ELSEVIER

Journal of Neuroscience Methods 164 (2007) 177–190

JOURNAL OF NEUROSCIENCE METHODS

www.elsevier.com/locate/jneumeth

# Nonparametric statistical testing of EEG- and MEG-data☆,☆☆

Eric Maris[a,b,*], Robert Oostenveld[b]

[a] *NICI, Biological Psychology, Radboud University Nijmegen, Nijmegen, The Netherlands*
[b] *F.C. Donders Center for Cognitive Neuroimaging, Radboud University Nijmegen, Nijmegen, The Netherlands*

Received 7 January 2007; received in revised form 19 March 2007; accepted 29 March 2007

**Abstract**

In this paper, we show how ElectroEncephaloGraphic (EEG) and MagnetoEncephaloGraphic (MEG) data can be analyzed statistically using nonparametric techniques. Nonparametric statistical tests offer complete freedom to the user with respect to the test statistic by means of which the experimental conditions are compared. This freedom provides a straightforward way to solve the multiple comparisons problem (MCP) and it allows to incorporate biophysically motivated constraints in the test statistic, which may drastically increase the sensitivity of the statistical test. The paper is written for two audiences: (1) empirical neuroscientists looking for the most appropriate data analysis method, and (2) methodologists interested in the theoretical concepts behind nonparametric statistical tests. For the empirical neuroscientist, a large part of the paper is written in a tutorial-like fashion, enabling neuroscientists to construct their own statistical test, maximizing the sensitivity to the expected effect. And for the methodologist, it is explained why the nonparametric test is formally correct. This means that we formulate a null hypothesis (identical probability distribution in the different experimental conditions) and show that the nonparametric test controls the false alarm rate under this null hypothesis.


*Keywords:* Nonparametric statistical testing; Hypothesis testing; EEG; MEG; Multiple comparisons problem

## 1. Introduction

The topic of this paper is the statistical analysis of ElectroEncephaloGraphic (EEG) and MagnetoEncephaloGraphic (MEG) data. These data will subsequently be denoted together as MEEG-data. MEEG-data have a spatiotemporal structure: the signal is sampled at multiple sensors and multiple time points (as determined by the sampling frequency). The data are typically collected in different experimental conditions and the experimenter wants to know if there is a difference between the data observed in these conditions. In most studies, the conditions differ with respect to the type of stimulus being presented immediately before or during the registration of the signal. In other studies, the conditions differ with respect to the type of response (e.g., correct or incorrect) that was given.

In the statistical analysis of MEEG-data we have to deal with the multiple comparisons problem (MCP). This problem originates from the fact that the effect of interest is evaluated at an extremely large number of (sensor, time)-pairs. This number is usually in the order of several thousands. The MCP involves that, due to the large number of statistical comparisons (i.e., one per (sensor, time)-pair), it is not possible to control the so-called family-wise error rate (FWER) by means of the standard statistical procedures that operate at the level of single (sensor, time)-pairs. The FWER is the probability under the hypothesis of no effect of falsely concluding that there is a difference between the experimental conditions at one or more (sensor, time)-pairs. A solution of the MCP requires a procedure that controls the FWER at some critical alpha-level (typically, 0.05 or 0.01).

In this paper, we discuss nonparametric statistical testing of MEEG-data. Contrary to the familiar parametric statistical framework, it is straightforward to solve the MCP in the nonparametric framework. Nonparametric tests were first proposed for testing the difference between MEEG-waveforms at a particular sensor (Blair and Karniski, 1993, elaborating on a parametric procedure proposed by Guthrie and Buchwald, 1991), then for MEEG-topographies at a particular time point (Achim, 2001; Galán et al., 1997; Karnisky et al., 1994), and finally also for whole spatiotemporal matrices (Maris, 2004).

☆ The methods described in this paper have been implemented in the Matlab toolbox *Fieldtrip*, which is available from http://www.ru.nl/fcdonders/fieldtrip.

☆☆ The authors would like to thank Ole Jensen for generously sharing his data.

* Corresponding author at: Nijmegen Institute of Cognition and Information (NICI), Radboud University Nijmegen, PO Box 9104, 6500 HE Nijmegen, The Netherlands. Tel.: +31 243612651.

*E-mail address:* maris@nici.ru.nl (E. Maris).

0165-0270/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jneumeth.2007.03.024

Nonparametric tests have also been used very successfully for frequency domain representations of EEG- and MEG-data (Kaiser and Lutzenberger, 2005; Kaiser et al., 2000, 2003, 2006; Lutzenberger et al., 2002). Recently, nonparametric tests were proposed for distributed inverse solutions obtained by a minimum variance beamformer (Chau et al., 2004; Singh et al., 2003) or a minimum norm linear inverse (Pantazis et al., 2005). Finally, nonparametric tests for fMRI-data were proposed by Holmes et al. (1996), Bullmore et al. (1996, 1999), Nichols and Holmes (2002), Raz et al. (2003), and Hayasaka and Nichols (2003, 2004).

The present paper contributes to the literature in several respects: (1) it explains how the sensitivity of the statistical test can be drastically improved by incorporating biophysically motivated constraints in the test statistic, (2) it is written in a tutorial-like fashion, enabling neuroscientists to construct their own statistical test, maximizing the sensitivity to the expected effect, and (3) it explains why the nonparametric test is formally correct, making use of the so-called conditioning rationale, a concept that is both rigorous and intuitive. The paper is written for two audiences: (1) empirical neuroscientists looking for the most appropriate data analysis method, and (2) methodologists interested in the theoretical concepts behind nonparametric statistical tests. With the empirical neuroscientist in mind, we have written Sections 2 and 3 in a tutorial-like fashion, and with the methodologist in mind, we have written Section 4 that is sufficiently rigorous.

## 2. Methods

We make use of an example data set that was obtained in a study on the semantic processing of sentences (Jensen et al., submitted). This study involved a comparison of two experimental conditions that differed with respect to the semantic congruity of the final word in a sentence with the first part of the sentence. As is the case for most neuroscience studies, the authors are interested in the difference between experimental conditions with respect to the biological data. A central point of the present paper is that there are many aspects of the biological data that may differ between the experimental conditions, and in the following we will give some examples. This puts serious demands on the statistical framework that will be used for the analysis of these data, and the nonparametric statistical framework meets these demands to a large extent.

For the sake of clarity and simplicity, we will in this section deliberately ignore three important issues: (1) the exact specification of the null hypothesis that is tested by the nonparametric statistical test, (2) the proof that this test controls the false alarm rate, and (3) the issue of how to choose a test statistic. We deal with these issues in Section 4.

### 2.1. Example: the magnetic N400

Semantic processing of sentences is often studied by manipulating semantic congruity. Typically, one compares sentences in which the last word is semantically congruent with the preceding sentence (e.g., "The climbers finally reached the top of the mountain."), with sentences in which the last word is semantically incongruent (e.g., "The climbers finally reached the top of the tulip."). The interest is in the electrophysiological activity that is observed while the subject processes the last word. EEG studies have convincingly demonstrated that semantic incongruity produces a negative going potential deflection with a maximal amplitude at 400 ms after the onset of the semantically incongruent word. Kutas and Hillyard (1980) termed this the N400 effect. The N400 effect has been replicated in MEG studies and its primary sources have been localized in the left superior temporal sulcus (Simos et al., 1997; Helenius et al., 1998, 2002; Halgren et al., 2002) and the left prefrontal cortex (Halgren et al., 2002).

Jensen et al. (submitted) conducted an MEG study in which subjects listened to sentences that had either semantically congruent or semantically incongruent sentence endings. We obtained the data of a single subject that participated in this study. We want to identify the signature of the differences between these two experimental conditions. As will be shown in the following section, these conditions can differ on several aspects.

We restrict our attention to the data of a single subject. This does not reflect current practice in neuroscience, in which typically multiple subjects are observed in the same experimental paradigm. However, it serves our purpose of illustrating the main statistical concepts by means of a simple example data set. In Section 5, we show that these statistical concepts also apply to multi-subject studies. In that section we will also deal with the issue of generalization to a population.

### 2.2. Aspects on which the conditions may differ

The N400 effect refers to an effect at the level of evoked responses: the average voltage (for EEG) or magnetic field (for MEG) differs for the two experimental conditions. The evoked responses at the different (sensor, time)-pairs can be considered as different aspects of the biological data with respect to which the experimental conditions will be compared. In the following, a (sensor, time)-pair will be denoted as a sample.

The statistical analysis is very simple if we know in advance where (at which sensor) and when an effect may be observed. In this case, it is sufficient to calculate a single $t$-value and its corresponding $p$-value. The situation is more complicated if the spatiotemporal locus of a possible effect is not known in advance. In that case, it is not sufficient to calculate multiple $t$-values, one for every sample, and their corresponding $p$-values. In fact, due to the large number of statistical comparisons (one per sample), it is not possible to control the FWER by means of the standard statistical procedures that operate at the level of single samples (e.g., the $t$-test). This is the multiple comparisons problem (MCP). A solution of the MCP requires a procedure that controls the FWER at some critical alpha-level. In the following, whenever we use the term false alarm (FA) rate in the context of a statistical comparison at multiple samples, we mean the FWER. The point to remember is the following: if the spatiotemporal locus of the effect is not known in advance, we need a specialized statistical procedure that takes our prior ignorance into account.

Modulation of evoked responses is just one neural index that informs us about the brain mechanisms that underly lan-

guage processing. Another index is the modulation of oscillatory brain activity. Oscillatory brain activity is measured by power estimates obtained from Fourier or wavelet analyses of the single-trial spatiotemporal data. Very often, power estimates are calculated for multiple data segments obtained by sliding a short time window over the complete data segment. In this way, spatio-spectral–temporal data are obtained from the raw spatiotemporal data. The spectral dimension consists of the different frequency bins for which the power is calculated.

Just as the spatiotemporal evoked responses, the spatio-spectral–temporal oscillatory power estimates require a specialized statistical procedure that takes prior ignorance about the locus of the effect into account (ignorance with respect to the spatial, temporal, and spectral dimension). As will be shown in the following, this can be realized by means of the same type of nonparametric statistical test as for the spatiotemporal evoked responses. Besides the spatiotemporal evoked responses and the spatio-spectral–temporal oscillatory power estimates, many other types of data can be compared statistically using a nonparametric statistical test. For instance, it is straightforward to construct a nonparametric statistical test for the difference between conditions with respect to between-sensor coherence, which is a spatio-spatio-spectral data structure (between-sensor measures that are frequency-specific).

### 2.3. The nonparametric statistical test

The nonparametric statistical test is performed in the following way:

(1) Collect the trials of the two experimental conditions in a single set.
(2) Randomly draw as many trials from this combined data set as there were trials in condition 1 and place those trials into subset 1. Place the remaining trials in subset 2. The result of this procedure is called a random partition.
(3) Calculate the test statistic on this random partition.
(4) Repeat steps 2 and 3 a large number of times and construct a histogram of the test statistics.
(5) From the test statistic that was actually observed and the histogram in step 4, calculate the proportion of random partitions that resulted in a larger test statistic than the observed one. This proportion is called the *p*-value.
(6) If the *p*-value is smaller than the critical alpha-level (typically, 0.05), then conclude that the data in the two experimental conditions are significantly different.

This six-step procedure results in a valid statistical test: under some well-specified null hypothesis (see Section 4), the probability of falsely rejecting this null hypothesis is equal to the critical alpha-level.

When based on an infinite number of random partitions, the histogram constructed in step 4 is called the permutation distribution. The corresponding *p*-value is often called the permutation *p*-value and the associated statistical test is called a permutation test. Besides the permutation test, there is another nonparametric test, which is called the randomization test. The permutation and the randomization test have a different rationale (Ernst, 2004), but in practice they often involve the same calculations. For MEEG-data, it is not necessary to make a distinction between these two types of statistical tests. Therefore, we restrict ourselves to the permutation test.

In practice, it is not possible to calculate the permutation *p*-value by repeating steps 2 and 3 an infinite number of times. Instead, this *p*-value is approximated by a so-called Monte Carlo estimate. This Monte Carlo estimate is obtained by repeating steps 2 and 3 a large number of times and comparing these random test statistics (i.e., draws from the permutation distribution) with the observed test statistic. The Monte Carlo estimate of the permutation *p*-value is the proportion of random partitions in which the observed test statistic is larger than the value drawn from the permutation distribution. The accuracy of the Monte Carlo *p*-value increases with the number of draws from the permutation distribution. Because the Monte Carlo *p*-value has a binomial distribution, its accuracy can be quantified by means of the well-known confidence interval for a binomial proportion (Ernst, 2004). By increasing the number of draws from the permutation distribution, the width of this confidence interval can be made arbitrarily small. To simplify the presentation of the results, in all our example analyses, the Monte Carlo *p*-values were calculated on 1000 random partitions.

As compared to parametric statistical tests, the nonparametric statistical test is extremely general. This is because the validity of the nonparametric test does not depend on the probability distribution of the data (i.e., whether it has a normal or some other distribution), nor on the test statistic on which the statistical inference is based (i.e., whether it is a *t*-, an *F*-, or some other statistic). The freedom to choose any test statistic one considers appropriate has important advantages, and these will be illustrated and discussed in Sections 3 and 4 The two important advantages are the following: (1) it provides a simple way to solve the MCP, and (2) it allows us to incorporate prior knowledge about the type of effect that can be expected. This prior knowledge may drastically increase the sensitivity of the statistical test.

## 3. Results

### 3.1. Evoked responses

#### 3.1.1. Single-sensor analyses

For well-studied experimental paradigms, one often knows at which sensor the strongest effect can be observed. For instance, previous EEG-studies in which semantic congruity was manipulated (for a review, see Kutas and Federmeier, 2000) have shown the maximum effect over parietal cortex near the midline (Pz and the surrounding electrodes). And previous MEG-studies (Simos et al., 1997; Helenius et al., 1998, 2002; Halgren et al., 2002) have shown a dipolar pattern over left temporal and left frontal cortex. In panel a of Fig. 1, we show the evoked responses at a sensor over left temporal cortex, separately for congruent and incongruent sentence endings.

Inspection of Fig. 1 reveals a raw effect of semantic congruity that is most prominent in the time interval between 400 and 800 ms (see Fig. 1, panel a). We need a statistical test to decide

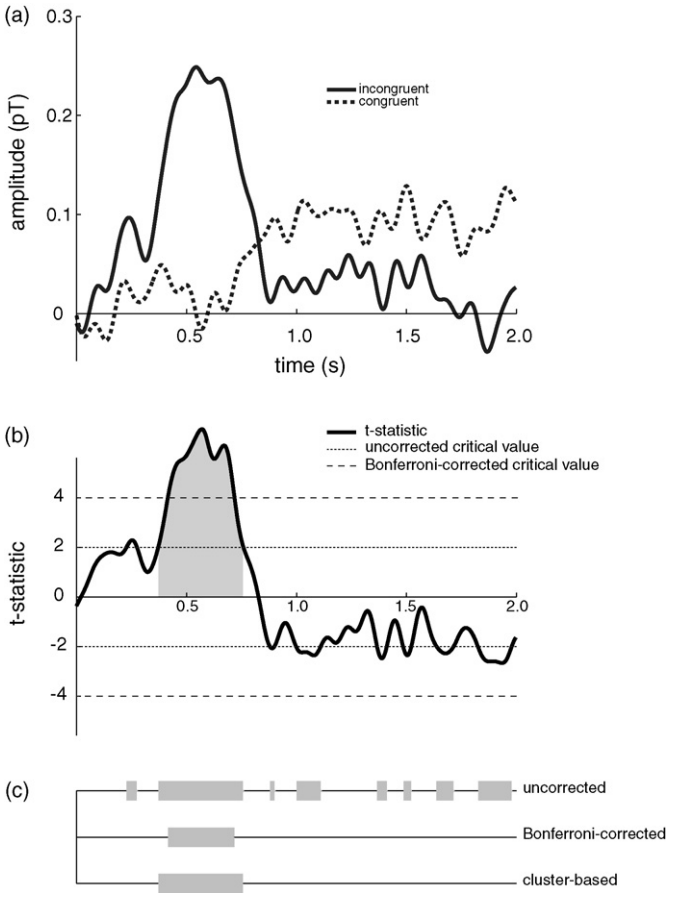


Fig. 1. Statistical testing of evoked responses at a single sensor. In panel a, the evoked responses are shown, separately for congruent (dotted line) and incongruent (solid line) sentence endings. In panel b, the time series of sample-specific $t$-values is shown. And in panel c, the significant samples are shown, separately for each of three statistical procedures: (1) sample-specific $t$-tests at the uncorrected 0.05-level (two-sided), (2) sample-specific $t$-tests at the Bonferroni-corrected level of $0.05/600 = 0.00008$ (two-sided), and (3) the cluster-based nonparametric test.

whether this raw effect is larger than can be expected on the basis of chance alone. Because the raw effect is observed at multiple time samples, an obvious first step is to calculate multiple sample-specific $t$-values (see Fig. 1, panel b). On the first line of panel c in Fig. 1, we show the time samples for which the sample-specific $t$-values exceed the critical value that corresponds to an alpha-level of 0.05. Equivalently, we could have calculated the sample-specific $p$-values and compared them with this critical alpha-level; in doing so, we detect the so-called significant $p$-values. Unfortunately, the FA rate of this procedure is larger than the critical alpha-level: under the null hypothesis, the probability of observing one or more significant $p$-values is larger than the critical alpha-level.[1] In fact, the larger the number of time samples, the more this probability approaches one. This is the MCP.

One way to solve the MCP is by lowering the critical alpha-level. Using the so-called Bonferroni inequality, it can be shown that a critical alpha-level of $0.05/C$ ($C$ is the number of time samples, which is 600 in our example data set) results in a FA rate that is less than 0.05. On the second line of panel c in Fig. 1, we show the time samples for which the sample-specific $p$-values exceed the Bonferrroni-corrected critical alpha-level. This approach is very conservative if the number of samples is large. We will illustrate this when we present the results of multi-sensor analyses.

We now consider the nonparametric statistical test. For this test, we use a test statistic that is based on clustering of adjacent time-samples that all exhibit a similar difference (in sign and magnitude). This test statistic was introduced by Bullmore et al. (1999) for the statistical analysis of structural MRI-data. In the fMRI-literature, it is called the cluster mass test. The calculation of this test statistic involves the following steps:

(1) For every sample, compare the MEG-signal on the two types of trials (semantically congruent versus semantically incongruent sentence endings) by means of a $t$-value (or some other number that quantifies the effect at this sample).
(2) Select all samples whose $t$-value is larger than some threshold. (This threshold may or may not be based on the sampling distribution of the $t$-value under the null hypothesis, but this does not affect the validity of the nonparametric test; see further.)
(3) Cluster the selected samples in connected sets on the basis of temporal adjacency.
(4) Calculate cluster-level statistics by taking the sum of the $t$-values within a cluster.
(5) Take the largest of the cluster-level statistics. (For our example data, the largest cluster-level statistic is indicated by the grey part in panel b of Fig. 1.)

The result from step 5 is the test statistic by means of which we evaluate the effect of semantic congruity. This is a test statistic for a one-sided test; for a two-sided test, we select test statistics whose absolute value is larger than some threshold (in step 2) and we take the cluster-level statistic that is largest in absolute value (in step 4). Also, for a two-sided test, the clustering in step 3 is performed separately for samples with a positive and a negative $t$-value.

The cluster-based test statistic depends on the threshold that is used to select samples for clustering. In our example, this threshold was the 97.5th quantile of a $\mathcal{T}$-distribution, which is used as a critical value in a parametric two-sided $t$-test at alpha-level 0.05. As will be shown later in Section 4, this threshold does not affect the FA rate of the statistical test. However, this threshold does affect the sensitivity of the test. For example, weak but long-lasting effects are not detected when the threshold is large.

The nonparametric statistical test is performed by calculating a $p$-value under the permutation distribution and comparing it with some critical alpha-level. The permutation distribution is obtained by the procedure described in Section 2.3: (1) collect the trials of the two experimental conditions in a single set, (2) randomly partition the trials in two subsets, (3) calculate the test statistic on this random partition, and (4) repeat steps 2 and

[1] To determine how much the FA rate exceeds the critical alpha-level, one has to know the statistical dependence between the $p$-values, which is typically unknown.

3 a large number of times and construct a histogram of the test statistics.

In the example data, there are eight clusters of time samples. These clusters are shown on the first line of panel c in Fig. 1. The first two of these clusters contain positive *t*-values and the others contain negative *t*-values. Using the cluster-based permutation test, only the largest positive cluster had a Monte Carlo *p*-value less than 0.025 (the critical alpha-level for a two-sided test). In fact, its Monte Carlo *p*-value was zero; none of the 1000 random partitions resulted in a cluster-level statistic that is larger in absolute value. This cluster is shown on the third line of panel c in Fig. 1.

It is important to note that the *p*-values for all eight clusters are calculated under the permutation distribution of the maximum (absolute value) cluster-level statistic and not under the permutation distribution of the second largest, third largest, etc. The choice for the maximum cluster-level statistic (and not the second largest, third largest, etc.) results in a statistical test that controls the FA rate for all clusters (from largest to smallest), but does so at the expense of a reduced sensitivity for the smaller clusters (reduced in comparison with a statistical test that is specific for the second, third, . . ., largest cluster-level statistic). We return to this point in Section 4.4

### 3.1.2. Multi-sensor analyses

Very often, one does not know at which sensors the effect can be observed. As compared to single-sensor analyses, our prior ignorance is much larger: instead of multiple time samples, we now have a much larger number of (sensor, time)-pairs (also called samples) at which we want to evaluate the effect. Again, it is an obvious step to calculate multiple sample-specific *t*-values in order to evaluate the reliability of the effect. However, as compared to the single-sensor analyses, the MCP is much larger here: we have 151 MEG sensors and 600 time samples, which results in 90,600 *t*-values. With this extremely large number of samples, Bonferroni correction results in a very conservative statistical test. In fact, in our example data, none of the sample-specific *p*-values exceeded the critical two-sided Bonferroni-corrected alpha-level of $0.025/90,600 = 0.0000003$. In contrast, the cluster-based permutation test turned out to be very sensitive.

The cluster-based permutation test for multi-sensor analyses is very similar to the one for single-sensor analyses. In fact, the calculation of the test statistic differs in a single aspect only: instead of clustering the selected time samples in connected sets on the basis of temporal adjacency, we now cluster the selected (sensor, time)-samples on the basis of spatial and temporal adjacency. (In the example study, we considered sensors to be neighbors if their distance is less than 4 cm.) We found 32 clusters of (sensor, time)-samples, 11 positive and 21 negative. Only two of these clusters have a Monte Carlo *p*-value less than 0.025, one positive and one negative. The combined (positive and negative) significant cluster extends over a time interval between 244 and 1630 ms after the onset of the last word.

In the top row of Fig. 2(a), we show the temporal evolution in the topography of the raw effect (i.e., the difference between the average MEG in the congruent and incongruent condition) over the time interval between 100 and 1300 ms. In the bottom row of Fig. 2(b), we show the same topography but now masked by the spatiotemporal pattern of the two significant clusters. This topography is consistent with the findings of previous studies that localized the primary source of the magnetic N400 effect in the left superior temporal sulcus (Simos et al., 1997; Helenius et al., 1998, 2002; Halgren et al., 2002) and the left prefrontal cortex (Halgren et al., 2002). By means of an arrow, we have indicated the approximate location of an equivalent current dipole that can explain the magnetic N400 effect.

## 3.2. Modulation of oscillatory activity

Just as the spatiotemporal evoked responses, the spatio-spectral–temporal oscillatory power estimates require a specialized statistical procedure that takes prior ignorance about the locus of the effect into account (ignorance with respect to the spatial, temporal, and spectral dimension). This can be realized by means of the same type of nonparametric statistical test as for the spatiotemporal evoked responses.

### 3.2.1. Single-sensor analyses

We used the multitaper method (Percival and Walden, 1993) to calculate time–frequency representations (TFRs) for our

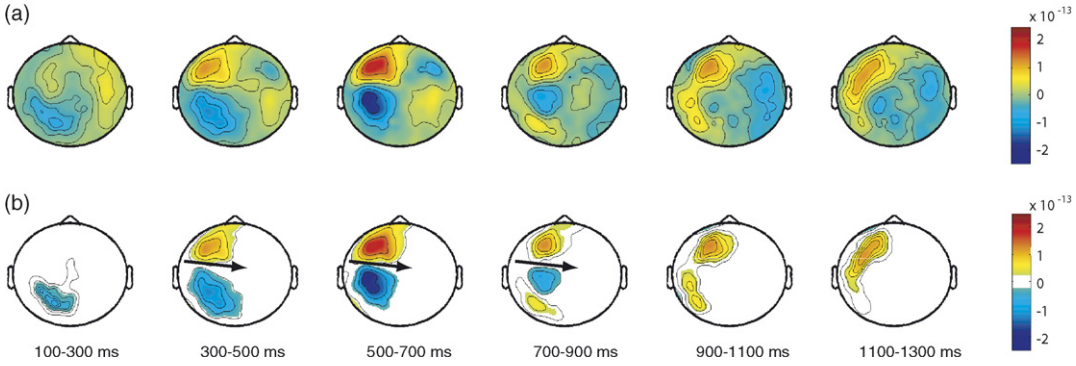


Fig. 2. (a) Temporal evolution of the topography of the raw effect (the difference between the average MEG in the congruent and the incongruent condition). Red denotes a positive and blue denotes a negative raw effect. The topography is shown for segments of 200 ms; within every segment, the raw effect was averaged. (b) Same topography as in the top row, but now masked by the spatiotemporal pattern of the two significant clusters. Masking involves that only the samples that belong to the two significant clusters are colored; all the other samples are transparent. The arrow denotes the approximate location of a dipole that may have generated the effect. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

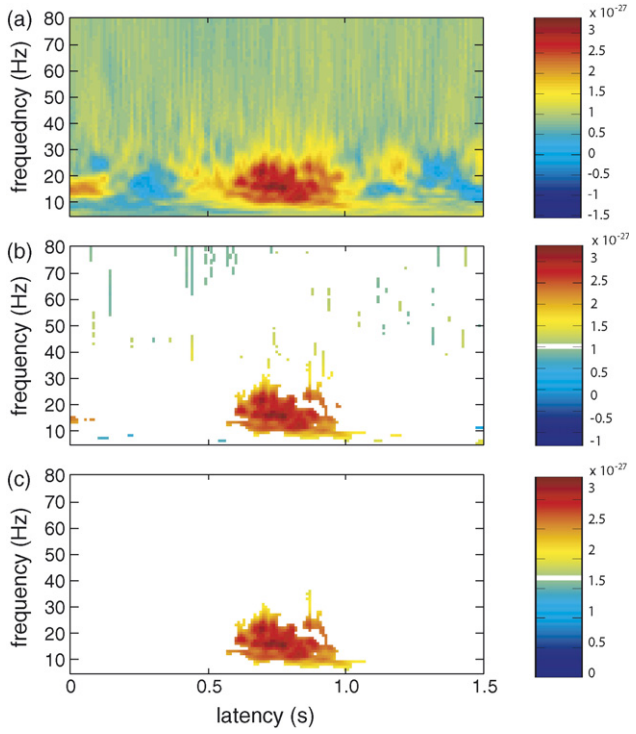


Fig. 3. The difference between the time–frequency representations (TFRs) for congruent and incongruent sentence endings, for a single sensor over left temporal cortex. (a) Simple difference between the two TFRs. (b) Difference between the two TFRs, masked by the spectral–temporal pattern of the significant sample-specific *t*-values (uncorrected). (c) Difference between the two TFRs, masked by the spectral–temporal pattern of the significant cluster. Red denotes a positive and blue denotes a negative raw effect. The TFRs are shown for the frequency range [5 Hz, 80 Hz] and the time interval [0 s, 1.5 s]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

example data. In Fig. 3, we show three ways of presenting the difference between the TFRs for congruent and incongruent sentence endings, for a single sensor over left temporal cortex. The upper panel (a) shows the raw effect, the simple difference between the two TFRs. Over the time interval from 0.6 to 1 s, brain responses to congruent sentence endings exhibit a stronger power in the lower beta band (from 10 to 20 Hz) than brain responses to incongruent sentence endings. To evaluate the reliability of this effect, we performed multiple sample-specific *t*-tests. The middle panel of Fig. 3(b) shows the TFR-difference masked by the spectral–temporal pattern of the significant sample-specific *t*-values (at the uncorrected alpha-level 0.05, two-sided). To solve the MCP, we applied Bonferroni correction. It turned out that none of (frequency, time)-specific *t*-values exceeds the Bonferroni corrected alpha-level. In contrast, the cluster-based permutation test turned out to be very sensitive.

The cluster-based permutation test for single-channel TFRs is very similar to the one for multi-sensor evoked responses. In fact, the calculation of the test statistic differs in a single aspect only: instead of clustering selected (sensor, time)-samples in connected sets on the basis of spatial and temporal adjacency, we now cluster the selected (frequency, time)-samples on the basis of spectral and temporal adjacency. In the example data, there are 75 clusters of (frequency, time)-samples, 42 positive and 33 negative. Only the largest positive cluster has a Monte Carlo *p*-value that is less than 0.025. The raw TFR-difference was masked by this spectral–temporal cluster and the resulting pattern is shown in the bottom panel of Fig. 3. This pattern is very similar to the one in the middle panel. Thus, we draw almost the same conclusion as on the basis of the middle panel, but do not suffer from the MCP.

### 3.2.2. *Multi-sensor analyses*

In multi-sensor analyses, we evaluate the effect at a much larger number of samples than in single-sensor analyses: instead of multiple (frequency, time)-samples, we now have a much larger number of (sensor, frequency, time)-samples. The cluster-based permutation test for multi-sensor analyses is very similar to the one for single-sensor analyses. In fact, the calculation of the test statistic differs in a single aspect only: instead of clustering the selected (frequency, time)-samples in connected sets on the basis of spectral and temporal adjacency, we now cluster the selected (sensor, frequency, time)-samples on the basis of spatial, spectral, and temporal adjacency.

In the example data, there are 519 clusters of (sensor, frequency, time)-samples, 309 positive and 210 negative. Only the largest positive cluster has a Monte Carlo *p*-value that is less than 0.025. The vast majority of the (sensor, frequency, time)-triplets in this cluster are in the beta band ([15 Hz, 30 Hz]). In the top row of Fig. 4(a), we show the temporal evolution in the topography of the raw effect for overlapping segments of 400 ms. This raw effect is the difference between the TFRs for the congruent and the incongruent condition, averaged over the frequencies in the beta band. In the bottom row of Fig. 4(b), we show the same topography but now masked by the significant cluster: the difference between the two TFRs was masked by the spatio-spectral–temporal pattern of the significant cluster, and this structure was subsequently converted into a spatiotemporal structure by averaging over the frequencies in the beta band. Two aspects of this topography should be mentioned: (1) over time, the location of the effect changes from the right to the left hemisphere, and (2) in the time interval [500 ms, 1500 ms], the effect is largest over the area that also shows an effect with respect to the evoked responses (see Fig. 2).

## 4. Justification

Until now, we have deliberately ignored three important issues: (1) the exact specification of the null hypothesis that is tested by the nonparametric statistical test, (2) the proof that this test controls the FA rate, and (3) the issue of how to choose a test statistic. The theory of nonparametric statistical tests is not well documented and not very accessible. Surprisingly, the central argument in this theory (the so-called conditioning rationale, see further) is rather intuitive and can easily be made accessible to the neuroscience community. This argument can be found in the introductory chapter of a recent book on permutation tests (Pesarin, 2001), but it also appears in the context of parameter estimation for models of achievement test data (Maris, 1998). However, it is not clear who deserves the credit for this argument.

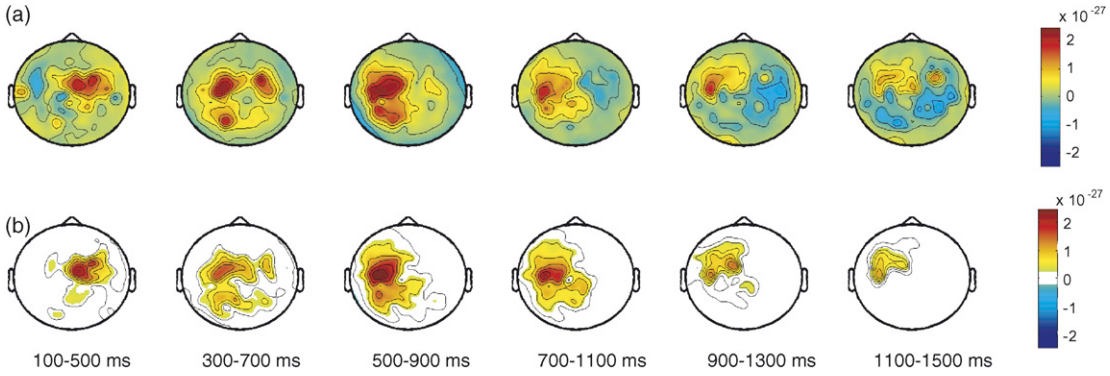


Fig. 4. (a) Temporal evolution of the topography of the raw effect (the difference between the TFRs for the congruent and the incongruent condition, averaged over the frequencies in the beta band [15 Hz, 30 Hz]). Red denotes a positive and blue denotes a negative raw effect. The topography is shown for overlapping segments of 400 ms. These segments are overlapping because the power spectra were calculated on overlapping time-windows. (b) Same topography as in the top row, but now masked by the spatio-spectral–temporal pattern of the significant cluster. After masking, the spatio-spectral–temporal structure was converted into a spatiotemporal structure by averaging over the frequencies in the beta band. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

To make the argument accessible to the neuroscience community, we need some definitions, and these are introduced now. To keep this paper self-contained, there will be some overlap with Maris et al. (2007) in the remaining of this section.

### 4.1. The structure in the data

In this section, we only consider single-subject MEEG-studies. Multiple-subject studies will be considered in section 5. In single-subject MEEG-studies, the units of observation are trials that belong to different experimental conditions and the research question is about the effect of these experimental conditions on the MEEG. The trails can be assigned to the experimental conditions according to two schemes: (1) the between-trials design, in which every trial is assigned to one of a number of experimental conditions, or (2) the within-trials design, in which every trial is assigned to all experimental conditions in a particular order. The between-trials design is by far the most common in practice. In fact, there is only one type of within-trials study that is performed regularly: the within-trials activation-versus-baseline study. This type of study involves multiple trials that consist of a baseline (the interval preceding the stimulus) and an activation condition (the interval following the stimulus), which have to be compared. In this section, we only consider the between-trials design. Nonparametric statistical testing for single-subject within-trials studies proceeds along the same lines as for multiple-subject within-subjects studies. We will briefly return to this point in section 5.

To describe the structure in the data, we make the usual distinction between a dependent and an independent variable. In MEEG-studies, the dependent variable is the recorded EEG or MEG. This variable is denoted by $D$, and it is assumed to be a random variable. This means that we consider $D$ as a variable whose value is the result of a random process. The value of $D$ that was actually observed in the experiment (the realization of $D$) is denoted by $d$. In a between-trials MEEG-study, the dependent variable $D$ is an array of $n$ smaller component data structures $D_r$ $(r = 1, \ldots, n)$, each one corresponding to one trial: every component $D_r$ is a spatiotemporal data matrix observed in a given trial.

The independent variable specifies the different experimental conditions. In the example, there are two experimental conditions: semantically congruent and semantically incongruent sentence endings. In general, the experimental conditions can differ with respect to a number of factors: stimulus type, task type, response type, characteristics of the data in an epoch prior to the dependent variable, etc. The independent variable is denoted by $I$. In a between-trials study, $I$ is an array of $n$ smaller components $I_r$ $(r = 1, \ldots, n)$, each one corresponding to one trial: every component $I_r$ denotes the condition to which the trial belongs. For instance, $I_r$ equals 1 if the trial belongs to the semantically congruent condition and 2 if it belongs to the semantically incongruent condition. The independent variable $I$ can be both random and fixed, but at this point it is not necessary to make this distinction. Later, we will return to this issue.

### 4.2. The null hypothesis

#### 4.2.1. Formulation

The null hypothesis of a permutation test involves the probability distributions of the trial-specific data structures $D_r$. These probability distributions are denoted by $f(D_r = d_r)$, and abbreviated by $f(D_r)$. These probability distributions do not have to be of a familiar type (e.g., normal, binomial, Poisson). Instead, we only need the assumption that there is some rule $f$ that assigns probabilities $f(D_r = d_r)$ to all possible realizations $d_r$; we do not have to know what this rule is. Now, the null hypothesis of a permutation test involves that all $n$ probability distributions $f(D_r)$ $(r = 1, \ldots, n)$ are equal:

$$f(D_1) = f(D_2) = \cdots = f(D_n). \quad (1)$$

In other words, the null hypothesis involves that all trial-specific data structures $D_r$ are drawn from the same probability distribution, regardless of the experimental condition in which they were observed ($I_r = 1$ or $I_r = 2$).

The same null hypothesis is also tested using the familiar parametric statistical tests (i.e., the *t*-, the *F*-test, and their multivariate generalizations). This may sound unfamiliar, because in statistics handbooks the parametric null hypothesis is formulated as equality of the two conditions with respect to some parameter of the probability distribution (typically, the expected value, but also the variance, the covariance, etc.). However, the familiar parametric statistical tests also make auxiliary assumptions about the probability distributions in the two conditions (i.e., normality and equal variances), and together with the null hypothesis of interest (equality with respect to some parameter of interest) this implies equality of the complete probability distributions.

#### 4.2.2. Strong and weak control of the FA rate

It is important to keep in mind that the data structures $D_r$ are spatiotemporal. This implies that, under the null hypothesis in Eq. (1), the probability distribution of the MEEG is identical for all (sensor, time)-pairs. If this null hypothesis is rejected, one concludes that the probability distribution of the MEEG is modulated by the experimental conditions for at least some (sensor, time)-pairs. With respect to the localization of this effect, it is not possible to control the FA rate at the level of a single (sensor, time)-pair. This issue was introduced in the fMRI-literature by Holmes (1994) (see also, Friston et al., 1996). In the fMRI-literature, the MCP involves inference at a large number of voxels in a three-dimensional volume. In this context, it makes sense to distinguish between weak and strong control of the FA rate. The null hypothesis behind weak control of the FA rate is equivalent to the one in Eq. (1): no difference between the experimental conditions for none of the voxels. We will call this the global null hypothesis. The null hypothesis behind strong control of the FA rate is voxel-specific: no difference between the experimental conditions for a given voxel but unknown differences for the other voxels. If it is possible to control for all voxels the probability of incorrectly rejecting this voxel-specific null hypothesis, then we have strong control of the FA rate. With strong control of the FA rate, we can quantify the uncertainty in the localization of the effect. The price that has to be paid for this stronger control, is a reduced sensitivity for detecting violations of the global null hypothesis.

In contrast to fMRI-data, for MEEG-data it does not make sense to distinguish between a global and a sample-specific null hypothesis. This is because the spatial correlation between MR-signals is much weaker than between MEEG-signals. In fact, the MR-technology makes use of spatial magnetic field gradients that allow the BOLD-response at a particular voxel to be measured with high spatial specificity. In contrast, the MEEG at a particular sensor is produced by physiological sources (usually characterized mathematically as current dipoles) that also affect the MEEG at the other sensors. As a consequence, if the experimental conditions differ with respect to the physiological sources that produce the MEEG, then this effect is present at all sensors (however, with a different magnitude at different sensors). This has an important implication: if a sensor-specific null hypothesis is false for one sensor, then it is also false for the other sensors. For this reason, it does not make sense to distinguish between a global and a sensor-specific null hypothesis.

For the temporal dimension in the MEEG-data, the argument against sample-specific null hypotheses (a sample is a (sensor, time)-pair) is a bit different. Although the temporal resolution of the MEEG is excellent, it very often does not make sense to test null hypotheses at the level of individual time points (also called time samples). This is because the duration of most effects involves several tens and often several hundreds of time points. If we want strong control of the FA rate, we have to test time-point-specific null hypotheses. Given the duration of most effects, this would result in a strongly reduced sensitivity for detecting violations of the global null hypothesis. In the following, we will restrict ourselves to this global null hypothesis and the associated weak control of the FA rate.

#### 4.2.3. Exchangeability

Very often, researchers are willing to make the assumption of statistical independence between the trials. In fact, this assumption is always made if one uses parametric statistical tests in between-trials studies. The assumption of statistical independence will be violated if the MEEG-data in one trial depend on the MEEG-data in another trial. A biologically plausible form of statistical dependence is temporal autocorrelation: correlation between the MEEG-data in neighboring trials. To avoid temporal autocorrelation, it is good practice to have the trials separated by some minimum time interval (determined by the lag of the temporal autocorrelation). In this paper, as in parametric statistics, we make the assumption of statistical independence between the trials. We need this assumption to show that the permutation test is a valid test of the null hypothesis of identical distributions in Eq. (1).

From the null hypothesis of identical distributions together with the assumption of statistical independence, it follows that the probability distribution of the dependent variable $D$, $f(D) = f(D_1, D_2, \ldots, D_n)$, is exchangeable. Exchangeability means that the probability of $D$ is invariant under permutation of the component data structures $D_r$. Exchangeability is a useful concept because it allows us to show the validity of the permutation test in a straightforward way. In the following, we will present the permutation test as a statistical test of exchangeability, and not as a statistical test of the null hypothesis of identical distributions. However, this is just a matter of presentation: under the assumption of statistical independence, the null hypothesis of identical distributions and exchangeability are equivalent.

### 4.3. The permutation test

In principle (but not in practice), one could test the hypothesis of exchangeability by constructing the probability distribution of some test statistic under this hypothesis, and by evaluating the actually observed test statistic under this distribution. However, it turns out to be much easier to construct a particular conditional probability distribution of the test statistic (also under the hypothesis of exchangeability). This conditional probability distribution is the permutation distribution and the resulting statistical test is the permutation test. As will be shown in the

following, using a conditional instead of the unconditional probability distribution results in exactly the same FA rate. Before introducing the permutation distribution, we first describe a procedure that effectively draws from it.

#### 4.3.1. Drawing from the permutation distribution

Drawing from the permutation distribution involves randomly permuting the components of $d$, the realization of the random dependent variable $D$. For instance, in a study with four trials, $d$ has the following structure: $(d_1, d_2, d_3, d_4)$. In a permutation test, the data matrices in $d$ are randomly permuted in such a way that every permutation of $d$ has the same probability. With four trials, there are $4! = 24$ different permutations, and they all have a probability of 1/24.

Very often, it is sufficient to perform random partitions instead of random permutations. This is the case for all test statistics for which the order of the trial-specific data matrices within the conditions is irrelevant. For instance, the cluster-based test statistic that was used in the example analyses (i.e., the maximum over the clusters of the cluster-level statistics) is of this type. To show this, assume that the first two trials belong to the semantically congruent condition, and the last two belong to the semantically incongruent condition. Now, the cluster-based test statistic is identical for the following four permutations: $(d_1, d_3, d_2, d_4)$, $(d_3, d_1, d_2, d_4)$, $(d_1, d_3, d_4, d_2)$, and $(d_3, d_1, d_4, d_2)$. This is because the sample-specific $t$-values for the trial pairs $(d_1, d_3)$ and $(d_2, d_4)$ do not depend on the order of the trials within the pairs, and therefore the same holds for the cluster-level statistics (sums of sample-specific $t$-values) and their maximum. As a consequence, the permutation distribution of the test statistic is identical to the so-called partitioning distribution, which is obtained by randomly partitioning the trials into two sets. The number of different partitions is equal to the so-called multinomial coefficient, which depends on the number of trials in each of the two conditions. In the mini-example above, there are two trials in every condition, and the multinomial coefficient is equal to $(4!/(2!2!)) = 6$. In the following, we will not make a distinction between the permutation and the partitioning distribution; one should remember that the permutation and the partitioning distribution are identical if the test statistic is independent of the order of the trials within the conditions.

#### 4.3.2. The permutation p-value is a conditional p-value

The permutation $p$-value is the $p$-value that is obtained in a permutation test. The permutation $p$-value is a conditional $p$-value because it is calculated under a conditional distribution. To show this, let $f(D)$ be the unknown probability distribution of the dependent variable $D$. Exchangeability involves that $f(D)$ is invariant under permutation of the trial-specific data matrices $D_r$. Now, the permutation distribution is the conditional distribution of $D$ given the unordered set of trial-specific data matrices $D_r = d_r$. This unordered set is denoted by $\{D\} = \{d\}$. In a study with four trials, $d = (d_1, d_2, d_3, d_4)$, the unordered set $\{d\}$ is the collection of all permutations of $(d_1, d_2, d_3, d_4)$: $(d_1, d_2, d_3, d_4)$, $(d_1, d_2, d_4, d_3)$, $(d_1, d_4, d_3, d_2)$, plus 21 more. The conditional distribution of $D$ given the unordered set $\{D\} = \{d\}$ is denoted by $f(D|\{D\} = \{d\})$. Now, if the unknown distribution $f(D)$ is exchangeable, then the conditional distribution $f(D|\{D\} = \{d\})$ is the permutation distribution, which is known. In other words, if $f(D)$ is exchangeable, then the draws from $f(D|\{D\} = \{d\})$ are permutations of the observed array $d$, and each of these permutations has the same probability.

The previous paragraph was about a conditional probability distribution of the dependent variable $D$. However, in statistical testing, we are not interested in the complete $D$, but in some test statistic, which is a function of $D$ and $I$, the independent variable. This test statistic is random and it is denoted by $S(D, I)$. The test statistic that was actually observed in the experiment (the realization of $S(D, I)$) is denoted by $S(d, I)$. Now, because we can draw from the conditional distribution $f(D|\{D\} = \{d\})$, we can calculate $f(S(D, I)|\{D\} = \{d\})$, the conditional distribution of $S(D, I)$ given $\{D\} = \{d\}$. In Section 2, we have described how $f(S(D, I)|\{D\} = \{d\})$ can be approximated by randomly partitioning the trials and constructing a histogram of the test statistics $S(D, I)$. The Monte Carlo $p$-value is calculated under this histogram, and therefore it is a conditional $p$-value. In Fig. 5, we give a schematic representation of the permutation test in which we refer to the fact that, under exchangeability, $f(D|\{D\} = \{d\})$ is the permutation distribution.

The permutation test is based on a $p$-value that is calculated under the conditional distribution $f(S(D, I)|\{D\} = \{d\})$. Therefore, the permutation test controls the FA rate in the following conditional sense: given the unordered set $\{D\} = \{d\}$, under exchangeability, the probability of observing a $p$-value that is less than the critical alpha-level is exactly equal to the critical alpha-level.

#### 4.3.3. The permutation test controls the false alarm rate unconditionally

At first sight, controlling the FA rate in this conditional sense (i.e., conditional on $\{D\} = \{d\}$) is not very appealing. After all, who is interested in the conditional FA rate of a statistical test

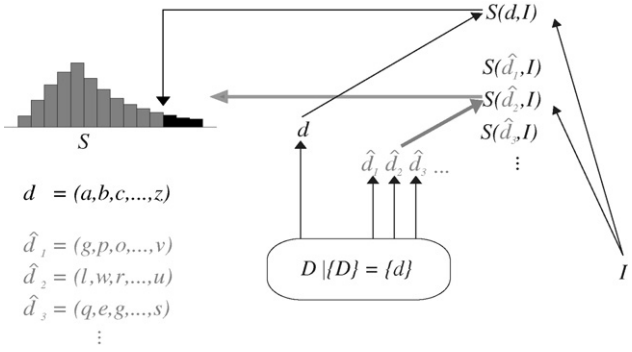


Fig. 5. Schematic representation of the permutation test. We use a box to denote the random variable $D|\{D\} = \{d\}$. The observed realization of $D$ (i.e., $d$) is printed black, and the draws from $f(D|\{D\} = \{d\})$ that are used to construct the permutation distribution of the test statistic (i.e., $\hat{d}_1, \hat{d}_2, \hat{d}_3$, etc.) are printed grey. The observed test statistic is denoted by $S(d, I)$, and the draws from the permutation distribution by $S(\hat{d}_1, I)$, $S(\hat{d}_2, I)$, $S(\hat{d}_3, I)$, etc. The permutation distribution of the test statistic is shown as an histogram, and the $p$-value is denoted by the black tail-area under the permutation distribution. In the lower-left corner, we show possible values for $d$, $\hat{d}_1$, $\hat{d}_2$, and $\hat{d}_3$, which are all permuted versions of the same set of lowercase letters. Each lowercase letter represents the data that was observed in a single trial.

given an event that occurs so rarely ($\{D\} = \{d\}$, the data that were observed in this experiment, but regardless of the trial order)? However, what matters is not this rare event, but the properties of a decision that is made on the basis of this $p$-value. The decision is about exchangeability of the probability distribution of $D$: if the permutation $p$-value is less than some critical alpha-level, this hypothesis is rejected; otherwise, it is maintained. The FA rate is a property of this decision rule. Now, the FA rate is equal to the critical alpha-level, regardless of whether the $p$-value has a conditional or an unconditional interpretation. This is because, for each of the events $\{D\} = \{d\}$ on which we condition, the FA rate is equal to the same critical alpha-level. Therefore, if we average over the probability distribution of $\{D\}$, the FA rate remains equal to this critical alpha-level.

This can also be shown in a short derivation. In this derivation, the FA rate under the conditional distribution $f(D|\{D\} = \{d\})$ is denoted by $P(\text{Reject } H_0|\{D\} = \{d\}, H_0)$ and the FA rate under the distribution $f(D)$ by $P(\text{Reject } H_0|H_0)$. We also use $\sum_{\{d\}}$ to denote the sum over all realizations of $\{D\}$ and $\alpha$ to denote the critical alpha-level:

$$
\begin{aligned}
&P(\text{Reject } H_0|H_0) \\
&\quad = \sum_{\{d\}} P(\text{Reject } H_0|\{D\} = \{d\}, H_0) f(\{D\} = \{d\}) \\
&\quad = \sum_{\{d\}} \alpha f(\{D\} = \{d\}) = \alpha
\end{aligned}
$$

In the first line of this derivation, we make use of the following equality from elementary probability theory: $P(A) = \sum_b P(A|B = b)P(B = b)$. And in the third line, we make use of the fact that the probabilities $f(\{D\} = \{d\})$ sum to 1.

We can conclude that an FA rate that is controlled under the conditional distribution $f(D|\{D\} = \{d\})$ is also controlled under the corresponding unconditional distribution $f(D)$. This conclusion is a special case of the following general fact: for every event (in our case, falsely rejecting the null hypothesis) whose probability is controlled under a conditional distribution, also the probability under the corresponding unconditional distribution is controlled. This general fact will be called the conditioning rationale. Note that the conditioning rationale is used to prove the unconditional control of the FA or type 1 error rate, and not the unconditional control of the type 2 error rate (i.e., the probability that null hypothesis is maintained while in fact the alternative hypothesis is true). This is similar to classical parametric statistics, in which only the type 1 error rate is controlled.

#### *4.3.4. The permutation test for a random independent variable*

Until now we have not made a distinction between random and fixed independent variables. For an empirical neuroscientist who only wants to apply permutation tests there is no need to make this distinction, because the calculations are identical for both types of independent variables. However, a methodologist may be interested in the rationale behind this fact. We now describe the difference between random and fixed independent variables. An independent variable $I$ is random if a replication of the experiment may show a different value of $I$ with some probability (possibly unknown). This can happen in two ways: (1) the experimenter assigns the trials to the experimental conditions by means of a randomization mechanism (which usually calls a random number generator), and (2) the independent variable depends on the subject's behavioral response (e.g., accuracy, speed). When $I$ is a random variable, we have to make a distinction between the random variable itself and its realization, i.e. the value that was actually observed. The realization of $I$ is denoted by $i$.

An independent variable $I$ is fixed if a replication of the experiment always shows the same value of $I$. This is the case if the experimenter assigns the trials to the experimental conditions according to a fixed scheme (e.g., a fixed pattern that is repeated every $x$ trials). Until now, we have tacitly assumed that the independent variable was fixed; only the dependent variable $D$ was considered random.

If both the dependent and the independent variable are random, then we have to give a rationale for the permutation test in terms of the joint probability distribution $f(D, I)$ instead of $f(D)$. It turns out that this rationale is very simple if the random independent variable is treated as if it is fixed. In probability theory, this conceptual move is called conditioning on the random independent variable. Conditioning on the random independent variable involves that we express our hypothesis in terms of the conditional probability distribution of the biological data $D$ given the assignment $I = i$, which is denoted by $f(D|I = i)$. Now, our hypothesis involves that $f(D|I = i)$ is exchangeable for all realizations $i$.

We can use the conditioning rationale to show that conditioning on a random independent variable does not affect the FA rate. We begin by observing that the permutation $p$-value is calculated under the double conditional distribution $f(D|\{D\} = \{d\}, I = i)$, which is the permutation distribution under exchangeability of $f(D|I = i)$. A statistical test based on this $p$-value controls the FA rate under the conditional distribution $f(D|\{D\} = \{d\}, I = i)$ and, because of the conditioning rationale, also under the unconditional distribution $f(D)$.

### *4.4. The choice of a test statistic*

False alarm rate control by means of a nonparametric statistical test does not depend on the test statistic that is used. This is an enormous advantage of nonparametric over parametric statistical testing. In parametric statistics, one can only use test statistics whose sampling distribution under the null hypothesis is known. In practice, this constraint forces us to use a test statistic whose sampling distribution is known under multivariate normality. In contrast, in nonparametric statistics, one is free to choose any test statistic that one considers appropriate. This freedom has at least four advantages: (1) it provides a simple way to solve the MCP, (2) it allows us to incorporate prior knowledge about the type of effect that can be expected, (3) it allows us to localize the effect on the spatial, spectral, and temporal dimension (however, without strong control of the FA rate), and (4)

it allows us to make use of test statistics with a vector-valued outcome.

#### 4.4.1. A solution for the MCP

In the nonparametric statistics, the MCP is solved in the following way: instead of evaluating the difference between the experimental conditions for each of the samples separately, it is now evaluated by means of a single test statistic for the complete spatio-(spectral)–temporal grid. Thus, the multiple comparisons (one for every sample) are replaced by a single comparison, and therefore the MCP does not exist any more.

#### 4.4.2. Incorporating prior knowledge

Incorporating prior knowledge about the type of effect that can be expected will increase the sensitivity of the test. For instance, when comparing spatiotemporal data matrices in two experimental conditions, one can make use of the fact that adjacent (sensor, time)-pairs are likely to exhibit the same effect. Therefore, it makes sense to use a test statistic that is based on a clustering of these adjacent (sensor, time)-pairs, such as the size of the largest connected cluster that exceeds some threshold, or the sum of the $t$-values in that cluster.

To examine the differential sensitivity of the different test statistics, we analyzed the example data set by means of eight different test statistics, two cluster-based and six non-cluster-based test statistics. The sensitivity of a test statistic was quantified by the minimum number of trials that is required to obtain a significant effect of semantic congruity in the multi-sensor analysis of the evoked responses. Details of this small sensitivity study are given in the supplementary material. It was found that the two cluster-based test statistics (the largest within-cluster summed $t$-value, and the size of the largest cluster) required approximately 30% of the number of trials that is required by the non-cluster-based test statistics.

#### 4.4.3. Localization by means of the maximum-statistic

When comparing spatiotemporal data structures in two conditions, one is almost always interested in the spatiotemporal localization (where and when) of the effect. Usually, the interest in the null hypotheses of exchangeability is only indirect: one is interested in these null hypotheses because they are violated by some localized effect. There is a conflict between this interest in localized effects and our choice for a global null hypothesis: by controlling the FA rate under this global null hypothesis one cannot quantify the uncertainty in the spatiotemporal localization of the effect. On the other hand, as argued in Section 4.2.2, it usually does not make sense to test sensor-specific or time-point-specific null hypotheses.

Instead of opting for strong control of the FA rate over the spatial and the temporal dimension, we propose a localization procedure that is much less ambitous. An important motivation for this procedure is the fact that cluster-based test statistics turn out to be very sensitive. After thresholding, cluster-level statistics are calculated by taking, for instance, the size of the cluster or the sum of the $t$-values within the cluster. These cluster-level statistics will be called ClusterStats. Our localization procedure involves identifying clusters on the basis of their ClusterStat. To control the FA rate of the localization procedure, we need a critical value for the ClusterStats with the following property: under the null hypothesis, the probability that one or more ClusterStats exceed the critical value CV, is controlled at some critical alpha-level. Formally,

$$P(\text{at least one ClusterStat} \geq \text{CV}) = \alpha.$$

This is equivalent to an equation in terms of the maximum-function (Max):

$$P(\text{Max(ClusterStat)} \geq \text{CV}) = \alpha.$$

Thus, the critical value for the Max(ClusterStat)-statistic can be used to identify significant clusters while controlling the FA rate.

### 4.5. Vector-valued test statistics

In situations where several different effects can co-occur, it is natural to use a vector-valued test statistic. For instance, when comparing the spatiotemporal data matrices of two experimental conditions, there may be multiple spatiotemporal clusters that exhibit a significant difference. In this situation, it makes sense to quantify the effect by an ordered sequence of ClusterStats. This is our vector-valued test statistic. Because very small clusters are unlikely to reflect important physiological activity, this vector only contains the ClusterStats of clusters that have some minimum size (chosen a priori).

Contrary to a parametric statistical test, it is straightforward to construct a nonparametric statistical test on the basis of a vector-valued test statistic. In particular, with a vector-valued test statistic, we obtain a multivariate permutation distribution under which we can identify a multivariate tail area that contains a probability volume equal to the desired FA rate. This multivariate tail area is defined by a vector-valued critical value that controls the following probability: the probability of observing a vector-valued test statistic whose elements exceed the critical value on one or more dimensions. If the vector-valued test statistic is an ordered sequence of ClusterStats, the first dimension corresponds to the cluster with the largest ClusterStat, the second dimension to cluster with the second largest ClusterStat, etc.

We applied such a vector-valued statistical test to the evoked responses in the example data. Clusters of less than 250 (sensor, time)-pairs were considered too small to be of interest. There were three significant clusters: the two clusters that were also significant in the analysis with the Max(ClusterStat)-statistic, plus an additional negative cluster over right temporal sensors in the time interval 900–1100 ms. In the supplementary material, we show a figure of the topography of the raw effect masked by the spatiotemporal pattern of the three significant clusters.

## 5. The permutation test for multiple-subject MEEG studies

In practice, one is often interested in a null hypothesis about a population of subjects, instead of a single subject. In a very

similar way as for a single-subject MEEG study, the null hypothesis about a population can be tested by means of a permutation test. To test the null hypothesis at the level of a population, a sample of subjects is drawn from this population. The first step involves taking the average[2] over all trials within every subject, which produces subject-specific evoked responses or average power. There are two types of multiple-subject studies: in a between-subjects study, every subject is observed in one experimental condition, and in a within-subjects study, every subject is observed in all experimental conditions (in a particular order). We will first consider between-subjects and then within-subjects studies.

A permutation test for a between-subjects MEEG study involves exactly the same calculations as a permutation test for a between-trials single subject MEEG study. The only difference is that the calculations are now performed on the subject-specific averages of a group of subjects instead of the trial-specific MEEG data of a single subject. Typically, the number of subjects in a between-subjects study is much smaller than the number of trials in a between-trials single subject study. This has consequences for the calculation of the permutation $p$-value: if the number of subjects is not too large, the permutation $p$-value can be calculated exactly by enumeration. For instance, with 16 subjects, 8 in every experimental condition, there are 12870 (i.e., $\binom{16}{8}$) possible values under the permutation distribution for one of the cluster-based test statistics. Note that, with a small number of subjects, it may be impossible to reach significance, regardless of the difference between the conditions. For instance, with four subjects, two in every condition, the smallest possible $p$-value is 0.16667.

With this permutation test, we test a null hypothesis about the probability distributions of the subject-specific averages. This null hypothesis involves that all subject-specific averages are drawn from the same probability distribution, regardless of the experimental condition in which they were observed. Thus, the null hypothesis is about the probability distribution from which the subjects are drawn. Because we can make a statement about a probability distribution that characterizes a population of subjects, the permutation test allows us to generalize from a sample to a population. This is an interesting conclusion because the permutation distribution under which we calculate our $p$-value is specific for our sample. (This is very different from the sampling distribution of a parametric statistical test, which does not depend on the values observed in our sample.)

In a within-subjects MEEG study, every subject has one subject-specific average for each experimental condition. For simplicity, we assume there are only two experimental conditions. Then, the subject-specific averages for the $r$ th subject are a pair $(D_{r1}, D_{r2})$, with $D_{r1}$ the data observed in the first experimental condition, and $D_{r2}$ the data observed in the second. In the calculation of the cluster-based test statistic, the sample-specific statistical values are now obtained from the formula for the paired-samples (dependent-samples) instead of the independent-samples $t$-value, which is used in a between-subjects MEEG study. The rest of the calculation is identical. The construction of the permutation distribution is again different for a between- and a within-subjects MEEG study. Instead of randomly permuting the subjects (such that they become associated with different experimental conditions), we now randomly permute the subject-specific averages $(D_{r1}, D_{r2})$ within every subject. Moreover, this random permutation is performed independently for every subject. With $n$ subjects, this results in a permutation distribution with $2^n$ possible values that are all equally probable under the null hypothesis (see next paragraph). The $p$-value is then obtained by locating the observed test statistic under this permutation distribution.

The hypothesis of interest in a within-subjects MEEG study is about the probability distribution of the subject-specific averages in the different experimental conditions. Let the joint probability distribution of the subject-specific averages be denoted by $f(D_{r1}, D_{r2})$. Now, the null hypothesis of a within-subjects permutation test involves that this joint distribution is exchangeable:

$$f(D_{r1}, D_{r2}) = f(D_{r2}, D_{r1}).$$

The null hypothesis of exchangeability implies that the marginal distributions for the two experimental conditions, $f(D_{r1})$ and $f(D_{r2})$, are equal. This is our hypothesis of interest. Thus, the permutation test for a within-subjects MEEG study tests the null hypothesis of exchangeability of the joint distribution $f(D_{r1}, D_{r2})$, and this null hypothesis will be violated if the marginal distributions $f(D_{r1})$ and $f(D_{r2})$ are different.

The permutation test for a within-subjects MEEG study can also be applied to the data of a single-subject within-trials MEEG study. In practice, there is only one common type of within-trials MEEG study: the within-trials activation-versus-baseline study, in which every trial consists of a baseline (the interval preceding the stimulus) and an activation condition (the interval following the stimulus). The permutation test for a within-trials study involves the same calculations as for a within-subjects study; instead of applying these calculations to subject-specific averages (in the different experimental conditions), they are now applied to trial-specific data (in the baseline and the activation condition). Note that this requires the baseline and the activation period to be equally long, since otherwise there is not a one-to-one correspondence between the samples in the baseline and the activation period.

## 6. Conclusions

We have shown how nonparametric statistical tests can be used to evaluate different effect types that are studied in the MEEG-literature (i.e., single-sensor and multi-sensor evoked responses and time–frequency representations). We have also presented a theory for these nonparametric statistical tests,

[2] In principle, the permutation test does not require that the trial-specific data have to be combined by means of an average; other ways of combining the trial-specific data are also possible. However, most null hypotheses of interest involve an average over the trial-specific data.

which demonstrates their validity in a rigorous way. This theory applies to both single-subject and multiple-subject studies. The null hypothesis of a nonparametric statistical test involves that the probability distributions of the MEEG-data in the different experimental conditions are equal. In the theoretical justification, this null hypothesis is linked to exchangeability, which plays the role of an intermediate concept that allows us to demonstrate the validity of the permutation test.

The main advantage of nonparametric testing is the freedom to use any test statistic one considers appropriate. This freedom allows us to solve the MCP in a simple way, and it also allows us to incorporate prior knowledge about the type of effect that can be expected. The latter may result in a drastic increase of the sensitivity of the statistical test, as is illustrated using the data of our example study.

Despite their advantages, cluster-based nonparametric tests are not a panacea. First, the results of the cluster-based nonparametric tests depend on the threshold that is used for selecting the samples that will subsequently be clustered. It is not clear how to choose this threshold to obtain maximum sensitivity for the unknown effect that is present in the data: for a weak and widespread effect, the threshold should be low, and for a strong and localized effect, the threshold should be high. Clearly, if the threshold is chosen on the basis of the data, the FA rate will not be controlled (unless Bonferroni-correction is used to control for the multiple thresholds). Second, the sensitivity of a cluster-based nonparametric test will always be less than that of the so-called uncorrected *p*-value approach, in which the null hypothesis is rejected if at least one sample-specific *t*-value exceeds the threshold. Clearly, the uncorrected *p*-value approach does not control the FA rate, and the cluster-based nonparametric test is therefore superior in this respect. In that sense, the cluster-based nonparametric tests trade in some sensitivity for FA rate control, just as other approaches that deal with the MCP.

## Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jneumeth.2007.03.024.

## References


Achim A. Statistical detection of between-group differences in event-related potentials. Clin Neurophysiol 2001;112:1023–34.

Blair RC, Karniski W. An alternative method for significance testing of waveform difference potentials. Psychophysiology 1993;30:518–24.

Bullmore E, Brammer M, Williams SCR, Rabe-Hesketh S, Janot N, David A. Statistical methods of estimation and inference for functional MR image analysis. Magn Reson Med 1996;35:261–77.

Bullmore E, Suckling J, Overmeyer S, Rabe-Hesketh S, Taylor E, Brammer M. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Trans Med Imag 1999;18:32–42.

Chau W, McIntosh AR, Robinson SE, Schulz M, Pantev C. Improving permutation test power for group analysis of spatially filtered MEG data. NeuroImage 2004;23:983–96.

Ernst MD. Permutation methods: a basis for exact inference. Stat Sci 2004;19(4):676–85.

Friston KJ, Holmes A, Poline JB, Price CJ, Frith CD. Detecting activations in pet and FMRI: levels of inference and power. NeuroImage 1996;4(3): 223–35.

Galán L, Biscay R, Rodríguez JL, Pérez-Abalo MC, Rodríguez R. Testing topographic differences between event-related brain potentials by using nonparametric combinations of permutation tests. Electroencephalogr Clin Neurophysiol 1997;102:240–7.

Guthrie D, Buchwald JS. Significance testing of difference potentials. Psychophysiology 1991;28:240–4.

Halgren E, Dhond RP, Christensen N, Petten CV, Marinkovic K, Lewine JD. N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. NeuroImage 2002;17:1101–16.

Hayasaka S, Nichols TE. Validating cluster size inference: random field and permutation methods. NeuroImage 2003;20:2343–56.

Hayasaka S, Nichols TE. Combining voxel intensity and cluster extent with permutation test framework. NeuroImage 2004;23:54–63.

Helenius P, Salmelin R, Service E, Connolly JF. Distinct time courses of word and context comprehension in the left temporal cortex. Brain 1998;121:1133–42.

Helenius P, Salmelin R, Service E, Connolly JF, Leinonen S, Lyytinen H. Cortical activation during spoken-word segmentation in nonreading-impaired and dyslexic adults. J Neurosci 2002;22:2936–44.

Holmes AP. Statistical issues in functional brain mapping. PhD thesis. University of Glasgow; 1994.

Holmes AP, Blair RC, Watson JDG, Ford I. Nonparametric analysis of statistic images from functional mapping experiments. J Cerebr Blood Flow Metabol 1996;16:7–22.

Jensen O, Weder N, Bastiaansen M, van den Brink D, Dijkstra T, Hagoort P. Modulation of the beta rhythm in a language comprehension task; submitted for publication.

Kaiser J, Hertrich I, Ackennann H, Lutzenberger W. Gamma-band activity over early sensory areas predicts detection of changes in audiovisual speech stimuli. NeuroImage 2006;30(4):1376–82.

Kaiser J, Lutzenberger W. Human gamma-band activity: a window to cognitive processing. NeuroReport 2005;16(3):207–11.

Kaiser J, Lutzenberger W, Preissl H, Mosshammer D, Birbaumer N. Statistical probability mapping reveals high-frequency magnetoencephalographic activity in supplementary motor area during self-paced finger movements. Neurosci Lett 2000;283(1):81–4.

Kaiser J, Ripper B, Birbaumer N, Lutzenberger W. Dynamics of gamma-band activity in human magnetoencephalogram during auditory pattern working memory. NeuroImage 2003;20(2):816–27.

Karnisky W, Blair RC, Snider AD. An exact statistical method for comparing topographic maps with any number of subjects and electrodes. Brain Topogr 1994;6:203–10.

Kutas M, Federmeier MD. Electrophysiology reveals semantic memory use in language comprehension. Trends Cognit Sci 2000;4:463–70.

Kutas M, Hillyard SA. Reading senseless sentences: brain potentials reflect semantic incongruity. Science 1980;207:203–5.

Lutzenberger W, Ripper B, Busse L, Birbaumer N, Kaiser J. Dynamics of gamma-band activity during an audiospatial working memory task in humans. J Neurosci 2002;22(13):5630–8.

Maris E. On the sampling interpretation of confidence intervals and hypothesis tests in the context of conditional maximum likelihood estimation. Psychometrika 1998;63(1):65–71.

Maris E. Randomization tests for ERP topographies and whole spatiotemporal data matrices. Psychophysiology 2004;41:142–51.

Maris E, Schoffelen JM, Fries P. Nonparametric statistical testing of coherence differences. J Neurosci Methods 2007;163:161–75.

Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 2002;15:1–25.

Pantazis D, Nichols TE, Baillet S, Leahy RM. A comparison of random field theory and permutation methods for the statistical analysis of MEG data. NeuroImage 2005;25:383–94.

Percival DB, Walden AT. Spectral analysis for physical applications. Cambridge: Cambridge University Press; 1993.

Pesarin F. Multivariate permutation tests. New York: Wiley; 2001.

Raz J, Zheng H, Ombao H, Turetsky B. Statistical tests for fMRI based on experimental randomization. NeuroImage 2003;19:226–32.

Simos PG, Basile LF, Papanicolaou AC. Source localization of the N400 response in a sentence-reading paradigm using evoked magnetic fields and magnetic resonance imaging. Brain Res 1997;762:29–39.

Singh KD, Barnes GR, Hillebrand A. Group imaging of taskrelated changes in cortical synchronization using nonparametric permutation testing. NeuroImage 2003;19:1589–601.