ELSEVIER

Journal of Neuroscience Methods 163 (2007) 161–175

JOURNAL OF NEUROSCIENCE METHODS

www.elsevier.com/locate/jneumeth

Nonparametric statistical testing of coherence differences☆

Eric Maris[a,*], Jan-Mathijs Schoffelen[b], Pascal Fries[c]

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

Received 12 October 2006; received in revised form 4 January 2007; accepted 7 February 2007

Abstract

Many important questions in neuroscience are about interactions between neurons or neuronal groups. These interactions are often quantified by coherence, which is a frequency-indexed measure that quantifies the extent to which two signals exhibit a consistent phase relation. In this paper, we consider the statistical testing of the difference between coherence values observed in two experimental conditions. We pay special attention to problems induced by (1) unequal sample sizes and (2) the fact that coherence is typically evaluated at a large number of frequency bins and between large numbers of pairs of neurons or neuronal groups (the *multiple comparisons problem*). We show that nonparametric statistical tests provide convincing and elegant solutions for both problems. We also show that these tests allow to incorporate biophysically motivated constraints in the test statistic, which may drastically increase the sensitivity of the test. Finally, we explain 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. The proposed methodology is illustrated by analyses of data collected in a study on cortico-spinal coherence [Schoffelen JM, Oostenveld R, Fries P. Neuronal coherence as a mechanism of effective corticospinal interaction. Science 2005;308(5718):111-3].

Keywords: Coherence; Nonparametric statistical testing; Multiple comparisons problem; Cortico-spinal coherence; Cortico-muscular coherence

---

1. Introduction

In neuroscience, one often encounters the problem of measuring the association strength between two signals. Very often, these signals are electrical recordings using electrodes (located on the scalp, on the dura, or in the grey matter) or magnetic recordings using SQUIDs. Because one wants to learn about interactions between brain areas (and sometimes, sensorimotor organs), it is of particular interest to understand the joint recordings from multiple sensors. A popular measure of association between signals is the coherence spectrum. Coherence is a frequency-indexed measure that quantifies the extent to which two signals exhibit a consistent phase difference. Coherence is often interpreted as a measure of effective interaction.

One problem that is still not fully resolved is the comparison of coherence values from two samples. Typically, these two samples are two sets of trials observed in different experimental conditions. Traditional statistical tests of coherence differences (Amjad et al., 1997; Brillinger, 1981; Enochson and Goodman, 1965) rely on the asymptotic normality of the Fourier transform, and this assumption can be questioned. Statistical testing of coherence difference is especially challenging with unequal sample sizes. This may happen, for instance, if the experimental conditions are defined by behavioral responses (e.g., correct/incorrect). Unequal sample sizes create a problem because coherence estimates are biased and the bias depends on the sample size: the smaller the sample size, the larger the bias. We need a statistical test that can cope with this differential bias.

Furthermore, in most neuroscience studies, coherence is evaluated in many frequency intervals (*bins*). This leads to a *multiple comparisons problem* (MCP): How can we perform a large number of statistical tests (one for every frequency bin) while at the same time controlling the false alarm rate for these statistical tests as a whole? An analogous problem is that, in most

---

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

* Corresponding author at: Nijmegen Institute of Cognition and Information (NICI), Radboud University, P.O. 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.2007.02.011

studies, much more than two signals are observed. As a consequence, the number of signal pairs can become very large, and so does the number of coherence values. This also leads to a MCP, but now over the spatial instead of the spectral dimension.

The objective of this paper is to show how nonparametric statistical tests offer a solution for these problems. First, a nonparametric statistical test of coherence differences does not depend on asymptotic normality. Second, it is valid for unequal sample sizes. And third, it deals with the MCP over the spectral as well as the spatial dimension. We will describe these statistical tests, apply them to an example data set on cortico-spinal interactions (Schoffelen et al., 2005), and present the theory that justifies their use.

The potential of nonparametric statistical tests for the analysis of EEG- and MEG-data has been noticed by several authors. These tests were first proposed for testing the difference between topographies at a particular time point (Achim, 2001; Galán et al., 1997; Karnisky et al., 1994) and later for whole spatiotemporal matrices (Maris, 2004). Nonparametric tests have also been used very successfully for frequency domain representations of EEG- and MEG-data (Kaiser et al., 2000, 2003, 2006; Kaiser and Lutzenberger, 2005; 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). Lee (2002) gives an interesting application of nonparametric tests to intracranial electrophysiological data observed in a single experimental condition. Finally, several authors have proposed nonparametric tests for the analysis of fMRI-data (Bullmore et al., 1996, 1999; Hayasaka and Nichols, 2003, 2004; Holmes et al., 1996; Nichols and Holmes, 2002; Raz et al., 2003).

The present 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 a Section 4 that explains why these nonparametric tests are formally correct. However, at no point in this paper, are concepts from mathematical statistics required.

2. Methods

2.1. The calculation of coherence

Although nonparametric statistical testing can also be applied to multiple-subject studies, for didactic reasons, we restricted our focus to single-subject studies (the subject can be an animal or a human). We consider the following case: a subject is observed in two experimental conditions and in every condition multiple trials are observed. The number of trials in the two conditions does not have to be equal. In every trial n, a bivariate time series (X_n, Y_n) is observed. These time series can be discretely sampled continuous processes (e.g., local field potentials, EEG, MR-signals) or point processes (e.g., spike trains). In the following, we assume the time series to be continuous processes. In the case of point processes and mixed continuous-point processes, coherence is calculated in a slightly different fashion; this is well described by Jarvis and Mitra (2001).

We now describe the calculation of coherence. This calculation involves trial-specific estimates of the power in the two signals (X and Y) and of their cross-spectrum: $s_n^{XX}(f)$, $s_n^{YY}(f)$, and $s_n^{XY}(f)$, with f denoting the frequency. These estimates can be calculated in several ways, and here we focus mainly on the so-called multitaper estimates (Percival and Walden, 1993). Multitaper estimation of power involves taking the average over a number of tapers (K) for a given signal X_n:

$$s_n^{XX}(f) = \frac{1}{K}\sum_{k=1}^{K}|F_f(t_k \otimes X_n)|^2,$$

in which $F_f(t_k \otimes X_n)$ denotes the Fourier transform of the tapered time series $t_k \otimes X_n$ (t_k is the k th taper) at frequency f. The power of the other signal (Y_n) is calculated in the same way. Multitaper estimation of the cross-spectrum also involves taking the average over a number of tapers (K):

$$s_n^{XY}(f) = \frac{1}{K}\sum_{k=1}^{K}F_f(t_k \otimes X_n)F_f(t_k \otimes Y_n)^*,$$

in which the asterisk (*) denotes complex conjugation.

Although we focus on multitaper estimation, the methods described in this paper also apply to other estimates of trial-specific power and the cross-spectrum. Two other estimates are described by Jarvis and Mitra (2001).

We introduce an independent variable I, whose length is equal to the total number of trials in the two conditions ($N_1 + N_2$, with N_1 and N_2 being the number of trials in the first and the second condition). The n th value of I, denoted by I_n, has the value 1 if the trial belongs to the first condition and it has the value 2 if it belongs to the second condition. The length of the time series, i.e. the number of time samples per trial, is equal for the two conditions.

The trial-specific estimates of power and the cross-spectrum are combined by averaging over the trials. These averages are calculated separately for each of the two conditions:

$$S_1^{XX}(f) = \frac{\sum_{n=1}^{N}\Im_1(I_n)s_n^{XX}(f)}{N_1},$$

$$S_2^{XX}(f) = \frac{\sum_{n=1}^{N}\Im_2(I_n)s_n^{XX}(f)}{N_2},$$

in which $\Im_a(x)$ is an indicator function that takes the value 1 if $x = a$ and 0 otherwise. Similar formulas hold for the power of the other signal ($S_1^{YY}(f)$ and $S_2^{YY}(f)$) and the cross-spectrum ($S_1^{XY}(f)$ and $S_2^{XY}(f)$).

Coherency is the normalized cross-spectrum, and it is calculated as follows:

$$C_1(f) = \frac{S_1^{XY}(f)}{\sqrt{S_1^{XX}(f)S_1^{YY}(f)}}.$$

A similar formula holds for $C_2(f)$, the coherency in the second condition. Coherency is a complex-valued measure whose magnitude ($|C_1(f)|$ and $|C_2(f)|$) quantifies the consistency of the phase differences between the two signals, and whose angle is equal to the average phase difference. Coherence is the magnitude of coherency. (Although it is not important for the present paper, it should be noted that coherency also depends on amplitude consistency; see Lachaux et al., 1999.)

2.2. *Coherence is biased*

The sample coherence is a biased estimate of the population coherence. This is because coherence is a positive quantity. To see this, assume that the population coherence is zero. In this case, the sample coherency has a uniform distribution in the complex plane, which means that it does not have a preferred angle. However, the magnitude of the sample coherency (i.e., the sample coherence) is always nonzero. The expected coherence depends on the variability of coherency: the more trials, the smaller the variability around zero in the complex plane, and the smaller its expected magnitude.

The coherence bias is illustrated in Fig. 1. In this figure, we show sample coherency values for two time series with a coherence of zero, separately for three different sample sizes ($N = 10$, $N = 40$, and $N = 100$). The cross-spectra were drawn from a uniform distribution in the complex plane. Every blue line in Fig. 1 represents the normalized cross-spectrum of a single trial. The normalization is performed by dividing the single-trial cross-spectrum by the square-root of the average power of each of the two signals, with the average being taken over the trials. Sample coherency is equal to the average (over the trials) normalized cross-spectrum. The sample coherency values are represented by the red lines. The important observation is that the sample coherence (i.e., the magnitude of the sample coherency) is a function of the sample size: the more trials, the smaller the sample coherence. This means that the coherence bias depends on the sample size.

2.3. *Parametric statistical tests for coherence differences*

Existing parametric statistical tests for coherence differences are based on asymptotic results (Amjad et al., 1997; Brillinger, 1981). More specifically, they are based on the asymptotic distribution of $\tanh^{-1}(|C(f)|)$, the inverse hyperbolic tangent of coherence. This asymptotic distribution depends on the number of degrees of freedom (d.f.) for the coherence estimates. For multitaper estimates, the d.f. for a coherence estimate is $2 \times N \times K$. For two conditions with equal d.f., Brillinger (1981) showed that, under the null hypothesis of equal population coherences, the difference $[\tanh^{-1}(|C_1(f)|) - \tanh^{-1}(|C_2(f)|)]$ asymptotically (for large d.f.) has a normal distribution with expected value 0 and variance 1/d.f. This result immediately leads to a statistical test for coherence differences. Amjad et al. (1997) extended this result to allow for a statistical test of the difference between three and more coherence values.

The statistical tests proposed by Brillinger (1981) and Amjad et al. (1997) can only be used if the conditions have equal d.f. For conditions with unequal d.f., we can make use of the following observation of Enochson and Goodman (1965): For d.f. > 20 and squared population coherence values $\gamma^2(f)$ between 0.4 and 0.95, $\tanh^{-1}(|C(f)|)$ is approximately normally distributed with mean $[\tanh^{-1}(|\gamma(f)|) + 1/(\text{d.f.} - 2)]$ (the term $1/(\text{d.f.} - 2)$ in this formula corrects for the bias) and variance $1/(\text{d.f.} - 2)$. It follows that, under the null hypothesis of equal population coherence values in the two conditions, the following test statistic (Z) is approximately normally distributed with expected value 0 and variance 1:

$$Z = \frac{(\tanh^{-1}(|C_1(f)|) - (1/\text{d.f.}_1 - 2)) - (\tanh^{-1}(|C_2(f)|) - (1/\text{d.f.}_2 - 2))}{\sqrt{(1/\text{d.f.}_1 - 2) + (1/\text{d.f.}_2 - 2)}}. \tag{1}$$

In this formula, d.f.$_1$ and d.f.$_2$ denote the degrees of freedom in, respectively, the first and the second condition. As far as we know, the properties of this test statistic have not been systematically evaluated. One reason of concern is that the squared population coherence values for which the observation of Enochson and Goodman (1965) is valid, are much larger than the squared sample coherence values that are typically observed in neuroscience experiments (Fries et al., 2001; Srinivasan et al., 1999; Tallon-Baudry et al., 2001).

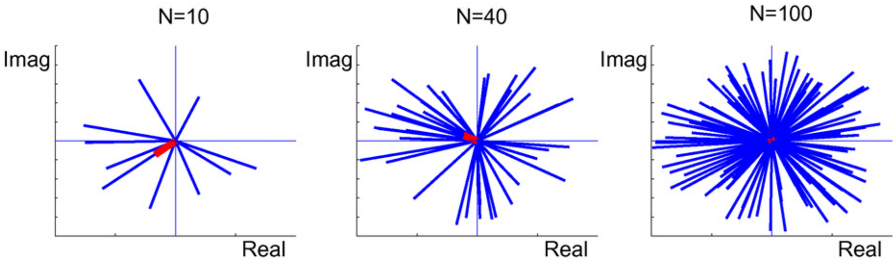

Fig. 1. Sample coherency values for three different sample sizes ($N = 10$, $N = 40$, and $N = 100$). Every blue line represents the normalized cross-spectrum of a single trial, and every red line represents a sample coherency. On the horizontal axes, we show the real part of the cross-spectra, and on the vertical axes, we show the imaginary part. The main observation is that the sample coherence (i.e., the magnitude of the sample coherency) is a function of the sample size: the more trials, the smaller the sample coherence. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Another reason of concern is that the d.f. of the multitaper estimates are only asymptotically valid. That is, the asymptotic distributions of $\tanh^{-1}(|C_1(f)|)$ and $\tanh^{-1}(|C_2(f)|)$ (normal with known means and variances) approximate the true sampling distributions only when the d.f. are large. Especially for sparse spike trains, the asymptotic distributions are likely to be poor approximations of the true sampling distributions. This problem was analyzed in detail by Jarvis and Mitra (2001). Under the assumption of a doubly stochastic inhomogeneous Poisson process, these authors showed that the asymptotic d.f. of the coherence estimates exceed the effective d.f. Moreover, the smaller the number of spikes, the larger the difference between the asymptotic and the effective d.f. The nonparametric methods proposed in this paper do not require parametric distributions (like the normal) and their parameters (mean and variance, which depend on the d.f.), and therefore they do not suffer from the problem mentioned here.

For conditions with unequal d.f., an obvious alternative to the coherence Z-test of Enochson and Goodman (1965) is to curtail one of the conditions to the same size as the other, and to use a statistical test based on Brillinger's (1981) result. However, this approach results in a loss of statistical power. If one uses the methods proposed in this paper, one does not have to tolerate a loss of statistical power.

2.4. A statistical test for coherence differences based on the jacknife

The issue of the appropriateness of the asymptotic d.f. for coherence estimates was taken up by Bokil et al. (2007). These authors proposed to use the Jacknife (Tukey, 1958) as a means to estimate the sampling variance of the Z-statistic in Eq. (1). The Jacknife is an obvious candidate here, because the regular formula for the sampling variance of the mean (the variance of the observations divided by N) cannot be used to estimate the sampling variance of the coherence estimates, nor of any function of these coherence estimates. Thus, instead of relying on the fact that the asymptotic variance of the Z-statistic is 1, these authors proposed to estimate it from the data. The Z-statistic is then divided by this Jacknife-estimated variance, and the resulting test statistic is assumed to be normally distributed with mean 0 and variance 1. In contrast to the approach in this paper, the approach of Bokil et al. (2007) is parametric. It is parametric because it assumes that the Z-statistic in Eq. (1) has a normal distribution with mean 0 and an unknown variance. However, it does not rely on the asymptotic sampling variance of the Z-statistic. The approach in this paper is fully nonparametric: no assumptions are made about the probability distributions of either the data or parameter estimates calculated on these data.

2.5. Type 1 and type 2 error rates

The quality of a statistical test is expressed in terms of its type 1 and type 2 error rate. The type 1 error rate is the probability of a false positive, rejecting the null hypothesis when it is true. And the type 2 error rate is the probability of a false negative, accepting the null hypothesis when the alternative hypothesis is true. The type 1 error rate is also called the *false alarm* (FA) rate of the test, and the complement of the type 2 error rate is also called the *sensitivity* of the test. In Section 4 of this paper, we deal with the FA rate: we show that nonparametric statistical tests of coherence differences effectively control the FA rate. Importantly, they control the FA rate for all sample sizes, including unequal and small ones, and for all spike rates.

Our treatment of sensitivity is much less complete than our treatment of the FA rate. In fact, our main point with respect to sensitivity is that it can be drastically increased by incorporating biophysically motivated constraints in the test statistic. Due to space limitations, no results will be presented that express the sensitivity of the test as a precise function of the factors on which it depends: true coherence, number of trials, spike rate, etc. Obtaining such results would require an extensive simulation study in which these factors are varied simultaneously. Without doubt, such a stimulation study would be very informative for neuroscientists interested in long-range coherence between spike trains. Especially for this type of coherence, it is important to know whether a null result can be due to the insensitivity of the statistical test.

2.6. Statistical testing of coherence differences in multiple frequency bins and multiple signal pairs

Very often, coherence is not calculated in a single frequency bin only. Instead, a coherence *spectrum* is calculated. A coherence spectrum is an array of frequency-indexed coherence values, sorted by frequency. We often do not know in which frequency bin an effect is likely to be observed. This makes the statistical analysis much more complicated than in the case of a single frequency bin. With multiple frequency bins, it is not sufficient to calculate multiple test statistics, one for every frequency bin, and their corresponding p-values. Due to the large number of statistical tests, it is not possible to control the *family-wise error rate* (FWER) by means of the standard statistical procedures that operate at the level of a single frequency bin. The FWER is the probability under the hypothesis of no effect of falsely concluding that there is a difference between the experimental conditions in one or more frequency bins. A solution of this multiple comparisons problem (MCP) requires a procedure that controls the FWER at some fixed level (typically, 0.05). In the following, whenever we use the term *false alarm* (FA) *rate* in the context of a statistical comparison involving multiple frequency bins, we mean the FWER.

The point to remember is the following: if the spectral locus of the effect is not known in advance, we need a specialized statistical procedure that takes our prior ignorance into account. It is very difficult to develop such a specialized procedure in the parametric statistical framework. The main obstacle is the unknown pattern of statistical dependence between the many frequency bins. Statistical dependence is not an important concern if the frequency bins are non-overlapping, because coherence estimates in non-overlapping frequency bins are statistically independent in large samples. In fact, Bokil et al. (2007) propose a way to deal with the MCP from exactly this perspective (statistically independent frequency bins). However, in many applications involving modern spectral analysis methods such as multitaper estimation and wavelets (Percival and Walden, 1993), the frequency bins do overlap, and therefore statistical dependence is an issue of concern.

We encounter a similar problem if more than two signals are observed and we do not know at which signal pair an effect is likely to be observed. With multiple signal pairs, it is not sufficient to calculate multiple test statistics, one for every signal pair, and their corresponding p-values. Instead, we need a specialized statistical procedure that takes our prior ignorance into account. Again, it is very difficult to develop such a specialized procedure in the parametric statistical framework. The main obstacle is the unknown pattern of statistical dependence between the spectral representations in the different signals.

We can now summarize the problems involved in the parametric statistical testing of coherence differences. First, all three statistical tests (Brillinger's, Amjad's, and the Z-statistic) depend on the asymptotic (complex-valued) normality of the Fourier transforms of the tapered time series. This property may not hold, and if it holds, it may be useless (if normality is only approximated in unrealistically large samples). As a consequence, the FA rate of these statistical tests may differ from their alpha-level. Second, for experimental conditions with unequal d.f., we have to rely on an observation by Enochson and Goodman (1965) of which we do not know whether it is a sufficient approximation of the true sampling distribution for all population coherence values between 0 and 1. Third, it is very difficult to develop a parametric statistical test of coherence differences in multiple frequency bins and multiple signal pairs. For all three problems, the nonparametric statistical framework provides a simple solution. This solution can compete with the approach of Bokil et al. (2007) because (1) it does not rely on the Jacknife estimate of the sampling variance (of which we know that it fails for some statistics; see Miller (1964)), and (2) it solves the MCP for all patterns of statistical dependence between different frequency bins and different signals.

2.7. *The nonparametric statistical test*

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 FA rate, and (3) the issue of how to choose a test statistic. These issues will be discussed in Section 4.

2.7.1. *A nonparametric statistical test for a single signal pair and a single frequency bin*

The first step in every nonparametric statistical test is the choice of a test statistic. Contrary to the parametric statistical framework, in the nonparametric framework, the scientist can use every test statistic that he believes to be sensitive to the effect of interest. If the interest is in the coherence difference for a single signal pair, then the obvious test statistic is $[|C_1(f)| - |C_2(f)|]$, the difference between the coherence values at one frequency in the two experimental conditions. Later, we will also consider a test statistic for the situation of multiple frequency bins and multiple signal pairs.

The nonparametric statistical test is performed in the following way:

1. (1) Collect the trials of the two experimental conditions in a single set.
2. (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. (3) Calculate the test statistic on this random partition.
4. (4) Repeat steps 2 and 3 a large number of times and construct a histogram of the test statistics.
5. (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 *Monte Carlo p-value*.
6. (6) If the Monte Carlo 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 using this procedure, is equal to the critical alpha-level.

Nonparametric statistical testing is extremely general because its validity 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. By using a different test statistic, one obtains a statistical test that is sensitive to other aspects of the data. For instance, the effect of the independent variable could (also) be reflected in the phase difference between the two signals. To capture this effect, one can use as a test statistic $[\Phi_1(f) - \Phi_2(f)]$, in which $\Phi_1(f)$ and $\Phi_2(f)$ are the phases in the two experimental conditions.

2.7.2. *A nonparametric statistical test for multiple frequency bins*

Very often, instead of calculating coherence in a single frequency bin, a coherence spectrum is calculated. Because the effect may be observed in any one of multiple frequency bins, we have to deal with the MCP. The nonparametric statistical test provides a solution for the MCP because it can be modified such that it calculates a single test statistic for the whole coherence spectrum. Several test statistics can be used for this purpose, and here we consider two: (1) the maximum (over the frequency bins) of the coherence Z-statistics, defined in Eq. (1), and (2) a test statistic that is based on clustering of adjacent frequency bins, which will be described after the maximum coherence Z-statistic. It is important to emphasize that, although we previously criticized the coherence Z-statistic, we have no problem using it as an element of a nonparametric statistical test. This is because the incorrectness of the assumed sampling distribution of the coherence Z-statistic (a normal distribution with expected value 0 and variance 1) does not affect the FA rate of the nonparametric test. This is because the nonparametric test controls the FA rate for all test statistics, regardless of their sampling distribution. For instance, we would also control the FA rate if, instead of the coherence Z-statistic, the plain coherence difference were used. The advantage of the Z-statistic is that it can more easily be used for thresholding, which is a step in the clustering procedure that will be described in the following. We will return to this point in Section 4.

The nonparametric statistical test that uses the maximum coherence Z-statistic is performed using the same recipe as the one in Section 2.7.1. The only difference in the actual computation is that, in step 3, the maximum coherence Z-statistic is calculated instead of the coherence Z-statistic for a single frequency bin. As an aside, it must be noted that this is a one-sided test; we obtain a two-sided test if the maximum is taken of the absolute values of the coherence Z-statistics.

A nonparametric statistical test controls the FA rate for all test statistics. We can take advantage of this fact by using a test statistic that is maximally sensitive to effects that are likely to occur. For instance, assuming that a hypothesized effect is broad-band, it is likely that adjacent frequency bins exhibit the same effect. To capture this phenomenon, we introduce a test statistic that is based on clustering of adjacent frequency bins. Instead of the maximum coherence Z-statistic, in step 3 of the recipe, a cluster-based test statistic is calculated. The calculation of this cluster-based test statistic involves several steps.

(1) For every frequency bin, calculate the coherence for each of the two experimental conditions.
(2) For every frequency bin, evaluate the coherence difference by means of a test statistic, such as the coherence Z-statistic in Eq. (1).
(3) Select all frequency bins whose coherence Z-statistic is larger than some threshold. For instance, this threshold can be some quantile of the normal distribution with expected value 0 and variance 1. (The incorrectness of the assumed sampling distribution of the coherence Z-statistic does not affect the FA rate of the nonparametric test.)
(4) Cluster the selected frequency bins in connected sets on the basis of adjacency; neighboring frequency bins are clustered in the same set.
(5) Calculate cluster-level statistics by taking the sum of the coherence Z-statistics within a cluster.
(6) Take the maximum of the cluster-level statistics.

This is a test statistic for a one-sided test; for a two-sided test, in step 3, we select test statistics whose absolute value is larger than some threshold, and in step 6, we take the cluster-level statistic that is largest in absolute value. Also, for a two-sided test, the clustering in step 5 is performed separately for frequency bins with a positive and a negative Z-statistic.

The cluster-based test statistic depends on the threshold that is used to select frequency bins for clustering. In our example, this threshold was the 95th quantile of the normal distribution with expected value 0 and variance 1. Although this threshold does not affect the FA rate of the statistical test, as will be shown in Section 4, it does affect the sensitivity of the test. For example, weak but widespread effects are not detected when the threshold is high.

Except for the fact that the test statistic is rather complicated, the nonparametric statistical test is performed in the same way as for a single signal pair: a Monte Carlo p-value is calculated by randomly partitioning the trials and if this p-value is less than the critical alpha-level, then conclude that the data in the two conditions are significantly different. If more than one cluster of frequency bins is identified, the p-values for all clusters are calculated under the histogram of the maximum cluster-level statistic and not under the histogram of the second largest, third largest, etc. The choice for the maximum cluster-level statistic (and not the second largest, third largest, ...) 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).

2.7.3. A nonparametric statistical test for multiple signal pairs

We now consider coherence in a single frequency bin for multiple sensor pairs. Here, the MCP can be solved in essentially the same way as for a single sensor pair and multiple frequency bins: instead of clustering neighboring frequency bins, we now cluster adjacent sensor pairs. To explain the clustering of sensor pairs, we must first describe the configuration of these sensor pairs. In this paper, we consider a set of signal pairs that involve one common sensor, which will be called the *reference sensor*. (The reason for this is in the example study that we use to illustrate the methodology. The example study will be described later.) Denoting the reference sensor by R and the target sensors by $M1$, $M2$, $M3$, ..., the set of signal pairs is the following: $(M1, R)$, $(M2, R)$, $(M3, R)$, ... In the example study, the non-reference sensors $M1$, $M2$, $M3$, ..., are the sensors of a magneto-encephalogram (MEG), and in the following we will denote them as *MEG sensors*.

To solve the MCP, we use a test statistic that is based on clustering of adjacent MEG sensors for which the coherence with the reference sensor exhibits a similar difference (in sign and magnitude). The calculation of this test statistic involves the following steps:

(1) For every MEG sensor, calculate the coherence with the reference sensor for each of the two experimental conditions.
(2) For every MEG sensor, evaluate the coherence difference by means of a test statistic, such as the coherence Z-statistic in Eq. (1).
(3) Select all samples whose Z-statistic is larger than some threshold.
(4) Cluster the selected samples in connected sets on the basis of spatial adjacency.
(5) Calculate cluster-level statistics by taking the sum of the Z-statistics within a cluster.
(6) Take the maximum of the cluster-level statistics.

So far, we have described approaches to treat the MCP for a single sensor pair and multiple frequency bins and also for multiple sensor pairs and a single frequency bin. The same rationale can be generalized to multiple sensor pairs and multiple frequency bins. The only difference is in the test statistic: in the latter case, cluster combinations of sensor pairs and frequency bins on the basis of spectral *and* spatial adjacency.

2.7.4. The reliability of the Monte Carlo p-value

Strictly speaking, the nonparametric statistical test is only valid if the Monte Carlo *p*-value is calculated on the basis of all possible partitions of the trials in two subsets (with each partition having the same probability), and not just some random subset of the collection of possible partitions. In fact, in the case of a complete enumeration of all possible partitions, the Monte Carlo *p*-value is the true nonparametric *p*-value, of which we will show that it controls the FA rate under some well-specified null hypothesis (see Section 4). However, the number of trials may be so large that it is infeasible to calculate the test statistic for all possible partitions. For example, with 200 trials, evenly distributed over the two conditions, the number of possible partitions is approximately 1.0e + 29. In this case, one can only calculate a Monte Carlo *p*-value on the basis of a random subset of all possible partitions.

To determine the necessary number of random partitions, it is useful to construct a confidence interval for the Monte Carlo *p*-value. 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. It makes sense to determine the number of draws in an adaptive way: increase this number until the confidence interval does not contain anymore the critical alpha level (0.01 or 0.05). With this strategy, the necessary number of random partitions is large if the Monte Carlo *p*-value is on the boundary of significance. In other words, when the difference is highly significant or far from significance, then this can be found with few random partitions. But when the true nonparametric *p*-value is just below or above the critical alpha level, then many random partitions are needed to attain certainty about the significance.

2.8. Example study: cognitive modulation of cortico-spinal coherence

We analyzed data of a study by Schoffelen et al. (2005) on cortico-spinal coherence. In this study, the interest was in the coherence between the electromyogram (EMG) over a muscle of the right forearm, which informs us about alpha motor neuron activity in the spinal cord (the reference sensor), and the MEG (especially the sensors over the contralateral motor cortex). The signals were observed in two experimental conditions that differed with respect to the subject's expectation of when a go-signal would occur. All responses were given with the right hand. The signals were taken from periods in which the right wrist was extended and the subjects held this wrist extension until a go-cue was given. This go-cue was a sudden change in the speed of moving concentric circles that were displayed on a screen in front of the subject. The two conditions differed with respect to when the response signal was most likely to occur. In the so-called UP condition, the probability that the response signal would occur in the time interval $[t, t + \Delta t]$, given that it had not yet occurred before time t, increased with t, and in the so-called DOWN condition it decreased with t. This instantaneous conditional probability is called the *hazard rate*. Subjects were trained beforehand on one or the other hazard rate and implicitly learned that hazard rate. This can be concluded from the fact that they modulated their reaction times accordingly while reporting to not being aware of it.

We analyzed the signal in the time interval between 250 and 850 ms after the onset of the visual stimulus. In this time interval, the hazard rate for the response signal was about three times higher in the DOWN condition than in the UP condition. Because Schoffelen et al. (2005) hypothesized that cortico-spinal coherence increases with the subject's readiness to respond, they expected a larger coherence in the DOWN than in the UP condition. These coherence values are denoted by, respectively, $C_{\text{DOWN}}(f)$ and $C_{\text{UP}}(f)$]. It is important to observe that the physical stimulation in the two experimental conditions is identical; only the subject's readiness to respond differs.

The statistical analysis of this data set is challenging for several reasons. First, there is a large difference in the number of trials in the two experimental conditions. This is because the go-signal (and thereby the trial end) occurred on average earlier in the DOWN condition than in the UP condition. For the parametric coherence Z-test this may be problematic, because it relies on the bias correction factors in the numerator of Eq. (1). Moreover, the sample coherence values are much smaller (approximately 0.2 over left motor cortex) than the minimum population coherence value for which Enochson and Goodman (1965) advocate the use of their approximation (i.e., $\sqrt{0.4} = 0.63$)).

Second, we would like to investigate the pattern of coherence without making strong prior assumptions as to which frequency bins and which MEG sensors to involve. Therefore, we need a statistical procedure that controls the FA rate over multiple frequency bins and multiple signal pairs.

3. Results

In this section, we will focus on three points: (1) we give the results of a nonparametric test for a single EMG-MEG signal pair and a single frequency bin (Section 3.1), (2) we show that, contrary to the parametric coherence Z-test, the nonparametric test controls the FA (Section 3.2), and (3) we give the results of the cluster-based nonparametric tests for multiple EMG-MEG signal pairs and multiple frequency bins (Sections 3.3, 3.4, and 3.5).

3.1. A single EMG-MEG signal pair and a single frequency bin

We investigated cortico-spinal coherence in the gamma frequency band, more specifically, in the range between 40 and

60 Hz. In this frequency band, the sensors over the left motor cortex (contralateral to the response) showed an increase in coherence when the stimulus appeared on the screen. We now want to know whether this gamma band coherence is modulated by the readiness to respond (i.e., the difference between the UP and the DOWN condition).

We selected the MEG sensor that had the largest gamma band coherence with the EMG signal, averaged over the two conditions. We used the multitaper method (Percival and Walden, 1993) to calculate coherence in the frequency bin from 40 to 60 Hz; this involved spectral smoothing over this interval. For the selected sensor over the left hemisphere, the nonparametric test of the coherence difference between the UP and the DOWN condition was significant: from the 10,000 random partitions, none resulted in a coherence difference that was larger than the observed coherence difference, and thus the Monte Carlo p-value equals 0. Thus, we can conclude that the readiness to respond increases the cortico-spinal coherence.

We also performed a statistical test of the difference between the UP and the DOWN condition with respect to the between-signal phase difference. The Monte Carlo p-value of this nonparametric test is 0.0557 and its 95% confidence interval is [0.0512, 0.0602]. Thus, there is no significant modulation of the between-signal phase difference by the readiness to respond.

3.2. An evaluation of the FA rate of the parametric and the nonparametric statistical test

Normally, we apply a statistical test to determine whether two conditions differ significantly. Any such test, parametric or nonparametric, will have an FA rate. We can determine this FA rate with a procedure that makes use of the trials of a single experimental condition. These trials come from some probability distribution f. It is possible to construct two samples from probability distribution f by randomly partitioning the trials in two samples. Building on this fact, we can evaluate the FA rate of a statistical test by means of the following procedure:

(1) Randomly partition the trials of one sample into two samples with fixed sizes.
(2) Perform a statistical test of the difference between the two samples. (If this statistical test is nonparametric, it may involve the calculation of a Monte Carlo p-value, which again requires random partitioning. It is important to distinguish between the first and the second random partition: the first is the mechanism that produces the two samples that will be compared, and the second is a part of the calculation of the Monte Carlo p-value.)
(3) Repeat steps 1 and 2 a large number of times and calculate the proportion of repetitions in which the statistical test is significant.

By construction, the trials in the two samples that will be compared come from the same probability distribution. Therefore, the proportion of repetitions in which the statistical test is significant (determined in step 3), is the FA rate of the statistical test. For a statistical test to be sound, this FA rate must be equal to its critical alpha-level (which is used in step 2).

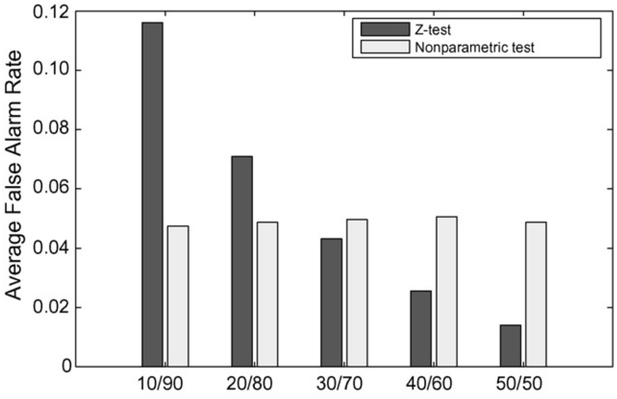

Fig. 2. Average (over the EMG-MEG sensor pairs) FA rates for the five sample size schemes. The parametric coherence Z-test is too liberal when the sample sizes are very unequal (10/90 and 20/80) and too conservative when the sample sizes are nearly equal (40/60 and 50/50). For the nonparametric test, the average FA rates are approximately equal to the critical alpha-level (0.05). This holds for all sample size schemes.

Because we are interested in the effect of unequal sample sizes on the FA rate, we constructed samples according to the following five schemes: 10/90 (10% of the trials in the first sample and 90% in the second), 20/90, 30/70, 40/60, and 50/50 (equal sample sizes). We used 196 trials of the UP condition. For each of the five schemes, we performed 1000 random partitions of these 196 trials. For every random partition, the resulting samples were compared by means of the parametric coherence Z-test and the nonparametric test of the previous section (Section 3.1). Both statistical tests were two-sided and used a critical alpha-level of 0.05. The statistical tests were performed for all 151 MEG sensors. The Monte Carlo p-value for the nonparametric test was calculated using 100 random partitions. This low number was chosen to show that the FA rate of the nonparametric test does not depend on the accuracy of the Monte Carlo p-value. Higher numbers of random partitions would increase the accuracy of the Monte Carlo p-value.

The results are shown in Fig. 2. The parametric coherence Z-test is too liberal when the sample sizes are very unequal (10/90 and 20/80) and too conservative when the sample sizes are nearly equal (40/60 and 50/50). This is confirmed by the results of two-sided binomial tests (one for every EMG-MEG sensor pair) of the null hypothesis that the FA rate is equal to the critical alpha-level (0.05). These results are shown in Fig. 3.

For the nonparametric test, the average FA rates are approximately equal to the critical alpha-level (0.05). This holds for all sample size schemes. This conclusion is confirmed by the results of the two-sided binomial tests (one for every EMG-MEG sensor pair) of the null hypothesis that the FA rate is equal to the critical alpha-level: the proportion of statistical tests that were either too liberal or too conservative is approximately 0.05 for all sample size schemes (results not shown).

3.3. Multiple frequency bins

We performed the statistical test for multiple frequency bins twice: once for a MEG-sensor over the contralateral (left) motor

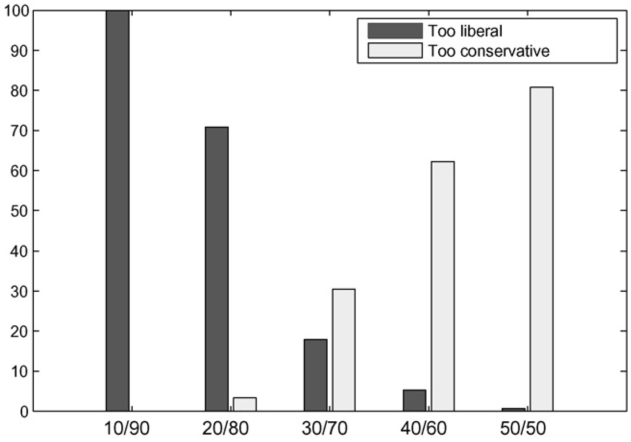

Fig. 3. Percentage of EMG-MEG sensor pairs for which the parametric coherence Z-test is too liberal (an FA rate that is significantly larger than the critical alpha-level of 0.05) and the percentage for which it is too conservative (an FA rate that is significantly smaller than the critical alpha-level).

cortex, and once for a MEG-sensor over the ipsilateral (right) motor cortex. We used the multitaper method (Percival and Walden, 1993) to calculate coherence in 46 frequency bins with center frequencies between 10 and 100 Hz. The frequency bins were chosen such that the largest frequency in a bin was 1.5 times the smallest frequency. For instance, for the frequency bin with center frequency 30 the smallest frequency was 24 and the largest frequency was 36. Thus, the frequency bins had variable widths.

The results for the MEG-sensor over the contralateral motor cortex are shown in Fig. 4. There are four clusters of connected frequency bins that exceed the threshold for the coherence Z-statistic, which was set at 1.96 (chosen a priori). The two largest clusters, one in the gamma band (36–70 Hz) and one in the beta band (21–34 Hz), are significant: for the largest cluster-level statistic, the Monte Carlo p-value is 0, and for the second largest it is 0.011 (calculated on 10,000 random partitions). In panel b of Fig. 4, the two significant clusters are indicated by the shaded area under the curve. The two non-significant clusters are between 80 and 95 Hz. The finding of two significant clusters confirms our previous conclusion on the basis of a single frequency bin: coherence is modulated by the readiness to respond. It is clear that in this subject the modulation is not only observed in the gamma band (the largest cluster), but also in the beta band (the second-largest cluster).

The results for the MEG-sensor over the ipsilateral motor cortex are shown in Fig. 5. There are four clusters of connected frequency bins that exceed the threshold for the coherence Z-statistic. Only the largest cluster-level statistic, the one in the gamma band, is significant: its Monte Carlo p-value is 0.001 (calculated on 10,000 random partitions). This cluster is indicated by the shaded area in panel b of Fig. 5. Despite the large coherence difference in the beta band, the corresponding cluster-level statistic is not significant (Monte Carlo p-value equal to 0.0769). The two smallest non-significant clusters are between 68 and 72 Hz, and they consist of one or two frequency bins only. On the basis of these findings, we conclude that in this subject the gamma coherence over the ipsilateral motor cortex is also modulated by the readiness to respond. This modulation may reflect effective connectivity between the two motor cortices or spatial non-specificity of the cognitive modulation.

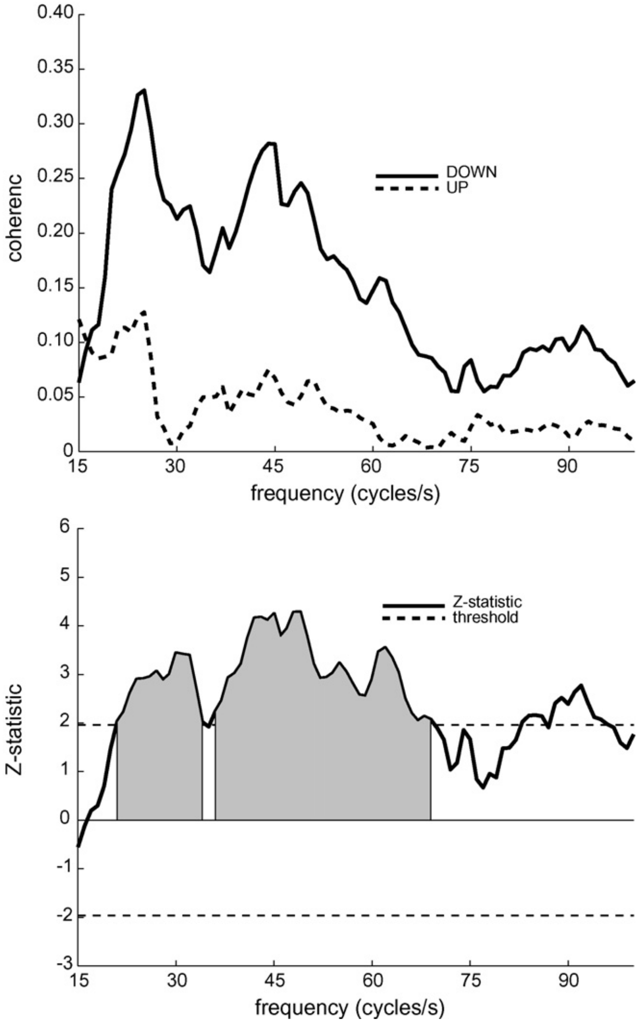

Fig. 4. Nonparametric statistical testing of coherence in multiple frequency bins for a sensor over the contralateral motor cortex. In panel a, the coherence spectra are shown, separately for the UP and the DOWN condition. In panel b, the solid line is the time series of sample-specific coherence Z-statistics, and the dashed line is the threshold that is used for selecting frequency bins that can subsequently be clustered. The shaded areas indicate the two significant clusters.

3.4. Multiple EMG-MEG signal pairs

The statistical test for multiple EMG-MEG signal pairs was applied to the frequency bin from 40 to 60 Hz. In the previous analyses, we observed an effect in this frequency bin, both over the contra- and the ipsilateral motor cortex. This finding was replicated in the present analysis that uses the data in all EMG-MEG signal pairs: in this subject, there are two spatially connected clusters in the data, one over the contra- and one over the ipsilateral motor cortex. Both clusters are significant: the largest cluster-level statistic has a Monte Carlo p-value of 0.009 and the second-largest has a Monte Carlo p-value of 0.0092 (calculated on 10,000 random partitions). This confirms our previous conclusion on the basis of single EMG-MEG sig-

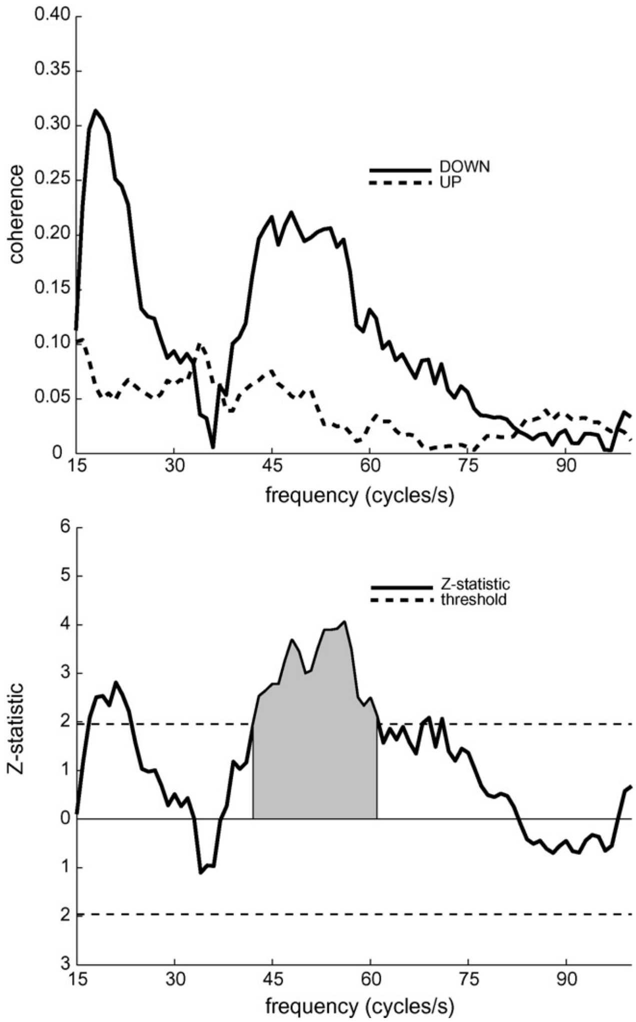

Fig. 5. Nonparametric statistical testing of coherence in multiple frequency bins for a sensor over the ipsilateral motor cortex. In panel a, the coherence spectra are shown, separately for the UP and the DOWN condition. In panel b, the solid line is the time series of the sample-specific coherence Z-statistics, and the dashed line is the threshold that is used for selecting frequency bins that can subsequently be clustered. The shaded area indicates the significant cluster.

nal pairs: gamma coherence over the ipsi- and the contralateral motor cortex is modulated by the readiness to respond.

In Fig. 6, we show the topography of the raw effect (i.e., the difference in cortico-spinal coherence between the DOWN and the UP condition). This topography is obtained by masking the raw effect by the spatial pattern of the significant cluster. This masking involves that the raw effects of all signal pairs that do not belong to the significant cluster are set equal to zero.

3.5. *Multiple EMG-MEG signal pairs and multiple frequency bins*

Finally, we also investigated the modulation of cortico-spinal coherence over all EMG-MEG signal pairs and over a large number of frequency bins. We calculated coherence in 17 frequency bins with center frequencies between 15 and 100 Hz. As before, the frequency bins were chosen such that the largest frequency in a bin was 1.5 times the smallest frequency. There are 14 clusters of connected (sensor,frequency)-pairs that exceed the threshold for the coherence Z-statistic. Four of theses clusters are significant: (1) one cluster in the beta band over the contralateral motor cortex ($p = 0.001$), (2) one cluster in the beta band over the ipsilateral motor cortex ($p = 0.007$), (3) one cluster in the gamma band over the contralateral motor cortex ($p = 0.008$), and (4) one cluster in the gamma band over the ipsilateral motor cortex ($p = 0.016$). These four clusters are shown in Fig. 7.

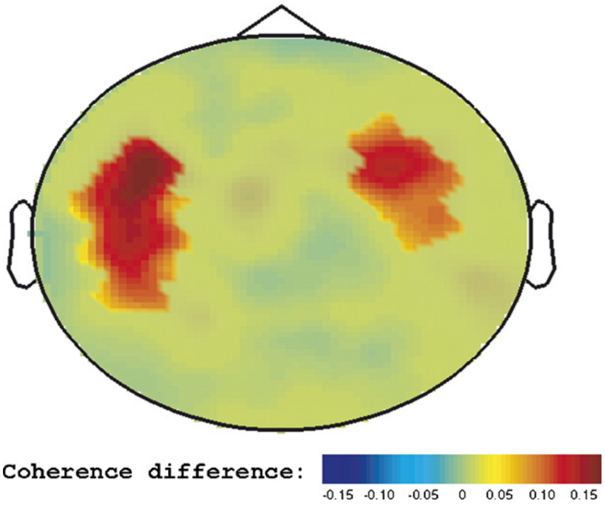

Fig. 6. Topography of the raw effect (the difference in cortico-spinal coherence between the DOWN and the UP condition) for the frequency bin [40, 60 Hz], masked by the spatial pattern of the significant cluster.

This finding confirms our previous conclusions: readiness to respond modulates coherence over the ipsi- and the contralateral motor cortex, and this modulation is observed in both the beta and the gamma band. This modulation is not a broadband phenomenon: in the coherence spectrum, there are two clear peaks, one over the beta and one over the gamma band. This is clear from the coherence spectra in Figs. 4 and 5, and from the fact that there are no significant clusters in the frequency bin [28, 42 Hz] (with center frequency 35 Hz), which is shown in Fig. 7. Note that, in the analysis of the single EMG-MEG signal pair over the ipsilateral motor cortex (see Fig. 5), there was no significant modulation in the beta band. In the last analysis, which involves all EMG-MEG signal pairs, the beta band modulation over the ipsilateral motor cortex *is* significant. This is because the cluster-based statistical test takes advantage of the fact that the beta band modulation is observed in multiple adjacent MEG sensors.

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. For the first two points, we can make use of the theory of nonparametric statistical tests. As compared to parametric statistics, the theoretical framework behind nonparametric statistical tests is not well documented and not very accessible. The central argument of this section (the so-called *conditioning*

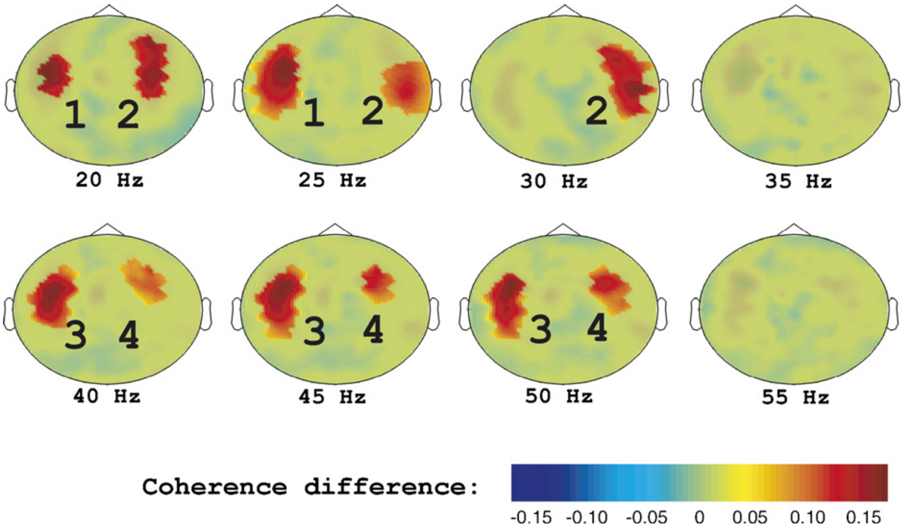

Fig. 7. Topography of the raw effect (the difference in cortico-spinal coherence between the DOWN and the UP condition) for eight frequency bins, masked by the spatio-spectral pattern of the four significant clusters. The four clusters are denoted by the numbers 1, 2, 3 and 4. The frequencies below the heads are the center frequencies of the frequency bins.

rationale, see further) is formally identical to an argument in the introductory chapter of a book by Pesarin (2001). The argument 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. Because this argument is extremely powerful, we have tried to make it accessible to the neuroscience community. When doing this, we needed some definitions, and these are introduced now.

4.1. *The structure in the data*

In this paper, we only consider single-subject studies. In this type of 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 signal that is observed in the trials. For completeness, it must be noted that 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 paper, we only consider the between-trials design.

To describe the structure in the data, we make the usual distinction between a dependent and an independent variable. In the example, the dependent variable is the complete set of (EMG,MEG)-signal pairs in the two conditions. 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 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. In our example study, the spatial dimension of the spatiotemporal matrix is formed by a single EMG channel and the 151 MEG channels.

The independent variable specifies the different experimental conditions. In the example, there are two experimental conditions: the upward going and the downward going hazard rate. 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 UP condition and 2 if it belongs to the DOWN 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*

The null hypothesis of a permutation test is about the probability distributions of the trial-specific data structures D_r, which 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:

$$f(D_1) = f(D_2) = \quad \ldots \quad = f(D_n). \tag{2}$$

In 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 null hypothesis of many familiar parametric statistical tests (i.e., the *t*-, the *F*-test, and their multivariate generalizations) also involves that all trial-specific data structures D_r are drawn from the same probability distribution, regardless of the experimental condition. This may sound unfamiliar, because in statistics handbooks the parametric null hypothesis is formulated as equality of the two conditions with respect some *parameter* of the probability distribution (typically, the expected value, but also the variance, the covariance, ...). 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.

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 signal in one trial depends on the signal in another trial. A biologically plausible form of statistical dependence is temporal autocorrelation: correlation between the signals 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. (2).

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 probability distributions. However, this is just a matter of presentation: under the assumption of statistical independence, the null hypothesis of identical probability 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 coherence difference $[|C_{\text{DOWN}}(f)| - |C_{\text{UP}}(f)|]$ is such a test statistic. To show this, assume that the first two trials belong to the DOWN condition, and the last two belong to the UP condition. Now, the coherence difference 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 coherence values for the trial pairs (d_1, d_3) and (d_2, d_4) are independent of the order of the trials within the pairs. 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. 8, 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 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.

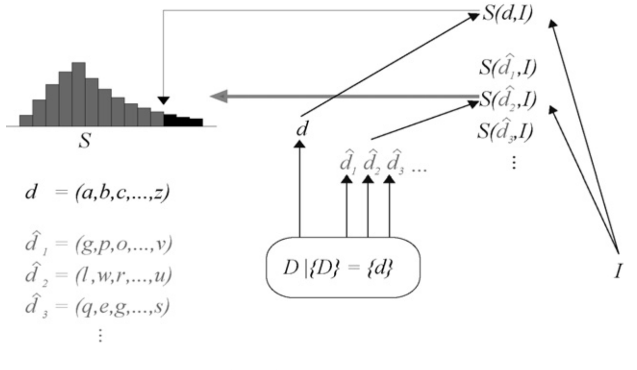
S(d,I)
$S(\hat{d}_1,I)$
$S(\hat{d}_2,I)$
$S(\hat{d}_3,I)$
⋮
S
d
$\hat{d}_1\ \hat{d}_2\ \hat{d}_3$ ...
$d = (a,b,c,\ldots,z)$
$\hat{d}_1 = (g,p,o,\ldots,v)$
$\hat{d}_2 = (l,w,r,\ldots,u)$
$\hat{d}_3 = (q,e,g,\ldots,s)$
⋮
$D|\{D\}=\{d\}$
I

Fig. 8. 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., $\hat{d}_1, \hat{d}_2, \hat{d}_3, \ldots$) are printed grey. The observed test statistic is denoted by $S(d, I)$, and the draws from the permutation distribution by $S(\hat{d}_1, I)$, $S(\hat{d}_2, I)$, $S(\hat{d}_3, I), \ldots$. 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.

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 false alarm 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 α to denote the critical alpha-level.

$$\begin{aligned} P(\text{Reject}H_0|H_0) &= \sum_{\{d\}} P(\text{Reject}H_0|\{D\}=\{d\}, H_0) f(\{D\}=\{d\}) \\ &= \sum_{\{d\}} \alpha f(\{D\}=\{d\}) \\ &= \alpha \end{aligned} \tag{3}$$

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 marginal distribution is controlled. This general fact will be called the *conditioning rationale*.

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 practical applications, 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

FA rate control by means of a permutation test does not depend on the test statistic. 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 contrast, in nonparametric statistics, one is free to choose any test statistic one likes. This freedom has several advantages and here we briefly discuss three of these advantages:

(1) It provides a simple way to solve the MCP: instead of evaluating the difference between the experimental conditions for each of the sensors separately, it is now evaluated by means of a single test statistic for the complete sensor array, for example the maximal value of the test statistic across the sensor array. Thus, the multiple comparisons (one for every sensor) are replaced by a single comparison, and therefore the MCP does not exist any more.

(2) It allows us to incorporate prior knowledge about the type of effect that can be expected. Incorporating prior knowledge will increase the sensitivity of the test. There are many biologically motivated constraints that can be incorporated. For example, when comparing sensor array data in two experimental conditions, one can make use of the fact that adjacent sensors 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 sensors, such as the size of the largest connected cluster that exceeds some threshold, or the sum of the coherence Z-statistics in that cluster.

To illustrate the differential sensitivity of different test statistics, we reanalyzed the example data set with a test statistic that is not based on clustering of adjacent sensors: the maximum of the sensor-specific coherence Z-statistics. We reanalyzed the multiple EMG-MEG signal pairs at 17 frequency bins with center frequencies between 15 and 100 Hz. On the basis of the critical value of the maximum coherence Z-statistic, only four (sensor,frequency)-pairs were significant. This contrasts with the cluster-based test statistic, that detected 211 significant (sensors,frequency)-pairs, distributed over four clusters.

(3) It is possible to localize the area of the largest effect by making use of the maximum-statistic. When comparing sensor array data in two conditions, one is almost always interested in the spatial and/or spectral localization of the effect. The maximum-statistic allows to bridge the gap between the interest in localized effects and the requirement to control the FA rate.

For localization, we need a method to identify significant parts of the spatio-spectral data structures. This is possible by means of a critical value that is applied at the level of the (sensor, frequency)-pairs or at the level of clusters. For every cluster, a cluster-level statistic is calculated by taking the sum of the t-statistics within the cluster. To control the FA rate of the localization procedure, we need a critical value for the cluster-level statistics with the following property: Under the null hypothesis, the probability that one or more cluster-level statistics exceed the critical value CV, is controlled at some critical alpha-level. Formally,

$$P(\text{at least one cluster-level statistic} \geq \text{CV}) = \alpha.$$

This is equivalent to

$$P(\text{Max(cluster-level statistics)} \geq CV) = \alpha.$$

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

4.5. Conclusions

We have shown that nonparametric statistical testing of coherence differences is a viable alternative to its parametric counterpart. First, the FA rate of nonparametric statistical tests does not depend on auxiliary assumptions about the probability distribution of the Fourier transforms. Second, nonparametric statistical testing offers complete freedom with respect to the choice of a test statistic. This property (1) provides a simple way to solve the MCP, (2) allows us to incorporate prior knowledge about the type of effect that can be expected, and (3) allows us to localize the effect.

We have presented a theory for these nonparametric statistical tests, which demonstrates their validity in a rigorous way. The null hypothesis of these statistical tests involves that the probability distributions of the signal in the different experimental conditions are equal. Under the assumption of statistical independence between the trials, this null hypothesis is equivalent with exchangeability of the dependent variable. Exchangeability is an intermediate concept that allows us to demonstrate the validity of the permutation test.

In this paper, we have only considered single-subject between-trial studies. However, the core of the theory is also applicable to single-subject *within*-trial studies and to *multiple*-subject studies (both between-subjects and within-subjects). In multiple-subject studies, we have to deal with the question whether the effect in the sample can be generalized to a population. This involves a so-called random-effect null hypothesis. Nonparametric testing of random-effect null hypotheses requires a separate paper.

References

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

Amjad AM, Halliday DM, Rosenberg JR, Conway BA. An extended difference of coherence test for comparing and combining several independent coherence estimates: theory and application to the study of motor units and physiological tremor. J Neurosci Meth 1997;73(1):69–79.

Bokil H, Purpura K, Schoffelen JM, Thomson D, Mitra P. Comparing spectra and coherences for groups of unequal size. J Neurosci Methods 2007;159:337–45.

Brillinger DR. Time series—data analysis and theory. San Francisco: Holden Day; 1981.

Bullmore E, Brammer M, Williams SCR, Rabe-Hesketh S, Janot N, David A, et al. 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.

Enochson LD, Goodman NR. Gaussian approximation for the distribution of sample coherence, Tech. Rep. TR 65–57. Ohio: Air Force Flight Dynamics Laboratory, Wright-Patterson AFB; 1965.

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

Fries P, Neuenschwander S, Engel AK. Rapid feature selective neuronal synchronization through correlated latency shifting. Nat Neurosci 2001;4(2):194–200.

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.

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.

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.

Jarvis MR, Mitra PP. Sampling properties of the spectrum and coherency of sequences of action potentials. Neur Comput 2001;13(4):717–49.

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.

Lachaux JP, Rodriguez E, Martinerie J, Varela FJ. Measuring phase synchrony in brain signals. Hum Brain Mapp 1999;8:194–208.

Lee D. Analysis of phase-locked oscillations in multi-channel single-unit spike activity with wavelet cross-spectrum. J Neurosci Meth 2002;115(1): 67–75.

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.

Miller RG. Trustworthy jacknife. Ann Math Stat 1964;35(4):1594.

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 randomisation. Neuro Image 2003;19:226–32.

Schoffelen JM, Oostenveld R, Fries P. Neuronal coherence as a mechanism of effective corticospinal interaction. Science 2005;308(5718):111–3.

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

Srinivasan R, Russell DP, Edelman GM. Increased synchronization of neuromagnetic responses during conscious perception. J Neurosci 1999;19(13):5435–48.

Tallon-Baudry C, Bertrand O, Fischer C. Oscillatory synchrony between human extrastriate areas during visual short-term memory maintenance. J Neurosci 2001;21(20):1–5.

Tukey JW. Bias and confidence in not-quite large samples. Ann Math Stat 1958;29(2):614.