisons, are overly conservative and can obscure real effects by dramatically reducing power. An alternative approach is to use cluster correction methods to control the Type I error rate, using mass univariate tests (e.g. Maris and Oostenveld, 2007). These typically involve summing test statistics such as T-values across adjacent significant locations and/or moments in time. The summed test statistic is compared to a null distribution generated from the same data by randomly permuting condition labels (or the sign of the data for one-sample tests). Such methods control the Type I error rate without substantially reducing statistical power. The same approach can be applied to the T^2 and T^2_{circ} statistics. This allows a principled method for identifying clusters of significant sensors, timepoints or frequencies responding to periodic stimuli. The *FourierStats* package includes an implementation of this method with options for multivariate statistics (the *clustercorrect* function).

9 Generalising to more than two conditions

Many studies involve more than two experimental conditions that need to be compared. Again, issues with familywise error will quickly become problematic if multiple pairwise T^2 or T^2_{circ} statistics are calculated. One possibility is to conduct a MANOVA, which takes covariances between dependent variables into account in much the same way as Hotelling's T^2, but permits independent variables with more than two levels, as well as factorial designs. However, if the assumptions of T^2_{circ} hold for a data set, it might alternatively be possible to extend the logic of the T^2_{circ} test (Victor and Mast, 1991) to the more general case, and obtain a sensitivity benefit similar to that shown in Figure 3.

The F-statistic for a one-way ANOVA is calculated by taking the ratio between the variance explained by the modelled group means, and the residual unexplained variance. For multivariate data, the change in group mean would be calculated using the vector distances between the complex means, and the residuals are the vector distances between each data point and its corresponding group mean. For an independent one-way design with k groups (or conditions) and N observations per group, the F-distribution will have $2(k-1)$ and $2((Nk)-k)$ degrees of freedom. The difference between traditional univariate ANOVA is the factor of two scaling, which accounts for the additional degree of freedom for each complex-valued number. A p-value can then be derived from the F-distribution in the usual way.

A suitable name for such a test might be $ANOVA^2_{circ}$, as this reflects the similarity to ANOVA, and the extension of the logic of T^2_{circ} (an alternative name might be $MANOVA_{circ}$, however this feels less appropriate given that many of the key features of MANOVA are absent). Figure 7 shows simulations analogous to those in Figure 3 for a one-way between-subjects design with three levels. MANOVA is directly compared to the $ANOVA^2_{circ}$ statistic across a range of effect sizes and sample sizes. Just as for the one-sample statistics, the advantages of $ANOVA^2_{circ}$ are particularly apparent for small sample sizes, and larger effect sizes (Figure 7f).

Following a significant $ANOVA^2_{circ}$ test, one could calculate T^2_{circ} statistics to make post-hoc pairwise comparisons between conditions, providing that appropriate multiple comparison correction is applied (e.g. Bonferroni correction). Univariate ANOVAs on the real and imaginary components are unlikely to be informative, as the relative magnitudes depend on stimulus phase (which is arbitrary; see Figure 1d). A repeated measures version of $ANOVA^2_{circ}$ can also be implemented following the same logic. The project repository contains a function (*anovacirc.test* in *R*, *anovacirc_test* in *Matlab*) to run both of these tests for one-way designs. In principle factorial versions might also be derived.
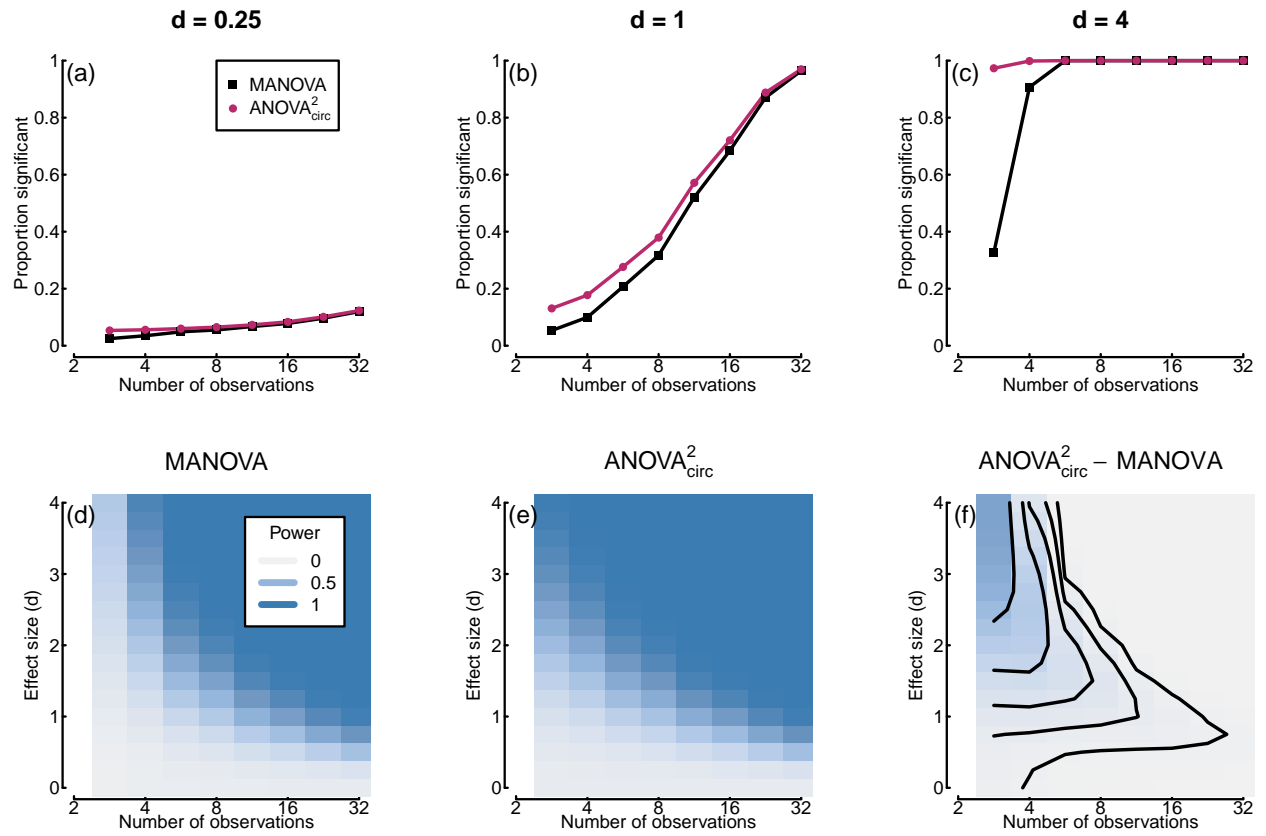
Figure 7: Simulations comparing the sensitivity of MANOVA and ANOVA2circ. The format mirrors that of Figure 3. In these simulations, there were three conditions, with the signal being added to one condition only.

10   Deciding which test to run

The flowchart in Figure 8 contains a proposed decision structure for the analysis of periodic data, once any outliers have been removed. Initially, data for each condition should be tested against the expected distribution of eigenvalue ratios using the condition index test. Comparisons with one or two conditions should be tested with the T^2_{circ} statistic if the eigenvalue ratios are consistent with circularity, and the T^2 statistic if they are not. Comparisons with more than two conditions should be tested with the $ANOVA^2_{circ}$ statistic if the eigenvalue ratios are consistent with circularity, or a MANOVA if not. Many MANOVA implementations cannot deal correctly with random factors (repeated measures), particularly in complex factorial designs. However the *multRM* function in the *MANOVA.RM* package (Friedrich et al., 2019) is able to handle such designs appropriately using a bootstrapping approach.
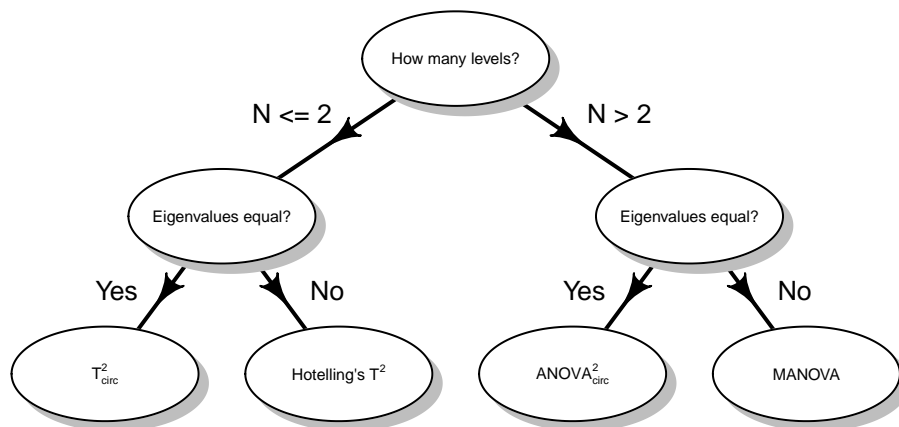

Figure 8: Flowchart illustrating how one might decide which test to conduct for a given data set, based on the study design and the outcome of the condition index test.

11   Applying multivariate methods to empirical data sets

Having developed some novel tools for the analysis of periodic data, in the following sections their use is demonstrated for two different publicly available empirical data sets. The first study recorded responses to auditory and optogenetic stimulation in mice. The second study measured visual responses to flickering grating patterns in humans. These examples also provide a demonstration of how the results of the tests described above might be appropriately reported.

11.1   Mouse auditory and optogenetic steady-state data

Hwang et al. (2019, 2020) measured steady-state responses using implanted scalp electrodes in 6 mice. The mice had previously been given a targeted virus that made parvalbumin neurons in their basal forebrain responsive to specific wavelengths of light, delivered through an optical fiber (a technique called optogenetics). Steady-state evoked potentials were recorded from 36 electrodes for 1 second epochs of 40 Hz auditory stimulation, and various schedules of optogenetic stimulation (including at 40 Hz). The data set is described more fully by Hwang et al. (2020), and was downloaded from: https://doi.gin.g-node.org/10.12751/g-node.e5tyek/. A processed data file is included with permission from the authors in the *FourierStats* package.

Figure 9a shows Fourier amplitude spectra at two frontal electrodes (marked black in the insets) for auditory stimulation (red) and optogenetic stimulation (blue), each at 40 Hz. There is a clear frequency-locked signal with approximately equal amplitude for each stimulation modality. Indeed, a paired univariate T-test on the amplitudes reveals no significant difference ($t = 0.84$, $df = 5$, $p = 0.44$). However, inspection of the complex
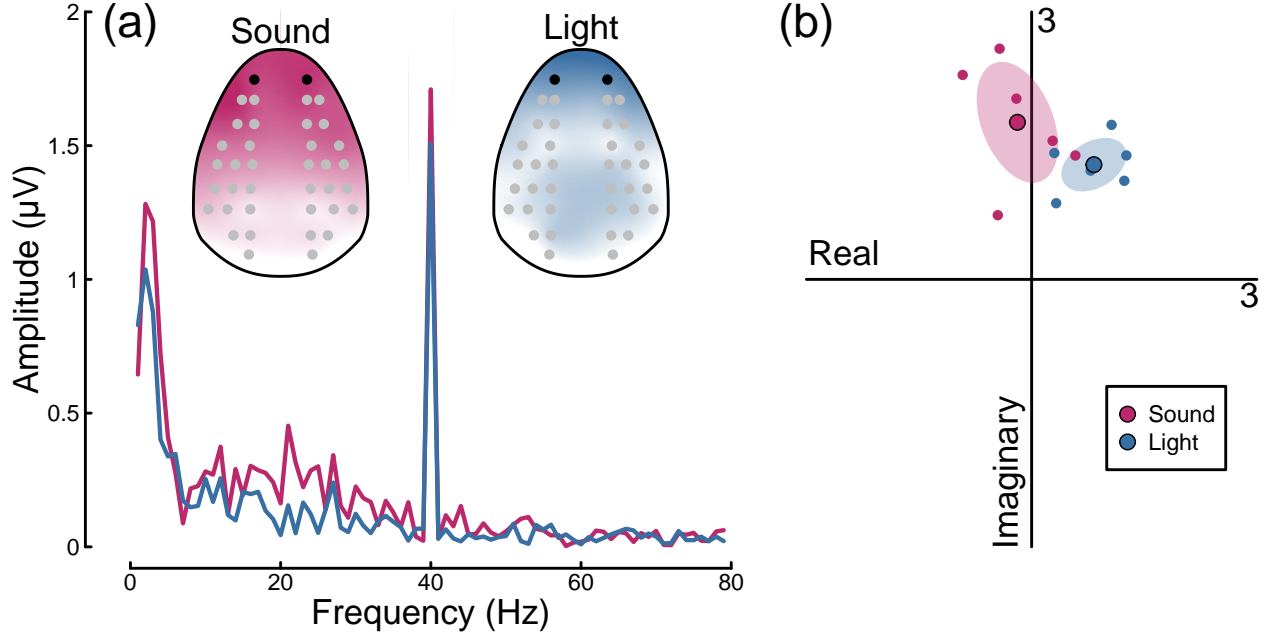
Figure 9: Summary of mouse steady-state responses to 40 Hz stimulation. Panel (a) shows the Fourier amplitude spectrum with inset scalp plots for sound (red) and light (blue) stimulation, averaged across repetitions and individuals. Grey and black points in the insets indicate electrode locations. Panel (b) shows complex (x = real, y = imaginary) Fourier components for 6 individual mice (small points) and their average (large points) for both conditions.

Fourier components for each condition suggests evidence of a phase difference between the two modalities (see Figure 9b). The condition index test was non-significant for both conditions (sound: CI = 1.59, $p = 0.66$; light: CI = 1.69, $p = 0.59$), so a paired-samples T^2_{circ} test was conducted. This revealed a significant difference between conditions ($T^2_{circ} = 1.39$, $F_{(2,10)} = 8.32$, $p = 0.007$) with an effect size of $D = 2.14$. This demonstrates that both sound and (optogenetic) light are able to entrain neural responses - the original study by Hwang et al. (2019) went on to explore interactions between these two signals.

11.2 Human visual steady-state data

Vilidaite et al. (2018) measured visual responses to flickering grating stimuli in a large sample of 100 adults. Each participant completed a series of 11-second trials, in which stimuli of different contrasts flickered at 7Hz (on-off sinusoidal flicker). Responses were strongest at occipital electrodes over visual cortex (see upper row of Figure 10), were well-isolated in the Fourier domain, and increased with stimulus contrast. Significant activity was evident at 4% contrast and above following cluster correction (with very stringent alpha levels given the high power of this data set), as indicated by the red electrodes in the upper row of Figure 10. For the main analysis, responses were taken from electrode Oz at the occipital pole (black points in the upper row of Figure 10), and averaged across repetition for each participant. Each condition included some outlier points with Mahalanobis distances exceeding 3, marked red in the lower row of Figure 10. Any participant that contributed at least one outlier was excluded, leaving a total of 89 participants for the main analysis.

With the outlier points removed, all seven conditions resulted in non-significant condition index tests (largest CI = 1.20, all $p > 0.23$). A repeated measures $ANOVA^2_{circ}$ test was conducted, revealing a significant effect of stimulus contrast ($F_{(12,1056)} = 38.9$, $p < 0.001$). Pairwise T^2_{circ} statistics comparing the baseline (0% contrast) condition to each subsequent condition (Bonferroni corrected for 6 tests to $\alpha = 0.08$) revealed significant differences at 8% contrast ($T^2_{circ} = 0.32$, $F_{(2,176)} = 28.43$, $D = 1.1$, $p < 0.001$), 16% contrast ($T^2_{circ} = 0.28$, $F_{(2,176)} = 25.25$, $D = 1.04$, $p < 0.001$), 32% contrast ($T^2_{circ} = 0.10$, $F_{(2,176)} = 8.55$, $D =$
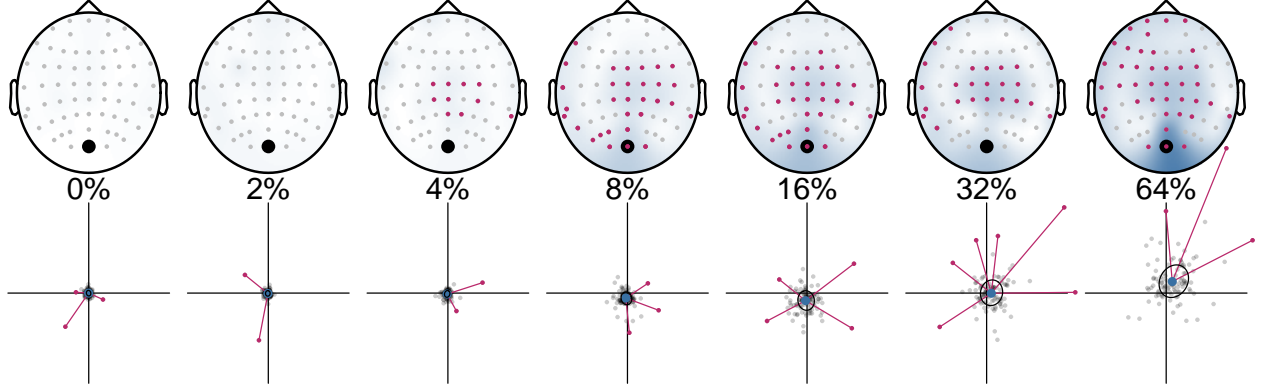
13

Figure 10: Summary of human SSVEP data. Upper row shows scalp distributions of Fourier amplitudes at 7Hz for stimuli of increasing contrasts (blue shading indicates higher amplitudes). Electrodes marked in red indicate cluster-corrected significance. Lower row shows scatterplots of complex (x = real, y = imaginary) Fourier components for 100 participants per condition, from electrode Oz (black point). Red points are outliers with Mahalanobis distances exceeding 3, and blue points mark the centroids.

0.63, $p < 0.001$) and 64% contrast ($T^2_{circ} = 0.40$, $F_{(2,176)} = 35.35$, $D = 1.17$, $p < 0.001$). The study by Vilidaite et al. (2018) compared SSVEP responses between individuals with and without autism, as well as in a *Drosophila* genetic model of developmental disorders. The raw data are available at: http://dx.doi.org/10.17605/OSF.IO/Y4N5K, and a processed version is included with the *FourierStats* package.

12 Further considerations

It is worth stating explicitly that the statistical tests discussed in this paper are applicable only when the signal phase is expected to be consistent across observations. This is the case for most paradigms in which the nervous system is driven by a periodic stimulus. However, they are less obviously applicable to the analysis of endogenous neural oscillations and brain rhythms (Berger, 1929; Buzsáki and Draguhn, 2004), which will typically have random phase and broader bandwidths in the Fourier domain: other analysis methods have been developed for such signals (e.g. Canolty and Knight, 2010). When phases are consistent across repetitions, greater statistical power can be obtained by coherently averaging across repetitions to obtain a participant-level average (Baker et al., 2021). This practice is also necessary in order to use the multivariate methods discussed here, as the alternative is to discard the phase information and average amplitudes instead, rendering the data univariate.

The present paper has focussed on the Frequentist statistical tradition. However there are many advantages to the Bayesian approach, in which one can make direct quantitative comparisons of the evidence supporting both the experimental and null hypotheses (Jeffreys, 1961). Subject to determining appropriate priors, Bayes factor scores might be calculated for all of the statistics considered here, much as has been done previously for univariate T-tests (Rouder et al., 2009) and ANOVA (Rouder et al., 2017). However this is a non-trivial undertaking, and is beyond the scope of the current paper.

Another possibility is to use machine learning techniques such as multivariate pattern analysis (MVPA) to analyse periodic data. This involves training a classifier algorithm to distinguish between two (or more) experimental conditions or states, and then assessing classifier accuracy for predicting the group labels of fresh data. If different conditions produce distinct patterns of neural response, then classifier accuracy will be above chance. Such methods have been hugely influential in the fMRI literature (Schwarzkopf and Rees, 2011), and for analysing event-related potential data collected using EEG or MEG (Grootswagers et al., 2017). However, they have not been widely applied to steady-state data (though see West et al. (2015) for one example). In principle, the real and imaginary Fourier components can be treated as separate dependent variables, along with different recording locations and/or frequencies. This approach has the potential to

offer sensitive, high-powered statistical tests that circumvent many of the shortcomings associated with traditional statistics.

Even when statistics are conducted using both the real and imaginary Fourier components, it is still typical to visualise the mean amplitudes. Several approaches to calculating appropriate error bars have been proposed. For example, Pei et al. (2017) suggest calculating the nearest and farthest points from the origin on the bounding ellipse, and using these to derive standard errors for the amplitude. This approach is somewhat demanding, though a function is provided to calculate error bars using this method (available through the *amperrors* function). However, as an alternative, bootstrap resampling offers a powerful and general method for calculating confidence intervals on amplitudes. This is achieved by resampling the complex data (with replacement) and calculating a resampled complex mean. The amplitude is then derived for this resampled mean, and the procedure repeated a large number of times (1000 or 10000 repetitions is typical) to build up a population of resampled mean amplitudes. Upper and lower confidence intervals on the amplitude can then be taken at appropriate quantiles of this population (68% and 95% are typical, with the 68% interval corresponding to the standard error).

13   General recommendations for analysing periodic data

The simulations reported here allow several recommendations to be made for how periodic data should be analysed. Multivariate statistics should be used for phase-locked Fourier data instead of univariate statistics such as T-tests and ANOVA. This avoids problems from non-normal distributions of amplitudes violating the test assumptions, and also provides a sensitivity benefit from the inclusion of phase information. Outliers should be removed when they have a Mahalanobis distance exceeding 3, and the pairwise Mahalanobis distance reported as a measure of effect size. For sample sizes of N < 32, the T^2_{circ} and $ANOVA^2_{circ}$ statistics can be used if the condition index test is non-significant for all conditions. Alternatively, the T^2 or MANOVA statistics should be used when these conditions are not met. The greater power afforded by these tests should in general lead to more accurate statistical inferences when analysing periodic data.

References

Afsari F, Christensen KV, Smith GP, Hentzer M, Nippe OM, Elliott CJH, Wade AR. 2014. Abnormal visual gain control in a Parkinson's disease model. *Hum Mol Genet* **23**:4465–78. doi:10.1093/hmg/ddu159

Baker DH, Vilidaite G, Lygo FA, Smith AK, Flack TR, Gouws AD, Andrews TJ. 2021. Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*. doi:10.1037/met0000337

Berger H. 1929. Über das elektrenkephalogramm des menschen. *Archiv f Psychiatrie* **87**:527–570. doi:10.1007/BF01797193

Braddick OJ, Wattam-Bell J, Atkinson J. 1986. Orientation-specific cortical responses develop in early infancy. *Nature* **320**:617–9. doi:10.1038/320617a0

Buzsáki G, Draguhn A. 2004. Neuronal oscillations in cortical networks. *Science* **304**:1926–9. doi:10.1126/science.1099745

Canolty RT, Knight RT. 2010. The functional role of cross-frequency coupling. *Trends Cogn Sci* **14**:506–15. doi:10.1016/j.tics.2010.09.001

Cohen J. 1988. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum, Hillsdale, New Jersey.

Del Giudice M. 2009. On the real magnitude of psychological sex differences. *Evolutionary Psychology* **7**:264–279. doi:10.1177/147470490900700209

Edelman A. 1988. Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications* **9**:543–560. doi:10.1137/0609045

Engel SA, Rumelhart DE, Wandell BA, Lee AT, Glover GH, Chichilnisky EJ, Shadlen MN. 1994. fMRI of human visual cortex. *Nature* **369**:525. doi:10.1038/369525a0

Enroth-Cugell C, Robson JG. 1966. The contrast sensitivity of retinal ganglion cells of the cat. *J Physiol* **187**:517–52. doi:10.1113/jphysiol.1966.sp008107

Friedrich S, Konietschke F, Pauly M. 2019. Resampling-Based Analysis of Multivariate Data and Repeated Measures Designs with the R Package MANOVA.RM. *The R Journal* **11**:380–400. doi:10.32614/RJ-2019-051

Gray KLH, Flack TR, Yu M, Lygo FA, Baker DH. 2020. Nonlinear transduction of emotional facial expression. *Vision Res* **170**:1–11. doi:10.1016/j.visres.2020.03.004

Grootswagers T, Wardle SG, Carlson TA. 2017. Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *J Cogn Neurosci* **29**:677–697. doi:10.1162/jocn_a_01068

Hotelling H. 1931. The generalization of Student's ratio. *The Annals of Mathematical Statistics* **2**:360–378.

Hwang E, Brown RE, Kocsis B, Kim T, McKenna JT, McNally JM, Han H-B, Choi JH. 2019. Optogenetic stimulation of basal forebrain parvalbumin neurons modulates the cortical topography of auditory steady-state responses. *Brain Struct Funct* **224**:1505–1518. doi:10.1007/s00429-019-01845-5

Hwang E, Han H-B, Kim JY, Choi JH. 2020. High-density EEG of auditory steady-state responses during stimulation of basal forebrain parvalbumin neurons. *Sci Data* **7**:288. doi:10.1038/s41597-020-00621-z

Jeffreys H. 1961. Theory of probability, 3rd ed. Oxford University Press.

Liu-Shuang J, Norcia AM, Rossion B. 2014. An objective index of individual face discrimination in the right occipito-temporal cortex by means of fast periodic oddball stimulation. *Neuropsychologia* **52**:57–72. doi:10.1016/j.neuropsychologia.2013.10.022

Mahalanobis P. 1936. On the generalised distance in statistics. *Proceedings of the National Academy of Sciences of India* **2**:49–55.

Maris E, Oostenveld R. 2007. Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods* **164**:177–90. doi:10.1016/j.jneumeth.2007.03.024

Morrone MC, Burr DC, Speed HD. 1987. Cross-orientation inhibition in cat is GABA mediated. *Exp Brain Res* **67**:635–44. doi:10.1007/BF00247294

Nakayama K, Mackeben M. 1982. Steady state visual evoked potentials in the alert primate. *Vision Res* **22**:1261–71. doi:10.1016/0042-6989(82)90138-9

Norcia AM, Appelbaum LG, Ales JM, Cottereau BR, Rossion B. 2015. The steady-state visual evoked potential in vision research: A review. *J Vis* **15**:4. doi:10.1167/15.6.4

Pei F, Baldassi S, Tsai JJ, Gerhard HE, Norcia AM. 2017. Development of contrast normalization mechanisms during childhood and adolescence. *Vision Res* **133**:12–20. doi:10.1016/j.visres.2016.03.010

Porciatti V, Fontanesi G, Raffaelli A, Bagnoli P. 1990. Binocularity in the little owl, athene noctua. II. Properties of visually evoked potentials from the wulst in response to monocular and binocular stimulation with sine wave gratings. *Brain Behav Evol* **35**:40–8. doi:10.1159/000115855

Rees A, Green GG, Kay RH. 1986. Steady-state evoked responses to sinusoidally amplitude-modulated sounds recorded in man. *Hear Res* **23**:123–33. doi:10.1016/0378-5955(86)90009-2

Regan D. 1966. Some characteristics of average steady-state and transient responses evoked by modulated light. *Electroencephalogr Clin Neurophysiol* **20**:238–48. doi:10.1016/0013-4694(66)90088-5

Rouder JN, Morey RD, Verhagen J, Swagman AR, Wagenmakers E-J. 2017. Bayesian analysis of factorial designs. *Psychol Methods* **22**:304–321. doi:10.1037/met0000057

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev* **16**:225–37. doi:10.3758/PBR.16.2.225

Schwarzkopf DS, Rees G. 2011. Pattern classification using functional magnetic resonance imaging. *Wiley Interdiscip Rev Cogn Sci* **2**:568–579. doi:10.1002/wcs.141

Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, Rosen BR, Tootell RB. 1995. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* **268**:889–93. doi:10.1126/science.7754376

Snyder AZ. 1992. Steady-state vibration evoked potentials: Descriptions of technique and characterization of responses. *Electroencephalogr Clin Neurophysiol* **84**:257–68. doi:10.1016/0168-5597(92)90007-x

Spitschan M, Jain S, Brainard DH, Aguirre GK. 2014. Opponent melanopsin and s-cone signals in the human pupillary light response. *Proc Natl Acad Sci U S A* **111**:15568–72. doi:10.1073/pnas.1400942111

Victor JD, Mast J. 1991. A new statistic for steady-state evoked potentials. *Electroencephalogr Clin Neurophysiol* **78**:378–88. doi:10.1016/0013-4694(91)90099-p

Vilidaite G, Norcia AM, West RJH, Elliott CJH, Pei F, Wade AR, Baker DH. 2018. Autism sensory dysfunction in an evolutionarily conserved system. *Proc Biol Sci* **285**:20182255. doi:10.1098/rspb.2018.2255

West RJH, Elliott CJH, Wade AR. 2015. Classification of Parkinson's disease genotypes in Drosophila using spatiotemporal profiling of vision. *Sci Rep* **5**:16933. doi:10.1038/srep16933