ELSEVIER

Electroencephalography and clinical Neurophysiology 102 (1997) 240–247

Testing topographic differences between event related brain potentials by using non-parametric combinations of permutation tests

Lídice Galán*, Rolando Biscay, Juan Luis Rodríguez, Maria Cecilia Pérez-Abalo, R. Rodríguez

Cuban Neuroscience Center, Av. 25, Playa, Havana, Cuba

Received 5 September 1995; revised version received 1 October 1996; accepted for publication: 10 October 1996

Abstract

MANOVA and repeated measures ANOVA approaches have provided evidence of a number of limitations in several event-related potential (ERP) studies due to violations of their statistical assumptions and the typically moderate size of the available sample. Alternative, computer-intensive methods based on permutation principles have recently been developed. Up to now this methodology has focused mostly on magnitude differences between scalp distributions as measured by t statistics. In this paper the scope of permutation techniques in ERP analysis was widened. A new statistic (D statistic) is introduced to compare the shapes of scalp distributions of ERPs. Additionally a general non-parametric combinatory technique is introduced to evaluate, by means of multivariate permutation tests, several time points and/or recording sites in ERP data. The methodology described here was used to test if two ERP components elicited during word-pair matching tasks to semantic or phonological incongruences had different scalp distributions. © 1997 Elsevier Science Ireland Ltd.

Keywords: Statistical analysis; Event-related potential; Brain topographic analysis; Permutation tests

1. Introduction

Event-related potential (ERP) components are usually identified by their polarity, latency, amplitude, and scalp distribution of brain electric potentials recorded under different experimental conditions (Reagan, 1989). In particular, scalp distributions are relevant for identifying ERP components due to the biophysical fact that different scalp distributions of potentials imply different underlying cortical current sources (Nunez, 1981). Thus, topographic differences between ERPs provide useful information on the existence of distinct neural generators involved in brain activity.

Unfortunately, the appropriate statistical methodology for testing topographic differences between ERPs remains problematic (see for example the editorial policy of the journal *Psychophysiology* on this subject published by Jenning et al., 1987). This is mainly due to the difficulties involved in the statistical treatment of the complex spatio-temporal structure of the dependencies present in ERP data.

Hitherto the most common approach to this testing problem has been repeated measures analysis of variance (rm-ANOVA). Usually, for each recording electrode p and experimental condition c, the electric response of each subject is averaged over trials and time instants (within a certain time window of interest). Baseline corrections are made subtracting the mean amplitude of a pre-stimulus time window. Finally to reduce subject-dependent and condition-related scale factors, affecting the ERPs (McCarthy and Wood, 1985; Carballo et al., 1992) which greatly increase the variance of the data, normalization procedures such as those proposed by McCarthy and Wood (1985) are commonly used. The resulting variables are then assumed to be described by a two-way rm-ANOVA model with two factors: *experimental condition* (which can be expanded into different effects), and *localization* on the scalp. The treatment of scalp location as a repeated (within) factor is an attempt to take into consideration the spatial correlation of the data. According to this model, differences in the scalp distribution across experi-

* Corresponding author.

PII S0921-884X(96)95155-3

EEG 95155

but we applied instead the general combinatory procedures described above to construct the multivariate test (see Section 3 on non-parametric combination of permutation dependent tests and Appendix A).

Permuted one-sided t statistics were calculated at Cz (the site of maximum negativity) across time and their corresponding significance values were combined separately for both components. This analysis was carried out within the time regions in which the negativities were visualized. Fig. 2 summarizes the results. For the SEM task (Fig. 2a) a significant N400 effect ($P < 0.05$) was present in a time region from 340 to 408 ms. The significant time region ($P < 0.05$) for the PHON task was at a later latency from 464 to 532 ms (Fig. 2b). The significant time regions thus selected will be used as the analysis windows from now onwards.

In the following sections the topographical differences between N400 and N450 will be examined with different methods. Differences in task difficulty, which could possibly influence the results, should be evaluated first. Thus behavioral data (number of errors) were analyzed for each subject and task. Mean error values (across sample) were SEM = 0.6, S.D. = 0.96 and PHON = 1.0, S.D. = 1.2. No significant differences were found using a non-parametric Wilcoxon test for matched pairs ($N = 10$, $T = 2.5$, $Z = 0.91$, $P < 0.36$).

4.4. Testing topographic differences between N400 and N450 with traditional methods

In this section we follow traditional methods to evaluate topographical differences between two ERPs, i.e. we carry out a repeated measures analysis of variance (Component × Site). The repeated factors Component and Site have two and 10 levels, respectively. Differences in the scalp distribution between components are detected in this model as significant interactions between the two factors.

In order to compare N400 and N450 topography, difference waveforms were calculated by subtracting match and mismatch ERPs for each task and subject. Mean amplitude values were obtained for each individual difference ERP waveform as follows. All time points within the preselected time windows (340–408 ms for N400 and 464–532 ms for N450) were averaged across time.

Geisser-Greenhouse correction of degrees of freedom was used when required. The rm-ANOVA did not detect any difference between the topographic distributions of the N400 and N450 components. The only significant effect corresponded to the main effect SITE ($F(9,1) = 415.91$, $P < 0.03$). Neither the Component effect ($F(1,9) = 0.03$, $P < 0.85$) nor the interaction Site × Component ($F(9,1) = 0.52$, $P < 0.79$) were significant. However, these results should be evaluated with caution due to the well-known drawbacks of the rm-ANOVA approach, especially taking into consideration that the sample size is not large.

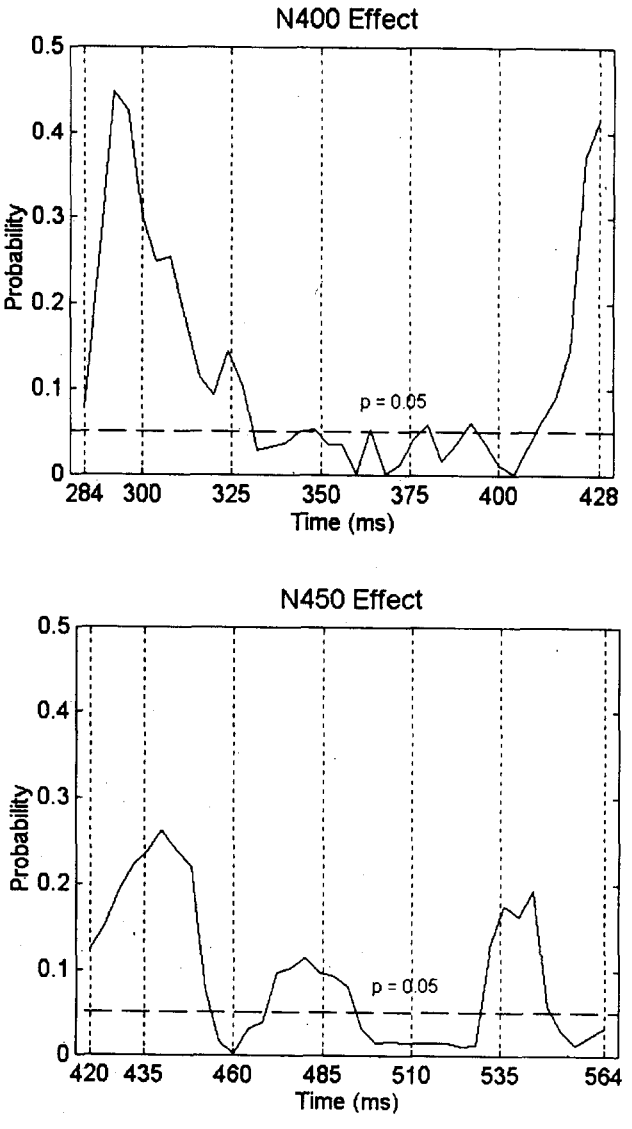

Fig. 2. Assessment of the mismatch effect in the semantic and rhyming tasks with one sided permuted t tests. (a) Results of the comparison between the ERPs elicited by match and mismatch word pairs in the semantic task. The observed significance values of the t test were plotted for each time point within the analyzed window. Significant P values (less than 0.05) are those under the thick line. Note that the two ERPs differed significantly in a time region from 340 to 408 ms. In this region an N400 effect would be present. (b) Results of the comparison between the ERPs elicited by rhyming and non-rhyming word pairs. The curve shows the observed significance value of the t test calculated for each time point within the analyzed window. Significant P values (less than 0.05) are those under the thick line. Note that the two ERPs differed significantly in a time region from 464 to 532 ms. In this region an N450 effect would be present.

4.5. Testing topographic differences between N400 and N450 with permutation techniques

4.5.1. Statistic for testing magnitude differences

Permutation t statistics were computed to evaluate possible differences in magnitude between N400 and N450 topography. Permuted t statistic between N400 and N450

nents will be used to demonstrate the application of the D statistic as well as the statistical methodology developed to combine permuted tests.

4.1. Experimental procedure

Subjects ($N = 11$, with ages ranging from 18 to 35 years) performed a semantic (SEM) and a phonological (PHON) matching task, during which ERPs were recorded. The order of the tasks was counterbalanced over subjects. In the matching tasks, two randomized sequences of stimulus pairs were presented, one for each task. The subjects had to detect 50% of the pairs in which the two stimuli were closely associated in meaning (in the SEM task), or 50% of the pairs which rhymed in the PHON task. The two stimuli in a pair were presented sequentially and with an onset asynchrony of about 2 s. The inter-pair interval lasted for about 5 s. All words employed were content words in Spanish and of high frequency of use (mean frequency: 44 per million). Rhyme and semantic association pairs were obtained from a larger pool of words rated by a different sample of 10 subjects.

4.2. Recording conditions

The electroencephalographic (EEG) activity was recorded with a MEDICID III/M system from 10 sites (F3, P4, F7, F8, Cz, Pz, T5, T6, O1, O2) of the international 10-20 system. Disk Ag/AgCI electrodes were used and interelectrode impedance was kept below 5 kΩ. Linked earlobes were used as reference and the forehead was grounded. Two bipolar derivations were used to monitor the horizontal and vertical electro-oculogram (EOG). The EEG after amplification and filtering from 0.05 to 30 Hz, was digitized with a 12 bit converter. Digitization was synchronized with the onset of the second stimulus in each pair, with a sampling period of 4 ms, and was stored on a magnetic disk for off-line analysis. Epochs of 1024 ms were selected on each trial, with a 100 ms pre-stimulus window. Each EEG segment was visually inspected and those with artifacts, eye movements, or incorrect responses were eliminated. For every subject, averaged ERPs were obtained separately for match and mismatch trials. The resulting vectors were baseline corrected by subtracting the average pre-stimulus amplitude value. When scaling was required, the minimum from each data point was subtracted and divided by the difference between maximum and minimum (where the minimum and maximum values are computed over all times and sites) as proposed by McCarthy and Wood (1985).

4.3. Assessment of the mismatch effect in the semantic and rhyming tasks

In order to visualize the negativities elicited by semantic or phonological incongruences, the ERPs to match and mismatch trails were subtracted. The difference waveforms averaged across subjects for each matching task are shown in Fig. 1. A late negative component can be visualized in both tasks (namely N400 and N450). For the SEM task this component occurs in a latency region between 284 and 428 ms and has a maximum amplitude at frontal and central recording sites. For the PHON task the negativity is at a later latency interval (420–564 ms) and has a central posterior scalp distribution.

The characterization of these components cannot be made only by visual inspection of the waveforms. The presence of an N400 or N450 should be demonstrated statistically, as a significant difference between the mean amplitude (within a time window of interest) of the ERPs to match and mismatch trials. To narrow down the time window at which these components were present, permutational techniques could be useful.

For this purpose, we followed a similar approach to Blair and Karniski (1993) (permuted one-sided t statistic),

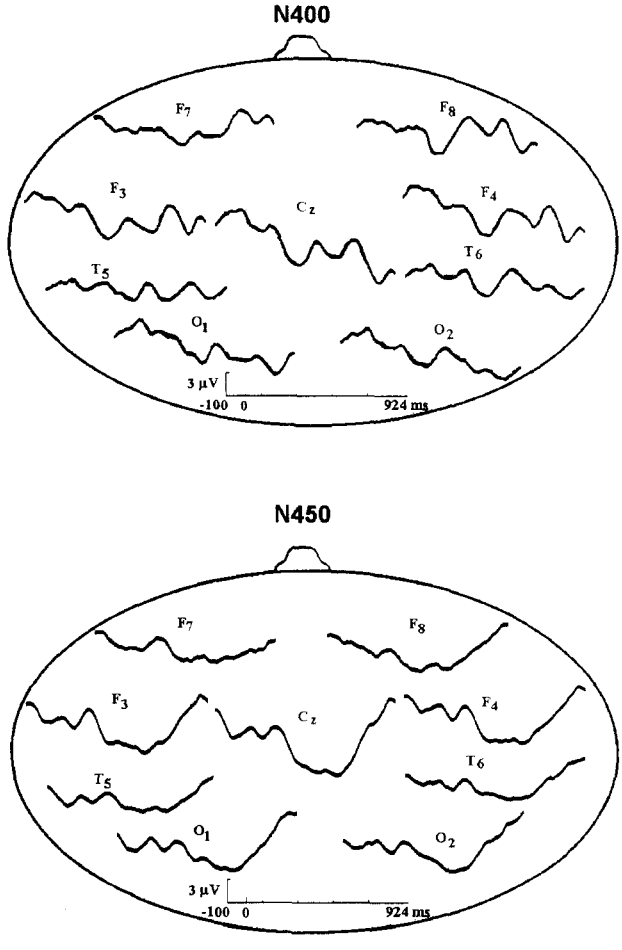

Fig. 1. The grand average (across subjects) difference ERPs (match-mismatch) were plotted at different recorded sites separately for each task. At the top the negativity elicited by semantic incongruences (N400) and at the bottom the rhyming incongruences (N450). Negativity was plotted downwards.

by the 10-20 international recording system). Let $f(1, \mu)$... $f(K, \mu)$ be the recording sites arranged in decreasing order of their potential values (i.e. components of μ). In a similar fashion, define $g(1, \beta)$,..., $g(K, \beta)$ for the ERP vector β. Then $f(1, \mu)$ and $g(1, \beta)$ are the locations on the scalp of the maximum values of the ERPs μ and β, respectively.

Consider the hypothesis of equality of locations of the maximum values, i.e. H_0: $f(1, \mu) = g(1, \beta)$. A sensitive measure of deviation from this hypothesis is: $D(f,g) = d(f(1, \mu), g(1, \beta))$, where $d(.,.)$ denotes the distance on the scalp between two sites. A simple way of calculating the distance d would be by means of the approximate representation of the head by a sphere of radius $R = 55/2\pi$ cm. The distance on the scalp between two sites is thus calculated as the arc distance between two points on the sphere. The estimation of this measure provides a sensitive statistic for testing this hypothesis. Specifically, suppose that both ERPs are recorded on each subject of a sample of size N; then, averaging across subjects, one obtains estimates $\bar{X}$ and $\bar{Y}$ of their mean values μ and β, and the estimate $D(\bar{X}, \bar{Y}) = d(f(1, \bar{X}), g(1, \bar{Y}))$ of D. Notice that this magnitude, which will be referred as the D statistic, is designed to compare a specific aspect of the shapes of the scalp distributions: it gives larger values as the distance between the locations of the maximum peaks of the two ERP landscapes increases.

A permutation test can be constructed on the basis of this statistic by using standard permutation techniques (Eddington, 1987). A number n of permuted samples is generated (each sample is obtained by random permutation of the two ERPs within each subject of the original sample), and the empirical probability distribution F^* of the resulting values of the D statistic is calculated. Then the test (at the significance level α) consists in rejecting H_0 when $1 - F^*(D(\bar{X}, \bar{Y})) < \alpha$.

This approach can be generalized for testing the hypothesis of equality of the locations of the $m(1 \leq m \leq K)$ largest values of two ERPs, i.e., H_0: $f(i, \mu) = g(i, \beta)$, $1 \leq i \leq m$. The generalized statistic is:

$$D(\bar{X}, \bar{Y}) = \frac{d(f(1, \bar{X}), g(1, \bar{Y})) + \ldots + d(f(m, \bar{X}), g(m, \bar{Y}))}{m}.$$

This type of statistic will also be referred to as the D statistics, and the associated tests as D tests. Notice that this statistic is rank-based, so it is insensitive to differences between ERP topographies when they have the same order of the potential values across electrodes.

The shape of the potential can be assessed by the D statistic considering not only its most prominent peaks but also the valleys. In this case, in the computation of the D statistic the estimated mean ERP vectors $\bar{X}$ and $\bar{Y}$ could be replaced by their inverted polarities $-\bar{X}$ and $-\bar{Y}$. Furthermore, if the D statistic is calculated with the absolute values of the mean potentials, sensitivity to both peaks and valleys is achieved.

3. Non-parametric combination of permutation dependent tests

Consider an experimental design in which multichannel ERPs (k recording sites and t time-points) are obtained under two different conditions (treatments) for each subject. The hypothesis H_0 of equality of the mean values of the two ERPs can be decomposed into the marginal hypotheses H_{0i}: $\mu_{dt1} = \mu_{dt2}$, where μ_{dtj} denotes the mean value of the ERP obtained at the scalp site d at time t under the experimental condition j.

Simple and consistent test statistics for the marginal hypotheses can be obtained by using permutation techniques (Eddington, 1987; Blair and Karniski, 1994) thus overcoming probability restrictions such as normal parent distributions and large sample sizes.

In order to make a decision for the overall hypothesis H_0, a possibility would be to summarize the marginal statistical tests. Blair and Karniski combined marginal permutation t statistics by calculating their maximum value, or sum of squares (Blair and Karniski, 1993; Blair and Karniski, 1994). A limitation of this procedure is that it requires homogeneity in the probability distribution of the marginal statistics. This condition is difficult to guarantee for ERP data.

Alternatively, we applied the general procedure for the non-parametric combination of marginal dependent permutation tests developed theoretically by Pesarin (1992) to ERP data. This procedure is not subjected to the limitation above mentioned, shows good theoretical properties, and under homogeneity conditions it can be reduced to the procedures used by Blair and Karniski. In our particular application with ERP data we propose to combine instead the observed significance values of the permuted statistic and not the values of the statistic. This would guard against possible non-homogeneities in the statistic probability distribution. The methodology used is detailed in Appendix A.

The non-parametric combination of marginal permutation tests offers great flexibility for testing global hypotheses in ERP analysis. It can be applied for: (i) a set of electrodes at a fixed time instant, (ii) a set of time points at a given electrode, and (iii) all the electrodes and time points (within a time window of interest). Furthermore, marginal tests based on any kind of statistics can be combined (t tests, D tests, etc.).

4. Application in a psychophysiological experiment

Two ERP components have been described during printed word-pair matching tasks to semantic (N400, Kutas and Hillyard, 1983) and rhyming incongruences (N450, Rugg, 1984). Experimental data on these compo-

mental conditions are interpreted as significant interactions of the two factors.

However, the rm-ANOVA approach has some well-known limitations. It assumes that error vectors have a multivariate normal distribution with a particular structure of covariance, namely the Huynh–Feldt structure (Huynh and Feldt, 1970). In general this structure is not present in spatial data where correlations decrease as the distances between the measurement points increase. Brain topographic data typically show complex spatial characteristics, which makes the Huynh–Feldt structure a questionable assumption (Vasey and Thayer, 1987; Valdés et al., 1992). Furthermore, the use of normalization procedures can make the assumption of normal distribution less justifiable. For example, subtracting the minimum value and dividing by the difference between the maximum and minimum values of the recording (a normalization procedure proposed by McCarthy and Wood, 1985) yields variables within the range (0–1), which obviously cannot have a normal distribution. The consequence of the violation of such assumptions are biased test levels; in other words, lack of control of Type I error probability (alpha level).

Two main approaches have been advocated to guard against such bias. One is the use of the general MANOVA (Vasey and Thayer, 1987). But this has the disadvantage that the tests for localization effects and for localization-by-experimental condition interactions are undefined when the number of recording electrodes exceeds the number of subjects, as is the case in many ERP studies. The other approach is the use of degrees of freedom corrections in the rm-ANOVA F tests (Greenhouse and Geisser, 1959; Huynh and Feldt, 1976). But such corrections should be used with caution because they only guarantee approximate F tests (Vasey and Thayer, 1987; Raz, 1989).

Recently, computer-intensive methods based on permutation principles (Eddington, 1987) have been proposed as an alternative statistical methodology for testing differences between ERPs waveforms and maps (Raz, 1989; Blair and Karniski, 1993; Blair and Karniski, 1994). The methodology has a number of advantages: the tests are distribution free, no assumptions of an underlying correlation structure are required, and it provides exact P values for any number of subjects, time points and recording sites.

However, up to now permutation tests to evaluate differences in ERP topography have been exclusively based on t statistics. The morphological aspects of the complex spatio-temporal structure of ERP differences, which are not captured by t statistics, have not been specifically addressed. Furthermore, the combination of permutation t tests corresponding to different time points and/or recording sites have been based only upon symmetric functions of t values, such as the maximum value or the sum of squares (Blair and Karniski, 1993; Blair and Karniski, 1994). This way of combining the marginal statistics deals with all the variables in an homogeneous fashion, thus it should be applied with caution to situations in which the marginal t statistics show different probability distributions.

In the present paper the use of permutation tests in ERP analysis was further developed to overcome some of the limitations above mentioned. First a new statistic (D statistic) was introduced for testing shape differences between ERPs maps. This statistic is defined as the sum of the distances between two ERPs, rank-ordered across derivations. Thus it is selectively sensitive to differences in the relative peaks and valleys of ERP landscapes. Using shape information to interpret brain maps in terms of electric sources have been advocated by Lehmann and Skrandies (1984). This information though would be poorly reflected by t statistics.

A second aspect addressed in this paper refers to the application of the general non-parametric methodology, described theoretically by Pesarin (1992) for combining permutation tests. The methodology was applied here for the first time to ERP data. Also instead of combining the permuted values of the statistic, we combined their observed significance values in several time points and/or electrodes. This way of making multivariate extensions would guard against the effects of non-homogeneities in the statistic probability distribution.

Finally, the main features of the statistical methodology described here are demonstrated with ERP data obtained in a psychophysiological experiment. The experiment addressed the question of whether the negative components elicited during word-pair matching tasks to semantic (Kutas and Hillyard, 1983, N400) or phonological (Rugg, 1984, N450) incongruences have different topographical distributions.

2. The D statistic: a permutation test for shape comparison of ERPs maps

Any ERP can be characterized as a two-dimensional matrix (time points by scalp recording sites) of amplitude (voltage) values. The most common approach to evaluate the topography of an ERP component is to average across time (within a window of interest), thus reducing this matrix to a vector (mean amplitude values at the different recording sites). However, useful information on ERP topography could be also extracted for each time point within a pre-selected window of interest (sometimes called the instantaneous landscape of the ERP, see Lehmann and Skrandies, 1984). The permutation test described below was designed for testing topographical differences between ERP vectors defined either as the voltage map at a single time or the mean voltage map within a window of interest. (The generalization of this test to include multiple times will be considered in Section 3)

Let $\mu = (\mu_1 \ldots \mu_K)$ and $\beta = (\beta_1, \ldots \beta_K)$ denote two ERP vectors (at a fixed time t) recorded under two different experimental conditions at K scalp sites (e.g., sites given

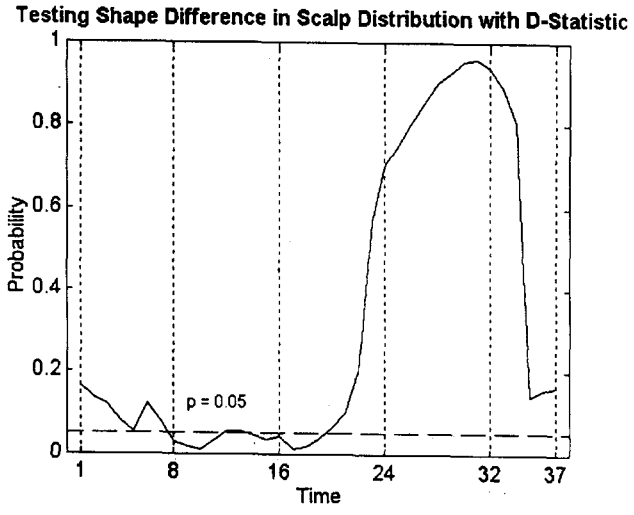

Fig. 3. Testing differences in topography between N400 and N450 with the t statistic. (a) Observed significance levels of permuted t tests (combined across recording sites) plotted for each time point within the pre-selected window. No significant values ($P < 0.05$) were obtained. (b) Observed significance levels of permuted t tests (combined across time) plotted for each recorded site. No significant values ($P < 0.05$) were obtained.

difference ERPs was calculated for each time point and all recorded sites. Results are summarized in Fig. 3. Significance values were plotted either as a function of time (combining for each time point across sites) (Fig. 3a) or recorded sites (combining for each site across time points) (Fig. 3b). Note that no significant differences were found between N400 and N450 with this methodology. These results are in agreement with those previously obtained with rm-ANOVA. No overall differences in magnitude between N400 and N450 scalp distributions were evidenced either with the more traditional approach (rm-ANOVA) or with permutational t test.

4.5.2. D Statistic for testing shape differences

To evaluate possible differences in the shape of the N400 and N450 scalp distributions, permuted D statistics were calculated.

First, in order to facilitate the comparison with results previously obtained by rm ANOVA, the ERP amplitude vector of sites was constructed in a similar way, i.e., averaging the amplitude value within the pre-selected time windows. Notice that scaling is irrelevant here, since the D statistic is not affected by scale factors.

To illustrate the scalp distribution of the two ERP vectors, their mean amplitude values were plotted superimposed at each of the ten recorded sites (Fig. 4).

Even though, there are no appreciable changes in the overall magnitude of their scalp distribution, the local valleys of the negativities are clearly differentiable. For N400 the maximal negativity (in decreasing order of magnitude) is reached at F4, F3 and Cz, whereas for N450 the more negative sites were O1, O2 and Cz.

These apparent differences could be statistically evaluated with the permutation D statistic (calculated from the inverted polarity potentials). The value of the D statistic was 7.015 cm corresponding to a probability $P < 0.02$. Thus, with this methodology we were able to demonstrate significantly different shapes for N400 and N450 mean amplitude scalp distributions. This could suggest distinct spatial patterns of activation for the two components.

A second way of using D statistic could be to evaluate the changes in shape occurring through time between two ERPs scalp distributions (namely the changes between N400 and N450 topographies across time). For this purpose, permutation marginal D statistics were calculated for each time point within a selected time window (340–408 ms for N400 and 464–532 ms for N450). The hypothesis (H_0) evaluated was the equality of the first 3 location values (see definition of the D statistic above). The statistic

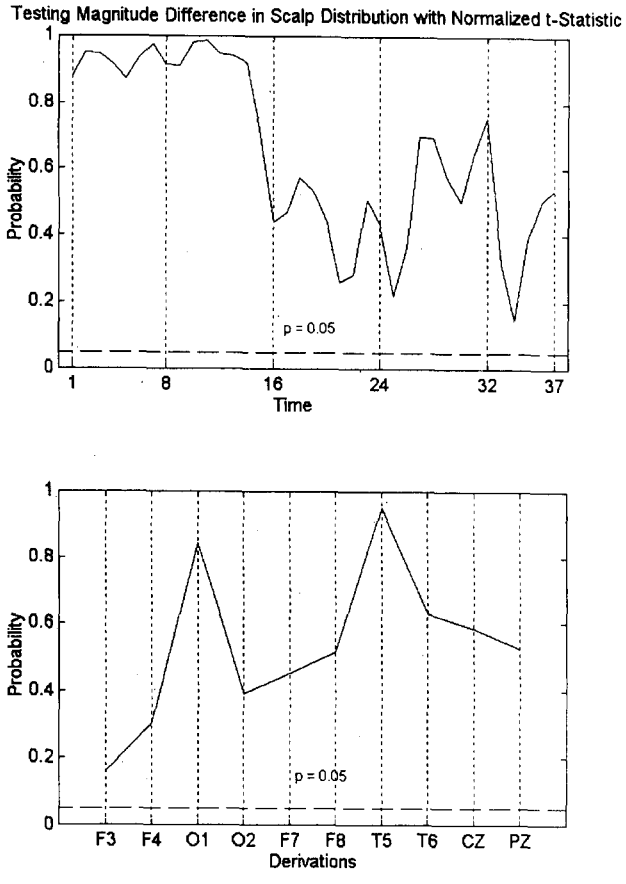

Fig. 4. Scalp distribution of N400 and N450. The vector across sites of mean amplitude values (calculated upon difference ERPs) was plotted superimposed for both tasks. Each curve represents the mean amplitude value calculated across time within a pre-selected window (340–408 ms for N400 and 464–532 ms for N450) for each of the 10 recorded sites. A continuous trace denotes the N400 and dashed lines were used for the N450. Note that N400 has a slightly more anterior scalp distribution than N450. This latter component has its maximum negativity at Cz, O2 and O1 while the N400 peaks at F3, F4, and Cz.

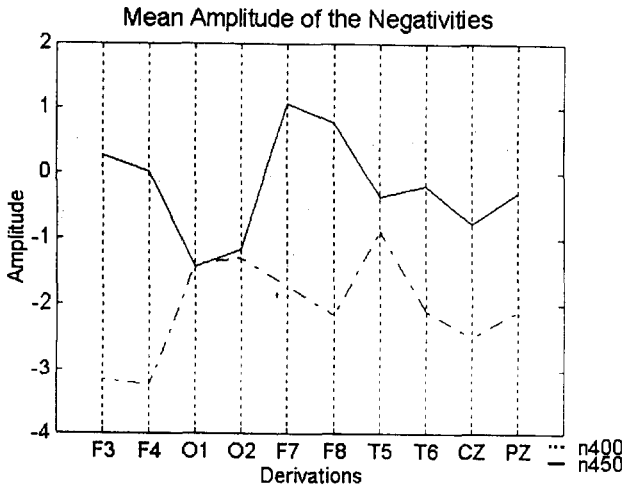

Fig. 5. Testing differences in topography between N400 and N450 with the D statistic. Observed significance levels of permutation D tests were plotted for each time point within the pre-selected window. Note that N40 and N450 differed significantly ($P < 0.05$) in the shape of their scalp distribution in a region from time point 8 to 19.

was combined across time to construct a multivariate test (using the maximum value as the combining function, see details in Appendix A).

The observed significance values of the univariate D statistic corresponding to each time point within the analyzed time window are shown in Fig. 5. Note that there is a region of significance extending from the 8 to the 19 time point. This suggests that there are also dynamic changes in the shape of the scalp topography or landscape of these components. Notice that the D statistic detected significant shape differences in their scalp distributions while no magnitude difference was detected by rm-ANOVA and permutation t tests.

The differences in topography evidenced with the D statistic suggest that the two components under study (N400 and N450) could reflect different underlying brain processes. Further studies should be undertaken to confirm the validity and consistency of this preliminary conclusion. The methods introduced and demonstrated in this paper may be used for this purpose.

5. Conclusions

A statistical methodology was described to evaluate relevant aspects in the complex spatio-temporal structure of ERPs. The introduced D statistic has all the well-known advantages of permutation techniques and is selectively sensitive to shape differences between ERP scalp distributions. Additionally, non-parametric combinations of univariate tests were further developed.

The analysis of ERP data with this methodology showed significant differences in the shape of N400 and N450 amplitude distributions. These results suggest that the two components could reflect different brain processes.

Acknowledgements

We are grateful to three anonymous referees and Ms. Thalia Harmony for their comments which contributed to improve the exposition of this paper.

Appendix A Non-parametric combination of dependent permutation tests

An overall hypothesis can be decomposed into a number of sub-hypotheses, or marginal hypotheses. We will describe here the specific algorithm developed for constructing a multivariate statistical test for the overall hypothesis.

This procedure is valid in the following general situation. The objective is to make a decision about the hypothesis H_0 that states the interchangeabilty of two (in general, dependent) random vectors $X = (X_1, \ldots, X_k)$ and $Y = (Y_1, \ldots, Y_k)$, on the basis of a sample Z of N independent observations of the whole vector of variables (X, Y). Two random vectors X and Y are regarded exchangeable if (X, Y) and (Y, X) have the same probability distribution. The exchangeability of X and Y implies that X has the same distribution as Y. It is assumed that the overall hypothesis H_0 can be decomposed into the sub-hypotheses H_{0i} each stating the equality in distribution of the variables X and Y which form the vectors. Thus, H_0 is true if and only if all the H_{0i} are true.

Under this situation it is usually easy to construct suitable univariate statistics to test the marginal hypotheses H_{0i}. Then the procedure combines these (possibly dependent) marginal tests to make a decision about the overall hypothesis H_0 according to the following steps:

1. Generate S randomly permuted sample of size N. each one obtained by means of a random permutation of the vectors X and Y within each observation of the original data.
2. Calculate the marginal statistics $d_i^*(Z^*)$ for each permuted sample Z^*, and calculate the empirical distribution function F_i^* of the obtained values for each statistic.
3. For each permuted sample Z^*, calculate the combined statistic $T^* = T(F_i^*(d_i(Z^*)))$, and calculate the empirical distribution function F^* of the resulting values. Here T is a suitable function, called a combining function (Pesarin, 1992).
4. Calculate the observed value of the combined statistic on the original sample, $T = T(F^*(d_i(Z))$
5. Reject H_0 if $1 - F^*(T) < \alpha$, where α is the significance level specified in advance. Reject H_{0i} if $1 - F^*(T_i) < \alpha$, where $T_i = F_i^*(d_i(Z))$.

The resulting multivariate test converges in distribution to an exact permutation test as the number S of simulations increases. Thus, it shows the well-known advantages of permutation tests. Under general conditions, the result-

ing test is consistent (i.e., the probability of detecting deviations from the null hypothesis tends to 1 as the sample size increases) if the marginal tests are consistent.The function T allows one to combine any simple marginal tests, regardless of the non-parametric dependencies between them, to obtain an overall test. The Tippet function $T(.) = \text{Max}(.)$, defined as the maximum value of its arguments, will be adopted. Thus, the observed overall statistic T is significant when at least one of the marginal tests $F_i^*(d_i(Z))$ has a significant (i.e., large) value according to the permuted distribution F^* of their maximum $T = \max(F_i^*(T))$. This has the advantage of showing which marginal hypotheses are rejected when the overall hypothesis is rejected.

References

Blair, R.C. and Karniski, W. An alternative method for significance testing of waveform difference potential. Psychophysiology, 1993, 30: 518–524.

Blair, R.C. and Karniski, W. Distribution-free statistical analyses of surface and volumetric maps. In: R.W. Thatcher, T.V. Hallett, T. Zeffiro, E.R. John and M. Huerta (Eds.), Functional Neuroimaging. Academic Press, New York, 1994, pp. 19–28.

Carballo, J.A., Riera, J.J., Biscay, R. and Valdés, P. Model parameters estimation when the evoked potential recordings are affected by a random scale factor. Int. J. Biomed. Comput., 1992, 30: 71–87.

Eddington, E.S. Randomization tests (2nd edn.), New York: Dekker, 1987.

Greenhouse, S.W. and Geisser, S. On methods in the analysis of profile data, Psychometrika, 1959, 24: 95–112.

Huynh, H. and Feldt, L.S. Conditions under which mean square ratios in repeated measurements design shave exact F-distributions. J. Am. Stat. Assoc., 1970, 65: 1582–1589.

Huynh, H. and Feldt, L.S. Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. J. Educ. Stat., 1976, 1: 69–82.

Jenning, J.R., Cohen, M.J., Ruchkin, D.S., Fridlund, A.J. Editorial policy on analyses of variance with repeated measures. Psychophysiology, 1987, 24(4): 474–478.

Kutas, M. and Hillyard, S.A. Event-related potentials to grammatical errors and semantic anomalies. Mem. Cognit., 1983, 11: 539–550.

Lehmann, D. and Skrandies, W. Spatial analysis of evoked potential in man: a review. Prog. Neurobiol., 1984, 23: 227–250.

McCarthy, G. and Wood, C.C. Scalp distributions of event-related potentials: an ambiguity associated with analysis of variance models. Electroencephalogr. Clin. Neurophysiol., 1985, 62: 203–208.

Nunez, P.L. Electric Fields of the Brain: The Neurophysics of EEG. New York: Oxford University Press, 1981.

Pesarin, F. A resampling procedure for non-parametric combination of several dependent tests. J. Int. Stat. Soc., 1992, 1: 87–101.

Raz, J. Analysis of repeated measurements using non-parametric smoothers and randomization tests. Biometrics, 1989, 45: 851–871.

Reagan, D. Human Brain Electrophysiology Evoked Potentials and Evoked Magnetic Fields in Science and Medicine. New York: Elsevier, 1989.

Rugg, M.D. Event related potentials in phonological matching task. Brain Language, 1984, 23: 225–240.

Vasey, M.W. and Thayer, J.F. The continuing problem of false positives in repeated measures ANOVA in psychophysiology: a multivariate solution. Psychophysiology, 1987, 24(4): 479–486.

Valdés, P., Bosch, J., Grave, R., Hernandez, J., Riera, J., Pascual R. and Biscay, R. Frequency domain models of the EEG. Brain Topopgr., 1992, 4: 309–319.