

# A Method to Determine the Presence of Averaged Event-Related Fields Using Randomization Tests

Thomas Koenig · Lester Melie-García

Received: 2 February 2010 / Accepted: 29 March 2010 / Published online: 8 April 2010  
© Springer Science+Business Media, LLC 2010

**Abstract** We present a simple and effective method to test whether an event consistently activates a set of brain electric sources across repeated measurements of event-related scalp field data. These repeated measurements can be single trials, single subject ERPs, or ERPs from different studies. The method considers all sensors simultaneously, but can be applied separately to each time frame or frequency band of the data. This allows limiting the analysis to time periods and frequency bands where there is positive evidence of a consistent relation between the event and some brain electric sources. The test may therefore avoid false conclusions about the data resulting from an inadequate selection of the analysis window and bandpass filter, and permit the exploration of alternate hypotheses when group/condition differences are observed in evoked field data. The test will be called topographic consistency test (TCT). The statistical inference is based on simple randomization techniques. Apart from the methodological introduction, the paper contains a series of simulations testing the statistical power of the method as function of number of sensors and observations, a sample analysis of EEG potentials related to self-initiated finger movements, and Matlab source code to facilitate the implementation.

Furthermore a series of measures to control for multiple testing are introduced and applied to the sample data.

**Keywords** Evoked potentials · Randomization · Global Field Power · Averaging · Signal-to-noise ratio

## Introduction

When investigating event-related EEG or MEG potentials, it is typically assumed that the signal emerges from noise sometime before or shortly after the event, lasts for some period of time and then disappears into the noise again. However, the question of when this actually takes place is almost never addressed.

Event related potentials are typically investigated by comparing two or more conditions, where one condition serves as baseline against which the conditions of interest are contrasted. This baseline can be a classical experimental control condition or a pre-stimulus period (e.g. Foxe et al. 2008; Lakatos et al. 2005; Murray et al. 2001, for stimulus related changes, or Sperli et al. 2006; Zumsteg et al. 2006, in the case of epileptic activity). Several authors have proposed custom-tailored methods for the statistical testing of differences between conditions in event related brain potentials (Blair and Karniski 1993; Galan et al. 1997; Greenblatt and Pflieger 2004; Guthrie and Buchwald 1991; Karniski et al. 1994; Koenig and Melie-Garcia 2009; Lobaugh et al. 2001). Thus, while this approach has produced an abundant body of knowledge of the electrophysiological changes induced by the event, and robust methods exist to test for the statistical significance of these changes, they obviously depend not only on the conditions of interest, but also on the baseline. Tests for the existence of an event related potential itself (in the absence

---

T. Koenig (✉)  
Department of Psychiatric Neurophysiology, University Hospital  
of Psychiatry, University of Bern, Bern, Switzerland  
e-mail: thomas.koenig@puk.unibe.ch

L. Melie-García  
Neuroinformatics Department, Cuban Neuroscience Center,  
Havana, Cuba

of a contrasting condition) are however almost inexistent in the present literature.

Nevertheless, having and applying an objective method to obtain evidence for the presence of an event-related EEG signal can have a significant impact on the further analysis and interpretation of the data.

First, by setting the analysis window arbitrarily, one may obtain falsely negative results. This can occur in the trivial case where a period with a significant signal has not been included in the analysis window. Another possibility for falsely negative results is that the analysis window that has been chosen is too large and corrections for multiple testing have been applied such that the number of time-frames, and thus the number of statistical tests is inflated above the necessary amount and over-correction occurs. Both of these cases can be avoided when the analysis window is set to the periods where there is positive evidence for a signal that's emerging from the noise.

Second, signal and noise often have different spectral distributions, and temporal filters are usually applied to improve the signal-to-noise ratio. Although it is obvious that applying a band-pass filter matching the spectrum of the signal reduces the noise much more than the signal, this task is a fly-by-night operation if the spectrum of the signal is unknown. As above, this may result in falsely negative findings, either by choosing a band-pass filter that is too narrow, significantly truncating the signal, or by applying a band-pass filter that is too broad, such that remaining noise obscures relevant aspects of the signal. By analogy to the above point about the temporal analysis window, having a method to tailor the spectral window of the analysis to the event-related signal could avoid such problems.

Third, if two cases (either conditions or groups) are compared, and one finds a significant difference between the two cases, this may either indicate that the two cases had different signals, or that one case had a significant signal and the other did not. Because in both of these situations, there is a significant difference, these possibilities cannot be distinguished by the comparison of the two cases alone, although this distinction might be important for the correct interpretation of the results. Applying statistics for the presence of a signal to the two cases can resolve this ambiguity and thus provide a means for a more precise interpretation of the encountered differences.

Finally, although this is more of a help than a necessity, noise may also be introduced by errors in the data analysis, and testing for the presence of the signal before continuing the analysis may signal such errors at a relatively early stage of the analysis.

The aim of the current paper is thus to introduce a simple methodology to test for the presence of a signal in multi-channel ERP data in the time, frequency, or time-

frequency domain. The method should be based on arguments that allow interpreting positive evidence of a signal as evidence for the activity of some stable set of active neurons. Furthermore, it should be global, i.e. across all channels, first, because a significant activation of some set of neurons will manifest itself as an electric field that is typically widely distributed across the scalp, and second, because temporal windows and band-pass filtered are always applied commonly to all channels of the data. And finally, the method should require only a minimum of statistical assumptions, because any assumption that has not been made is one assumption less that may be violated.

Once the method is developed, a further and obvious aim of the paper is to apply some benchmarking in order to know whether and how well the method works under different conditions.

## Materials and Methods

### General Principle

Let us assume we have a set of sources in the brain that activate at some time-period and in some frequency range in a constant relation to some repeated event. This set of sources will produce a scalp electric and magnetic field that is determined by the lead field of those active sources, and that is typically widely distributed across the scalp (Mosher et al. 1999). Assume furthermore that there is random, zero mean noise in the data.

For one given point in time and/or frequency, the null hypothesis we want to test is whether repeatedly measured scalp fields can be explained entirely by noise, which would in turn suggest that there is no evidence that at least partly overlapping sources were active across the repeated measurements. The evidence for at least partially overlapping sources is defined as evidence for a contribution of a scalp field with a constant spatial distribution across the repeated measurements. If we can reject this null hypothesis, we can accept the alternative hypothesis that at the investigated point in time/frequency, there is a scalp field that is constantly being observed in relation to the repeated event and therefore assumingly corresponds to a set of active neurons in the brain that are functionally associated with this type of event. In the current context, constancy is therefore investigated across repeated observations, and not across time or frequency ranges (Koenig et al. 2005).

So under the null hypothesis, averaging the data across observations should lead to increasing cancellation in the data and the mean should converge to zero. If parts of the data have been caused by a constant set of active sources, the mean will not converge to zero, because the signals

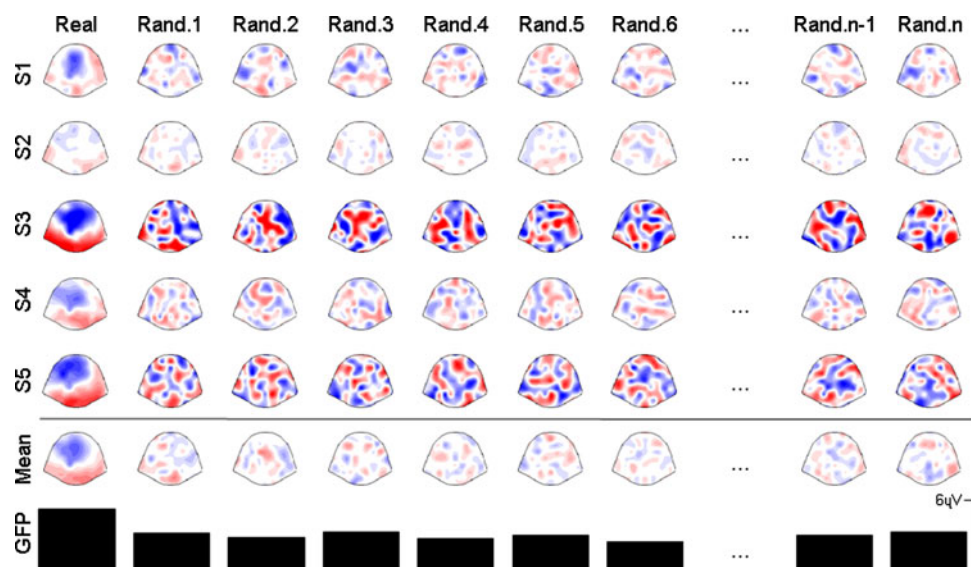
from these sources do not cancel out across observations. Therefore, the mean value of the averaged potential at some sensor is an index for the presence of a constant signal in the data at that sensor. Extending this argument to multi-channel data, we can conclude that having mean values different from zero across widely distributed scalp areas is an index for the presence of a constant scalp field, and therefore suggests the presence of a constant set of active sources.

In order to test the probability of the null hypothesis, we thus need a global (across all channels) index of the presence of a scalp field in the average across observations. Such an index is given by the Global Field Power (GFP, Lehmann and Skrandies 1980) that is computationally equivalent to the standard deviation across all channels and can be formulated in a reference independent way. The GFP of a mean ERP has previously been used as index of topographic consistency (Brandeis et al. 1992), and it could also be shown that when the topographies of all single observations are scaled to unity GFP before averaging, the resulting GFP of the average ERP is identical to the mean correlation coefficient of the mean ERP map with the maps of the single observations.

However, since the GFP of an average scalp field depends both on the amount of signal and on the overall variance of the data across channels and observations, the GFP value alone does not yield information about significance, i.e. the probability of the null-hypothesis. In order to

obtain the significance of a given GFP value obtained from some averaged scalp-field data, this GFP value has to be compared to a distribution of GFP values that is compatible with the null hypothesis.

The distribution of the GFP of some averaged scalp-field data under the null hypothesis can be obtained by randomization techniques (Manly 2007). The purpose of the randomization procedure is to modify the given data such that the overall variance across all channels and observations remains unchanged, but that a potential constancy of a signal across observations is eliminated. This can be achieved by randomly shuffling, in each observation, the data among channels. When this randomized data is averaged across observations, the GFP of the average data is by definition an instance of the GFP under the null-hypothesis. It depends solely on the variance across channels and observations and not on some constancy across observations. By repeating the randomization procedure many times, one obtains thus an empirical distribution of the GFP of an average scalp field under the null hypothesis. The probability of the null hypothesis is given by the percentage of cases were a GFP value obtained after randomization is larger or equal to the GFP value obtained in the original data (Manly 2007). The entire procedure is illustrated by a Matlab code snippet in Appendix and in Fig. 1. We will from now on refer to this test procedure as topographic consistency test (TCT).



**Fig. 1** Illustration of the procedure based on first five trials of the sample data, at time = 120 ms. For further information on the sample data, see text below. The *first column* shows the real data, the *other columns* instances of the randomized data. The *upper five rows* show the potential maps of the first five trials (S1–S5), the *6th row* shows the mean map of those five trials, and the *7th row* the GFP of the

mean map. It becomes apparent that the randomized maps of the single trials have by definition equal amplitudes and GFP as the real data (although obviously a different field), whereas the amplitudes and thus the GFP of the mean maps of the randomized data are consistently lower computed to the real data. All maps are scaled to  $\pm 16 \mu\text{V}$

## Specific Implementations

In the above section, we have outlined a general principle that can be adapted to specific cases. The repeated observations can be repeated stimuli within one subject, or they can be averaged ERPs measured across a series of subjects, (this is probably the most common case), or they can be grand-means from a series of studies. The maps upon which the procedure is based can be in the time-domain, obtained at a specific latency or latency-range from the event of interest, which gives us information about the time-window where a signal is present or absent. Alternatively, it can also be a map consisting of complex numbers that were obtained using an FFT or a complex wavelet analysis of the single observations, which yields information on the frequency or time–frequency window in which the signal is present. This can then be used to choose an appropriate band-pass filter.

## Control for Multiple Testing

Typically, ERPs contain many time frames, and if one wants to know at which of these time frames there is a consistent topography, and one therefore applies the TCT to all time frames separately. This may obviously lead to problems of multiple testing. Since in addition, the data submitted to these repeated tests is correlated, it is difficult to estimate the overall degrees of freedom of the entire dataset, which makes corrections for multiple testing such as the Bonferroni correction difficult. There are several possible solutions to this problem:

One option is to first attempt to reject the null-hypothesis for the entity of the data in a single test. Given that this test is significant (which is evidence for consistent topographies at least somewhere in the data), one performs the consistency test for all data points, considering them as post-hoc tests that precisely define where this consistency occurs. In order to perform a single test for the entire dataset, one can vectorize the channel by data-point matrices of the single observations and apply the test to these vectors. For example, if one has evoked potentials with 64 channels, 100 time frames and 25 subjects, one would first compute the TCT using all 6,400 values of each subject, and, given this is significant, one then computes 100 post-hoc tests with the 64 channels of each time frame separately.

Another option is to perform additional testing based on the count of significant results observed in the data in comparison to the count expected under the null-hypothesis (Koenig and Melie-Garcia 2009). We proceed as follows: First, given a threshold  $\alpha$  for significance, we count the number of significant results in our data. As usual in statistics, we also have to obtain the distribution

of this count under the null-hypothesis. In the present case, we can extract this distribution from the randomization runs. To begin sampling this distribution, we assume that the data of interest has not been the data that we have actually measured, but the data that we have obtained in the first randomization run. The GFP values of this first randomization are thus compared against all other randomization runs for each data-point, the  $p$ -values are computed in the way described above, and the sum of  $p$ -values smaller than  $\alpha$  is computed. This count is by definition a count of “significant” results under the null-hypothesis, i.e. a count of false positives. This count can now be computed for each randomization run, yielding the needed estimate of the distribution of the count of false positives. Finally, the count obtained in the real data is compared against the distribution of false positive counts; the  $p$ -value is then the percentage of counts under the null-hypothesis that are greater or equal to the count obtained in the real data. A similar procedure can also be applied when, instead of the count of significant results, the duration of continuous periods of significance is taken as basis for the overall test. Having the distribution of the duration of false positive results allows thresholding the significant periods of the real data, excluding periods of significance that are not longer than expected by chance alone (Koenig and Melie-Garcia 2009; Nichols and Holmes 2001).

Finally, one can also resort to thresholding  $p$ -values based on the false discovery rate (FDR, Genovese et al. 2002), which limits the number of false positives to a defined number.

The above described count-test and the threshold estimation based on the FDR are implemented in the Matlab code in Appendix.

## Simulations and Examples

The simulations were constructed as a series of observations consisting of multi-channel data with a noise and a signal component. The noise was constructed separately for each observation, using normally distributed random numbers that were normalized to unity variance across channels. The signal was a map with a constant anterior posterior gradient with systematically varying variance across channels. Simulations were computed for series of 10, 50 and 100 observations, and for datasets with 21, 64 and 128 channels, yielding nine sets of simulations with all combinations of number of observations and number of channels. In each of these sets of simulation, the variance of the signal across channels was systematically increased from 0.01 to 1 in steps of 0.01. For each set, a graph was constructed showing the  $p$ -value obtained by the TCT as a function of the

variance of the signal, i.e. the signal-to-noise ratio. All simulations were computed with 5,000 randomization runs.

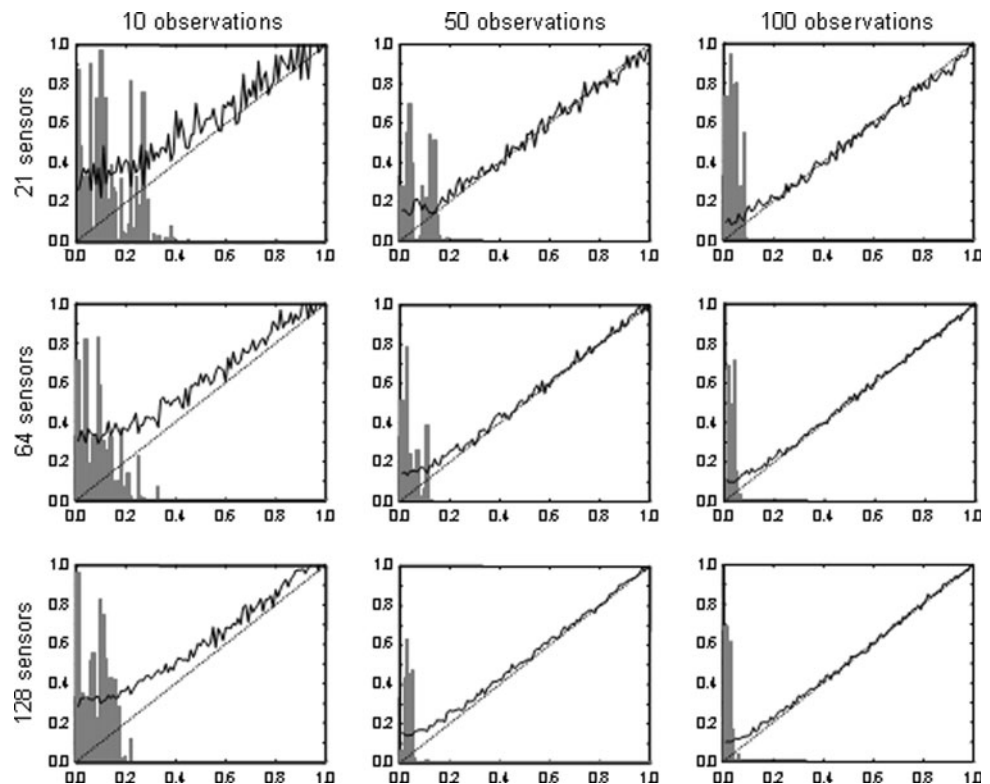
As an example for the application of the TCT in time-domain data, we analyzed EEG data related to self-initiated movements of the right index finger (Kornhuber and Deecke 1965), recorded from a healthy subject. The data were sampled from 62 channels, at 5,000 Hz digitization, with an analog bandpass filter from 0.1 to 2,500 Hz. The EEG was inspected for artifacts, recomputed to the common average reference, high-pass filtered at 70 Hz and (after EMG-onset detection) down-sampled to 250 Hz. 20 analysis windows were selected starting 2 s before and ending to 2 s after finger-related EMG onset. The TCT was applied time-instant by time-instant. The resulting  $p$ -value, the GFP of the average ERP, and  $t$ -maps against baseline were plotted.

To show the usefulness of the TCT for the investigation of the frequency extent of the signal, we employed a wavelet decomposition of the same data segments using complex Gabor functions (Durka and Blinowska 1995; Koenig et al. 2001), and applying the TCT to each time–frequency point. The wavelets covered a frequency range between 0.5 and 25 Hz. The resulting  $p$ -values as well as the GFP of the obtained averaged event-related signal were plotted as function of time and frequency.

## Results

The results of the simulations are shown in Fig. 2. As expected, the detection of a signal by our method increased both with the number of observations and the number of channels. Another observation is that the TCT becomes more robust against false positives when the number of sensors is increased. This is to be expected, because more channels allow for more possible permutations and the chances of obtaining permutations that resemble the measured data are smaller. Furthermore, when looking at sensitivity, and especially in the simulations with a high number of channels, the following natural relation between the number of observations and the significance of the test becomes apparent. For normally distributed data, the attenuation of noise by averaging is approximately proportional to the square-root of the number of observations. In the simulation with 10 observations, averaging across observations attenuates the noise approximately by a factor of  $\sqrt{10}$  or to about 30%. At the same time, the test started to give significant results when the variance of the signal was around 30% of the variance of the noise in the single observations. This holds also when larger numbers of observations were simulated. The TCT thus typically becomes significant when the signal-to-noise ratio is above one, and becomes non-significant when the signal-to-noise ratio is below one.

**Fig. 2** Significance of the TCT as function of signal to noise ratio, number of observations, and number of sensors. Each graph displays the result of a specific number of sensors and observations in the following way: the horizontal axis indicated the signal to noise ratio as defined in the text. The vertical axis indicates both the  $p$ -value and GFP. The lines represent the GFP of the signal alone (dotted line), and of the signal and noise mixture (continuous line). The gray bars indicate the  $p$ -value obtained by applying the TCT at the different signal to noise ratios. It becomes apparent that the test becomes more robust against false positives when more sensors were used, and it becomes more sensitive when the number of observations is increased



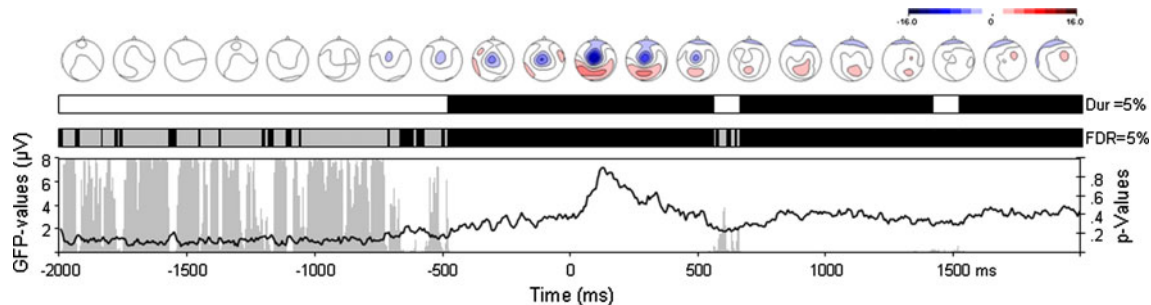
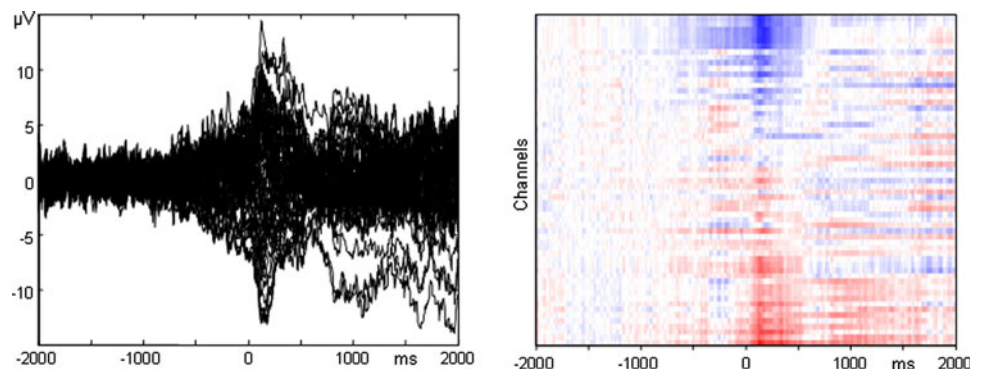
The averaged trials of the data measured for the application of the method are shown in Fig. 3. When the data was analyzed time domain, i.e. when the TCT was applied to each time-instant, we found evidence for the presence of a consistent topography across trials beginning around 500 ms before EMG onset (Fig. 4). When an FDR criterion of 5% was applied to control for multiple testing, a threshold of  $p \leq 0.032$  was found, indicating that  $p$ -values below that threshold had a below 5% likelihood to be false positive. Furthermore, the above described data-driven criterion for the count and duration of significance was computed using again a 5% threshold for significance of the individual tests. The test for overall significance using the count of locally significant results was significant ( $p \leq 0.001$ ), allowing us to reject the null-hypothesis on a global level. The duration threshold was found to be 328 ms, which indicates that there is a lower than 5% probability of finding periods of significance that last longer than 328 ms by chance alone. When this criterion was applied, the early and transient effects were eliminated.

The result was very consistent with the existing literature (Kristeva et al. 1979), especially when the duration

threshold was applied. The time-window where the test was significant coincided with the period where the  $t$ -maps showed areas covering several sensors that had  $t$ -values larger than around  $\pm 2$ , and covering several sensors. Furthermore, it seemed that the tests became significant when the GFP of the averaged ERP exceeded a certain threshold. Since in our test, GFP has been used to measure the presence of an ERP field, this suggests that the noise level in the data was approximately constant across time, and the significance of the test was mostly driven by the amplitude of the signal.

Figure 5 shows the significance of the TCT obtained when applied to the same data after having transformed the data to the time–frequency domain. There was a consistent signal across the entire time interval in the delta band that extended into the theta range around the time of the EMG onset. After the stimulus, there was also a consistent post-movement signal in the alpha range, which is a typical finding reported in the literature (Neuper and Pfurtscheller 2001). Interspersed across the analysis window, but somewhat concentrated around the EMG onset, were significant signals in the beta range.

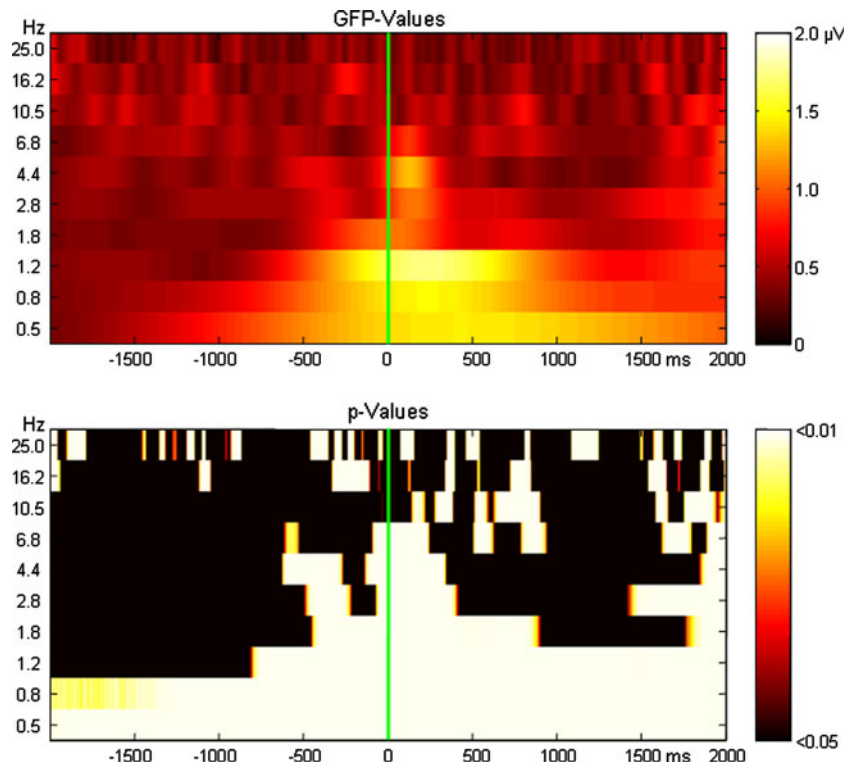
**Fig. 3** Butterfly-plot (left) and intensity plot of the unfiltered movement related potentials used for the sample analysis. In the intensity plot, the vertical axis indicates electrodes; the channels have been ordered according to their similarity. Colors have been scaled to cover a range from  $-15$  to  $15 \mu\text{V}$ . EMG onset was at time 0



**Fig. 4** Results of applying the TCT time-instant by time-instant to the recorded data. In the lower graph, the line shows the GFP of the averaged data (mapped to the left vertical scale), and the gray bars show the  $p$ -value of resulting from the tests (mapped to the right vertical axis). The black areas in the middle indicate periods where

the test met the threshold of a FDR of 5%, or of the duration criterion. The maps above the graph display the averaged scalp electric field for the different time periods. There is an almost continuous evidence of presence of a consistent topography beginning around 500 ms before EMG onset

**Fig. 5** Application of the TCT to the same data as in Figs. 3 and 4 in the time–frequency domain. The *horizontal axis* represents time, the *vertical axis* represents frequency, the *color codes* for the GFP (*upper graph*) and *p*-value obtained by the test (*lower graph*)



## Discussion

In the present paper, we have proposed a method to test for the presence of a consistent topography (TCT) in a series of repeated observations of multichannel EEG data related to some event. These observations can be set of un-averaged EEG epochs of a subject (testing for the existence of an ERP in this data, as in the given example), or a set of event-related potentials (usually recorded from different subjects, thus testing for the communality of activation across subjects) or even grand-averages of several studies (testing for the communality of activation across studies). The method is based on the well established relationship between intracerebral electric activity and the measurable scalp electric that such activity generates. If at a given point in time/frequency, the null-hypothesis of the test can be confidently rejected, this implies that at the investigated point in time/frequency, there was a set of active neurons in the brain that was functionally associated to the event the data was related to. For the statistical hypothesis testing, the TCT is based on randomization techniques that require very little assumptions.

Because of the ill-posed inverse problem of the EEG, one could argue that having evidence for a scalp field that remains constant across a series of observations does not necessarily impose that the distribution of intra-cerebral electric activity was also constant; different sources may

have produced identical scalp fields. However, assuming that a constant field corresponds to a constant set of neuronal sources is the basis of the computation of every averaged evoked potential, and averaged evoked potentials have produced an enormous body of convincing results, such that we think that this assumption is by far the most plausible one. Furthermore, the alternative hypothesis, namely that across trials, different sets of neuronal generators always produce the same scalp field is very unlikely.

Apart from having a well justified rationale, another advantage of the TCT is that by being based on GFP, it uses a reference independent index of presence of a scalp electric field. Therefore, the entire subsequent statistical inference and the resulting probability of the null-hypothesis is also reference independent, which is a very important point in view of all the confusion that different choices of references have brought to the ERP literature.

In the form that the TCT has been presented, it is global in space, because it considers all sensors simultaneously, but it is local in time and/or frequency, because it considers each time and/or frequency point separately. This is probably the most useful form if one wants to identify the appropriate analysis window in time and/or frequency, but obviously results in multiple tests of intrinsically correlated data. To assess the overall significance of a signal across multiple time/frequency intervals, one can apply the above described procedures to correct for the effects of multiple

testing (Koenig and Melie-Garcia 2009). In the sample data, this has proven to be efficient in removing short and potentially spurious findings at the beginning of the data. Please also note that applying the proposed consistency test to identify the analysis window for further comparisons among conditions does not lead to circular statistics, because the differences among conditions are not subject to the tests that define the analysis window.

The TCT complements other global procedures for statistical testing of ERPs that are also based on randomization and bootstrapping (Galan et al. 1997; Greenblatt and Pflieger 2004; Karniski et al. 1994; Koenig et al. 2008; Lobaugh et al. 2001), but that are used to compare different conditions. Both approaches (testing for presence of condition event-related activity in a single condition and testing for differences among conditions) can be mutually informative: In periods where there is no evidence for a consistent topography in either condition, further testing of differences among those conditions can be omitted, which limits the result space of the comparisons. If, for a given time period, there is evidence of a consistent topography in a subset of the conditions, further testing of differences among conditions may be dubious, because some of those conditions are not well defined, and one should rather discuss which conditions produce or don't produce a consistent response. Only if there is evidence for a consistent topography in all conditions entering a comparative analysis, these comparisons yield the typical interpretations of differently active neural sources. We therefore suggest as extended procedure in the analysis of average event-related potentials that the different conditions and groups are all first tested for consistent topography across observations (usually subjects) as function of time. The comparisons between conditions and groups are then limited to those periods where there is evidence for consistent topography in all conditions and groups included in the comparisons. Where this is not the case, the discussion of the data should focus on the absence of consistently activated sources in defined conditions/groups and not on differences among those.

There are some essential differences of the methods to compare conditions to the proposed test for topographic consistency also from their underlying rationale. While contrast-based methods refer to an experimentally defined

baseline that must not necessarily be zero, the method presented here explicitly relies on the assumption that under the null-hypothesis, the mean of the signal is distributed around zero at all electrodes. This assumption holds for evoked potentials and complex frequency or time–frequency data, but not for unsigned data, such as spectral power, or current densities. For such purposes, one may rely on methods as proposed by De Lucia et al. (2007), where a set of mutually exclusive features (spatial topographies obtained by cluster analysis) are extracted from single trial EEG data. Evidence for a stimulus-associated EEG event is then assessed by testing whether the probability of observing one specific feature at some latency is associated to the stimulus.

Interestingly, since the proposed method can also be applied to complex time–frequency data, evidence for a consistent (complex) topography at some time–frequency point is simultaneously evidence for consistent phase across trials, which may be useful to further investigate the issue of evoked versus induced evoked responses (e.g. Makeig et al. 2002; Shah et al. 2004). Another domain where the method may offer new and valid analysis options is in spike-aligned EEG data of patients with epilepsy, where the method offers a simple and powerful way of determining spike onset without having to apply inverse models (Lantz et al. 2003; Sperli et al. 2006; Zumsteg et al. 2006).

Finally, because the rejection of the null-hypothesis is based on the observation that the scalp field is consistently different from zero at some scalp location, it is obvious that the electrode montage must cover these locations to avoid a false negative finding. Thus, as repeatedly pointed out, it is essential that the electrode montage covers as much of the scalp as possible, and namely covers the essential peaks and troughs of the scalp field (Michel and Brandeis 2009).

**Acknowledgement** The authors wish to thank the two reviewers for their constructive and fast responses to the initial version of the manuscript.

## Appendix

Matlab code snippet illustrating the application of the method.



```

nObservations = size(data,1);
nElectrodes   = size(data,2);
nDataPoints   = size(data,3);

nIter         = 100;           % Precision of the p-values on the 1% level

gfp(1,:) = std(mean(data,1),1,2); % GFP of the mean across observations
rdata = zeros(size(data));      % Allocate space for the randomized data

for i = 2:nIter % This is the randomization loop
    for j = 1:nObservations
        rdata(j, :, :) = data(j, randperm(nElectrodes), :); % permute observations
    end
    GFP(i, :) = std(mean(rdata,1),1,2); % We re-compute the GFP
end

for t = 1:nDataPoints
    p(t) = (sum(GFP(:,t) >= GFP(1,t)))/nIter; % The p-values we want
end

% False discovery rate

FDR = 0.05; % This is the FDR that we choose to be acceptable

p_exp = (1:nDataPoints) / nIter; % This is the distribution of p-values under
% the null-hypothesis
p_corr = p_exp * FDR; % This is the distribution of the accepted
% false positives

p_sorted = sort(p);
FDR_threshold = p_sorted(max(find(p_sorted <= p_corr)));

% Count test
threshold = 0.05;

for i = 1:nIter
    for t = 1:nDataPoints
        p_fake(t,i) = (sum(GFP(:,t) >= GFP(i,t)))/nIter; % p-values under the
        % null-hypothesis
    end

    Hits(i) = sum(p_fake(:,i) < threshold); % Count of false positives
end

p_overall = sum(Hits(1) <= Hits) / nIter;

```

## References

- Blair RC, Karniski W (1993) An alternative method for significance testing of waveform difference potentials. *Psychophysiology* 5:518–524
- Brandeis D, Naylor H, Halliday R, Callaway E, Yano L (1992) Scopolamine effects on visual information processing, attention, and event-related potential map latencies. *Psychophysiology* 3:315–336
- De Lucia M, Michel CM, Murray MM (2007) Single subject EEG analysis based on topographic information. *Int J Bioelectromagn* 9:168–171
- Durka PJ, Blinowska KJ (1995) Analysis of EEG transients by means of matching pursuit. *Ann Biomed Eng* 5:608–611
- Foxe JJ, Strugstad EC, Sehatpour P, Molholm S, Pasiaka W, Schroeder CE, McCourt ME (2008) Parvocellular and magnocellular contributions to the initial generators of the visual

- evoked potential: high-density electrical mapping of the “C1” component. *Brain Topogr* 1:11–21
- Galan L, Biscay R, Rodriguez JL, Perez-Abalo MC, Rodriguez R (1997) Testing topographic differences between event related brain potentials by using non-parametric combinations of permutation tests. *Electroencephalogr Clin Neurophysiol* 3:240–247
- Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 4:870–878
- Greenblatt RE, Pflieger ME (2004) Randomization-based hypothesis testing from event-related data. *Brain Topogr* 4:225–232
- Guthrie D, Buchwald JS (1991) Significance testing of difference potentials. *Psychophysiology* 2:240–244
- Karniski W, Blair RC, Snider AD (1994) An exact statistical method for comparing topographic maps, with any number of subjects and electrodes. *Brain Topogr* 3:203–210
- Koenig T, Melie-Garcia L (2009) Statistical analysis of multichannel scalp field data. In: Michel CM, Koenig T, Brandeis D, Gianotti LRR, Wackermann J (eds) *Electrical neuroimaging*. MIT Press, Cambridge, pp 169–189
- Koenig T, Marti-Lopez F, Valdes-Sosa P (2001) Topographic time-frequency decomposition of the EEG. *Neuroimage* 2:383–390
- Koenig T, Studer D, Hubl D, Melie L, Strik WK (2005) Brain connectivity at different time-scales measured with EEG. *Philos Trans R Soc Lond B Biol Sci* 1457:1015–1023
- Koenig T, Melie-Garcia L, Stein M, Strik W, Lehmann C (2008) Establishing correlations of scalp field maps with other experimental variables using covariance analysis and resampling methods. *Clin Neurophysiol* 6:1262–1270
- Kornhuber H, Deecke L (1965) Hirnpotentialänderungen bei Willkurbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. *Pflügers Arch* 284:1–17
- Kristeva R, Keller E, Deecke L, Kornhuber HH (1979) Cerebral potentials preceding unilateral and simultaneous bilateral finger movements. *Electroencephalogr Clin Neurophysiol* 2:229–238
- Lakatos P, Pincze Z, Fu KM, Javitt DC, Karmos G, Schroeder CE (2005) Timing of pure tone and noise-evoked responses in macaque auditory cortex. *Neuroreport* 9:933–937
- Lantz G, Grave de Peralta R, Spinelli L, Seeck M, Michel CM (2003) Epileptic source localization with high density EEG: how many electrodes are needed? *Clin Neurophysiol* 1:63–69
- Lehmann D, Skrandies W (1980) Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalogr Clin Neurophysiol* 6:609–621
- Lobaugh NJ, West R, McIntosh AR (2001) Spatiotemporal analysis of experimental differences in event-related potential data with partial least squares. *Psychophysiology* 3:517–530
- Makeig S, Westerfield M, Jung TP, Enghoff S, Townsend J, Courchesne E, Sejnowski TJ (2002) Dynamic brain sources of visual evoked responses. *Science* 555:690–694
- Manly BFJ (2007) *Randomization, bootstrap and Monte Carlo methods in biology*. Chapman & Hall, Boca Raton
- Michel CM, Brandeis D (2009) Data acquisition and pre-processing standards for electrical neuroimaging. In: Michel CM, Koenig T, Brandeis D, Gianotti LRR, Wackermann J (eds) *Electrical neuroimaging*. MIT Press, Cambridge, pp 79–92
- Mosher JC, Leahy RM, Lewis PS (1999) EEG and MEG: forward solutions for inverse methods. *IEEE Trans Biomed Eng* 3: 245–259
- Murray MM, Foxe JJ, Higgins BA, Javitt DC, Schroeder CE (2001) Visuo-spatial neural response interactions in early cortical processing during a simple reaction time task: a high-density electrical mapping study. *Neuropsychologia* 8:828–844
- Neuper C, Pfurtscheller G (2001) Event-related dynamics of cortical rhythms: frequency-specific features and functional correlates. *Int J Psychophysiol* 1:41–58
- Nichols TE, Holmes AP (2001) Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp* 15:1–25
- Shah AS, Bressler SL, Knuth KH, Ding M, Mehta AD, Ulbert I, Schroeder CE (2004) Neural dynamics and the fundamental mechanisms of event-related brain potentials. *Cereb Cortex* 5:476–483
- Sperli F, Spinelli L, Seeck M, Kurian M, Michel CM, Lantz G (2006) EEG source imaging in pediatric epilepsy surgery: a new perspective in presurgical workup. *Epilepsia* 6:981–990
- Zumsteg D, Friedman A, Wieser HG, Wennberg RA (2006) Propagation of interictal discharges in temporal lobe epilepsy: correlation of spatiotemporal mapping with intracranial foramen ovale electrode recordings. *Clin Neurophysiol* 12:2615–2626