RUNNING HEAD: Comparing ERPs with successive univariate tests

Statistically comparing EEG/MEG waveforms through successive significant univariate tests: How bad can it be?

Vitória Piai[1,2], Kristoffer Dahlslätt[3] and Eric Maris[1]

In press, Psychophysiology

[1] Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands

[2] University of California Berkeley, Helen Wills Neuroscience Institute, Berkeley, CA, USA

[3] Independent Researcher

Correspondence concerning this article should be addressed to Vitória Piai, University of California, Helen Wills Neuroscience Institute, 132 Barker Hall, Berkeley, CA 94720-3190, USA, email: v.piai.research@gmail.com, phone : +1-510-6439744, and to Eric Maris, Radboud University Nijmegen, Montessorilaan 3, 6525HR, Nijmegen, the Netherlands, email: e.maris@donders.ru.nl, phone: +31-24-3612651.

Abstract

When making statistical comparisons, the temporal dimension of the EEG signal introduces problems. Guthrie and Buchwald (1991) proposed a formally correct statistical approach that deals with these problems: comparing waveforms by counting the number of successive significant univariate tests and then contrasting this number to a well-chosen critical value. However, in the literature, this method is often used inappropriately. Using real EEG data and Monte Carlo simulations, we examined the problems associated with the incorrect use of this approach under circumstances often encountered in the literature. Our results show inflated false-positive or false-negative rates depending on parameters of the data, including filtering. Our findings suggest that most applications of this method result in an inappropriate family-wise error rate control. Solutions and alternative methods are discussed.

Introduction

Event-related potentials/fields (ERP/Fs) are widely used to investigate brain processes. ERP/Fs are obtained from the electro/magnetoencephalogram (EEG/MEG) by averaging the signal over multiple occurrences of an event. Typically, researchers manipulate a variable to create experimental conditions to be compared.

ERP/Fs have a high temporal resolution and researchers like using this property for temporally specific statistical inferences (i.e., inferring an effect in some time interval but not in others). As a consequence, when researchers want to compare the time-resolved ERP/Fs of different conditions, they are faced with a multiple-comparisons problem (MCP). In this context, it is important that the statistical test employed controls the probability of a false positive (i.e., Type I) error in the set of multiple tests performed at some critical alpha-level: the family-wise error rate (FWER). In practice, researchers normally obtain ERP/Fs from multiple channels, and therefore, the MCP also emerges in the spatial dimension. However, in this paper, we only focus on the temporal dimension of ERP/Fs.

In a method proposed by Guthrie and Buchwald (1991), univariate *t*-tests between two conditions are performed for each time sample, the number of successive significant *t*-tests is counted (a so-called *run*), and the longest run is determined. Under the null hypothesis of no difference between the experimental conditions, the probability distribution of this longest run strongly depends on the autocorrelation of the noise signal, with noise denoting the part of the signal that is not evoked by the stimulus. Autocorrelation is the correlation of a signal with a time-shifted version of this same signal. With a time-shift of a single sample, this autocorrelation (i.e., *first-order* autocorrelation) measures how similar the signal is at two adjacent samples. The larger the *first-order* autocorrelation, the longer the intervals of successive between-condition differences that all deviate from zero in the same direction. As a consequence, if a sample shows a false-positive between-condition difference at a particular time sample then, with increasing first-order noise autocorrelation[1], the number of false-positive differences at adjacent samples also increases.

In line with these observations, Guthrie and Buchwald (G-B henceforth) proposed to reject or accept the null hypothesis on the basis of a critical-run value (based on the probability distribution of the longest run) that depends on the signal's noise autocorrelation. For a given univariate threshold alpha-level, this critical-run value depends on two parameters: the signal's length (in number of samples) and its noise autocorrelation[2], which was discussed above. With respect to the signal's length, note that it is measured in number of samples, and therefore depends both on the sampling rate and on the signal's length measured in seconds. It is important to stress that G-B's method is formally correct if these two parameters are known and subsequently used to determine the critical-run value.

An important problem with G-B's method is that it assumes the noise autocorrelation to be known, but it is unclear how it should be estimated from the data. G-B's proposal for estimating it consists of first removing systematic sources of variation from the signal. This is achieved by applying singular value decomposition to the signal and then removing the first *k* singular value component from the original signal. This residual signal is then used in the computation of the first-order autocorrelation. However, it is unclear how many components (parameter *k* in G-B's appendix) must be removed from the data.

G-B provided a table with critical-run values given the critical parameters only for a finite set of parameter combinations. In principle, for every parameter combination, the critical-run value can be calculated, but researchers may decide not to and simply choose a number from G-B's table.

We surveyed studies in the literature employing G-B's method to assess the rate of compliance to the method's requirements. For each year between 2000 and 2013, we collected information from four highly cited articles[3], resulting in 58 studies employing G-B's method in the past 14 years. We tried to select the articles with the highest number of citations for each year, but this turned out not to be ideal because in many cases, this strategy yielded many articles by the same research group. When that was the case, we selected other highly cited articles published by a different research group. With these criteria, we are confident that the articles surveyed provide a good representation of the (high-impact) literature. Notably, some studies did not report enough details to allow for a reconstruction of how the method was applied. For example, for one study, it could not be determined whether the noise autocorrelation had been calculated or just assumed. For two studies, the signal's length could not be derived from the information provided in the article, and for a large number of studies, the signal's length had to be inferred from the figures. For 14% of the studies, the critical-run value was not explicitly stated. Only 10% of the studies reported calculating the noise autocorrelation. The signal's length varied between 11 and 1024 samples (mean = 341, median = 250, *sd* = 253), with 73% of the studies using lengths larger than the maximum value of 150 samples provided in G-B's table. The critical-run value employed in these studies varied between 4 and 30 samples (mode = 11, mean = 12). Finally, low-pass filter cut-off values ranged from 20 to 300 Hz. In this context, the use of low-pass filtering is particularly interesting as this type of filtering temporally smooths the signal. Contrary to high-pass filtering, which does not temporally smooth the signal, low-pass filtering (and by extension band-stop filtering, commonly used to remove power line noise) increases the first-order autocorrelation. For our purposes, we focus on filtering issues only with respect to the autocorrelation. If the autocorrelation is calculated and taken into account in the statistical test, the impact of filtering on the false alarm rate is also accounted for. However, given our observation that only 10% of the surveyed studies calculated the autocorrelation when using G-B's method, the impact that filtering can have on the statistical test seems often not to be accounted for. In total, only 7% of the studies we surveyed reported the autocorrelation and subsequently employed a critical-run value that was appropriate given the signal length.

Below, we demonstrate the problems associated with the inappropriate use of G-B's method. We show how disregarding the noise autocorrelation and the signal's length, both very common practices in the literature, impacts the FWER control of G-B's method.

Simulation Protocol

We employed real EEG data[4] (Piai, Roelofs & Maris, 2014) of 14 participants reading sentences (sampled at 500 Hz with a 0.016–100 Hz online band-pass filter during recording). For our purposes, the artefact-free EEG segments comprising the first word of each sentence[5] were used, baseline-corrected using the average EEG between -150 ms and word onset (mean number of trials per participant = 98). Data from six channels (Cz, C3, C4, Pz, P3, P4) were averaged per participant, representing one channel. The signal was analysed between 150 ms pre- to 600 ms post-stimulus, yielding 375 sample points. For comparison with G-B's table, which extends only until 150 sample points, we also downsampled the signal to 250 Hz, yielding 187 sample points. To compare the results of our simulations across different pre-processing pipelines, these data were low-pass filtered at cut-offs of 30 and 15 Hz (-6 dB attenuation) using a zero-phase Hamming windowed finite impulse response filter of order 100 (Matlab 2010b, fir1 function, default parameters, no zero-padding, one-pass, 53 dB stopband attenuation).

We used the Monte Carlo method to evaluate the FWER control of G-B's method. The trials of each participant were randomly partitioned into two sets and then averaged to form two participant-average ERPs, representing two surrogate conditions. The random partitioning of the data from the same condition ensures that the null hypothesis is true. Then group-level (across the 14 participants) between-condition *t*-tests were performed and the longest run was compared to the critical-run values of 5 (a common low value in G-B's table for 150 sample points), 11 (middle-range, most common value used in the studies we surveyed), and 17 (highest value in G-B's table), without taking the autocorrelation into account. This procedure was meant to illustrate the behaviour of this test under conditions often used in the literature (i.e., disregarding the autocorrelation and the signal's length). The simulation procedure (random partitioning followed by group-level *t*-tests) was repeated 1,000 times. By calculating the proportion of simulations (i.e., random partitions) that yielded a significant difference between the two surrogate conditions (with a true null hypothesis), we can assess the test's FWER control. Under a critical alpha-level of .05, about 5% of the random partitions should yield a significant difference between the two conditions.

Results and Discussion

Figure 1 shows the results of the simulations. Horizontal solid lines indicate the 5% critical alpha-level. The results can be summarised as follows: the FWER increases with (1) shorter critical-run values, (2) smaller low-pass filter cut-offs, and (3) longer signals. These results emphasise the need to use the correct critical-run value when using this method, that is, the value that controls the FWER at 5%. We note that very similar results were obtained when the simulations were run on ERFs from MEG data with participants viewing pictures (from Piai, Roelofs, Jensen, Schoffelen & Bonnefond, 2014), and when we used an infinite impulse response Butterworth filter (default option in Fieldtrip, Oostenveld, Fries, Maris & Schoffelen, 2011, and in Brain Vision Analyzer).

The present results highlight the importance of adjusting the critical-run value as a function of the signal's length and the filter parameters. In this regard, our results reinstate the original conclusion by G-B regarding the effect of the signal's length and autocorrelation on the test's performance. It is clear that simply copying a critical number from G-B's table, ignoring the factors on which this choice should be based, can result in either an increased false alarm rate or decreased sensitivity. At present, a typical ERP/F study uses 30-Hz low-pass filter. In studies using G-B's method, a critical run of 11 samples is often used over segments comprising more than 250 sample points on average. Our results show that, for this typical ERP/F study one may find in the literature, current practices likely have an inappropriate FWER control.

To avoid having to remove an arbitrary number of principal components from the data or having to use a critical number from G-B's table while the autocorrelation is unknown, one can calculate the critical-run length based on the *observed* data using the Monte Carlo approach we adopted for our simulations. An improved alternative to this approach is implemented in the cluster-based permutation test (Maris & Oostenveld, 2007). Under this approach, adjacent time-points that exhibit a similar difference across conditions are clustered together, just as in G-B's method. Then, for each cluster, a cluster-level statistic is calculated (e.g., by taking the sum of the *t*-values within that cluster) and the largest cluster (either in size or in *t*-value) is selected. The crucial difference with G-B's method is that a permutation distribution is then created by calculating the largest cluster-level statistic for all random permutations *of the data being analysed*, with which the observed statistic is then compared. This obviates the need to know the autocorrelation and to consult a look-up table of critical values. Another important advantage of this approach is that can deal with the MCP not only in the temporal domain, but also in the spatial and spectral domains with the clusters being determined on the basis of spatial and spectral adjacency.

From the studies we surveyed, 82% used G-B's method for temporally specific inferences. That is, in these studies, inferences were made about specific time windows within the epoch analysed. Even more specifically, in 32% of the studies, inferences were made about the precise time-point when the signals from two conditions first diverge or converge, indicating exactly *when* the onset/offset of the effect is. From a statistical point of view, this type of temporally specific inference requires that a particular false alarm rate is controlled: the probability of identifying as significant an earlier time point than the first one for which the null hypothesis does not hold. This false alarm rate is not controlled by G-B's method, and neither is it by cluster-based permutation tests (Maris & Oostenveld, 2007). In fact, these approaches only control the false-alarm rate under the omnibus null hypothesis involving no effect for *none* of the time points. This point is discussed in more detail in Maris (2012), focusing on spatially specific inference. With respect to onset latency differences, methods should be preferred that have been constructed and validated for that specific purpose (Kiesel, Miller, Jolicoeur & Brisson, 2008; Miller, Patterson, & Ulrich, 1998).

In sum, G-B's method provides a formally correct way of statistically comparing waveforms. However, it assumes a known noise autocorrelation, which we do not know how to estimate from the data. For a given noise autocorrelation, the critical-run value can then be calculated (or looked-up in a table) as a function of the signal's length. However, in practice, a large number of studies uses G-B's method inappropriately, often resulting in increased false-positive rates. An alternative approach that circumvents the need to know the noise autocorrelation and the signal length is to calculate the *p*-value under a permutation distribution generated from the observed data, as with the Monte Carlo approach implemented in this article or in the cluster-based permutation test.

References

Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology*, *28,* 240-244.

Kiesel, A., Miller, J., Jolicoeur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology*, *45*, 250-274.

Maris E. (2012). Statistical testing in electrophysiological studies. *Psychophysiology*, *49,* 549–565.

Maris, E. & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164,* 177–190.

Miller, J., Patterson, T., & Ulrich, R. (1998). Jackknife-based method for measuring LRP onset latency differences. *Psychophysiology*, *35,* 99–115.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, *2011*, 1-9.

Piai, V., Roelofs, A., & Maris, E. (2014). Oscillatory brain responses in spoken word production reflect lexical frequency and sentential constraint. *Neuropsychologia*, *53,* 146-156.

Piai, V., Roelofs, A., Jensen, O., Schoffelen, J.M., & Bonnefond, M. (2014). Distinct patterns of brain activity characterise lexical activation and competition in spoken word production. *PLoS ONE* 9(2): e88674.

Footnotes

[1] We are not aware of any study that has investigated the sources of the noise autocorrelation. However, when group-average evoked responses are compared between two conditions (typically manipulated within participants), an important source of the noise autocorrelation will be the individual differences in the timing and amplitude of the evoked responses. These individual differences are typically visible by eye when the group-average difference waveform is subtracted from the participant-specific difference waveforms. Instead, when participant-specific evoked responses are compared between two conditions (manipulated between trials), an important source of the noise autocorrelation will be the between-trial differences in the timing and amplitude of the signal-trial evoked responses. Due to the lower signal-to-noise ratio at the single-trial level, these between-trial differences are typically *not* visible by eye when the evoked response is subtracted from the single trials. However, these between-trial differences will contribute to the noise autocorrelation.

[2] With the simulations, G-B found that the number of participants (i.e., degrees of freedom) was not a parameter influencing the critical-run value.

[3] The journals in which these articles were published are the following: Acta Psychiatrica Scandinavica, Behavioral Brain Research, Biological Psychiatry, Cerebral Cortex (4 articles), Clinical Neurophysiology (5 articles), Cognitive Brain Research, Developmental Neuropsychology, Frontiers in Human Neuroscience, Frontiers in Psychology, Hippocampus, Human Brain Mapping, Integrative Physiological & Behavioral Science, International Journal of Psychophysiology, Journal of the Acoustical Society of America, JAMA Psychiatry, Journal of Cognitive Neuroscience (4 articles), Journal of Neuroscience (10 articles), Neurobiology of Aging, NeuroImage (11 articles), Neuropsychologia (2 articles), Neuroscience, PLoS ONE, Proceedings of the National Academy of Sciences of the United States of America (2 articles), Psychiatry Research, Psychophysiology.

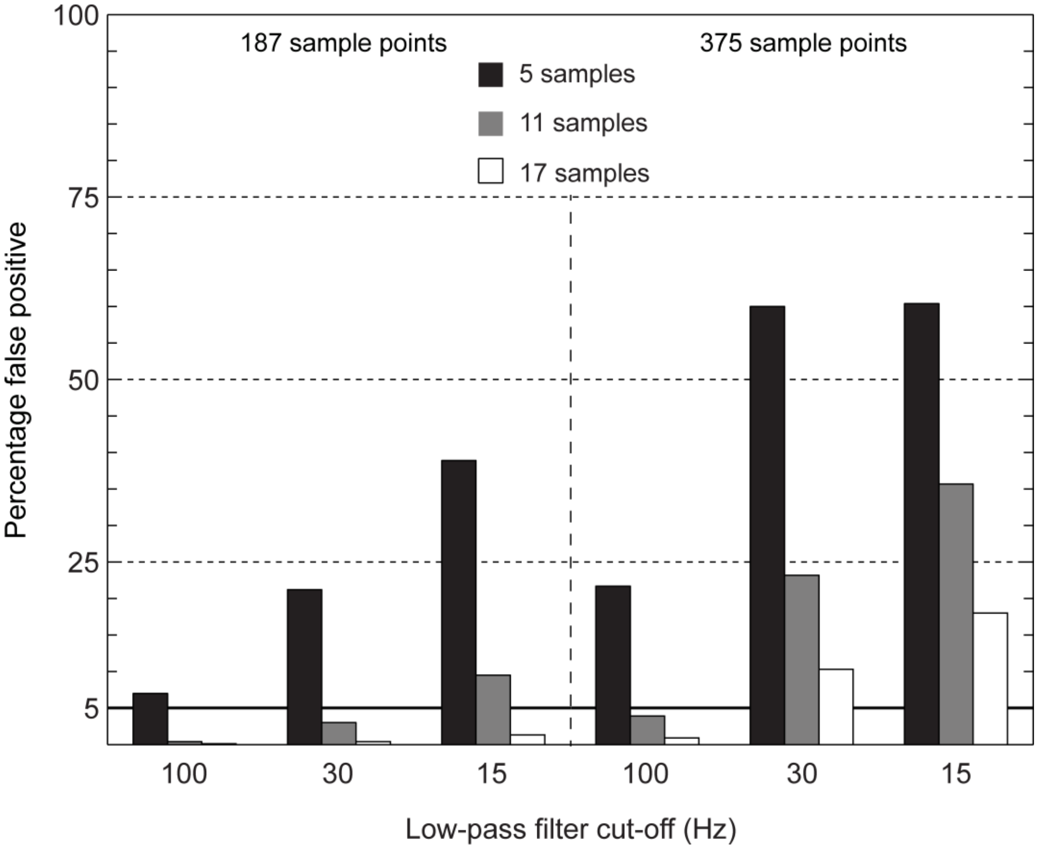
[4] We opted for avoiding assumptions with respect to the properties of the electrophysiological signal, which could erroneously influence the outcome of the simulations, and therefore employ real

EEG data in our simulations.

[5] The segments of Piai et al.'s nonconstraining condition were used.

Author Notes

We are very thankful to Andreas Widmann for helpful discussion on data filtering and to Ivar Clemens for technical support. This research was supported by a grant from the Netherlands Organization for Scientific Research (#446-13-009) to the first author.



Results of the simulations for signal's length of 187 (left) and 375 (right) sample points with critical-run values of 5, 11, and 17 samples. Horizontal solid lines indicate the 5% critical alpha-level. 100 Hz = data filtered only during recording (0.016–100 Hz band-pass) with no additional low-pass filtering.