Research Article

Statistics in Medicine

Received 7 August 2008, Accepted 8 October 2009 Published online 25 November 2009 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3784

False discovery rate and permutation test: An evaluation in ERP data analysis

Agustín Lage-Castellanos,[a]*† Eduardo Martínez-Montes,[a]
Juan A. Hernández-Cabrera[b] and Lídice Galán[c]

Current analysis of event-related potentials (ERP) data is usually based on the *a priori* selection of channels and time windows of interest for studying the differences between experimental conditions in the spatio-temporal domain. In this work we put forward a new strategy designed for situations when there is not *a priori* information about 'when' and 'where' these differences appear in the spatio-temporal domain, simultaneously testing numerous hypotheses, which increase the risk of false positives. This issue is known as the problem of multiple comparisons and has been managed with methods that control the false discovery rate (FDR), such as permutation test and FDR methods. Although the former has been previously applied, to our knowledge, the FDR methods have not been introduced in the ERP data analysis. Here we compare the performance (on simulated and real data) of permutation test and two FDR methods (Benjamini and Hochberg (BH) and local-fdr, by Efron). All these methods have been shown to be valid for dealing with the problem of multiple comparisons in the ERP analysis, avoiding the *ad hoc* selection of channels and/or time windows. FDR methods are a good alternative to the common and computationally more expensive permutation test. The BH method for independent tests gave the best overall performance regarding the balance between type I and type II errors. The local-fdr method is preferable for high dimensional (multichannel) problems where most of the tests conform to the empirical null hypothesis. Differences among the methods according to assumptions, null distributions and dimensionality of the problem are also discussed. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: ERP; false discovery rate; permutation tests; multiple comparisons

Introduction

Event Related Potentials (ERP) are obtained by averaging epochs of electric potential measured on the scalp that are time-locked to the onset of a stimulus or motor event. The statistical analysis of ERP waveforms answers two main questions: (a) Is there any particular signal related to the stimulus, i.e. an ERP component? and (b) Is there any difference between ERPs of two experimental conditions? In most of cases the researcher knows the answer to the first question and it is just interested in knowing the answer to the second question. The latter can be looked up at a particular recording site (channel or derivation) and a particular time window, which implies the use of an experimental model and the testing for specific hypothesis. However, in exploratory analyses it is more useful to look in the entire spatio-temporal domain for the possible differences, which leads to a high-dimensional problem of simultaneous testing of hypothesis [1].

In this view, the statistical interrogation of ERP waveforms depends first on the domain of the analysis (temporal, spatial or spatio-temporal) and second, on whether the hypothesis being tested makes or not specific predictions about how a component should vary between experimental conditions [2]. Temporal interrogation refers to find 'when' significant effects appear in the ERP signals, whereas spatial interrogation refers to 'where' those effects are present. In the spatio-temporal case, both questions need to be answered simultaneously. On the other hand, testing specific hypotheses usually implies the *ad hoc* selection of one or a few channels of interest, as well as a small time window, in which the ERP is averaged in order to reduce the dimensionality of the problem. In the case of the search for differences in the whole temporal or spatio-temporal domains, the experiment is said to have an observational design [3]. In this type of studies there is no *a priori* knowledge about 'when' and 'where' the differences between ERPs appear, thus leading to effect-unspecific hypotheses.

[a] *Neurostatistics Department, Cuban Neuroscience Center, Havana, Cuba*
[b] *Universidad de la Laguna, Tenerife, Spain*
[c] *Department of Cognitive Neuroscience, Cuban Neuroscience Center, Havana, Cuba*
[*] *Correspondence to: Agustín Lage-Castellanos, Neurostatistics Department, Cuban Neuroscience Center, Havana, Cuba.*
[†] *E-mail: mlsttin@gmail.com*

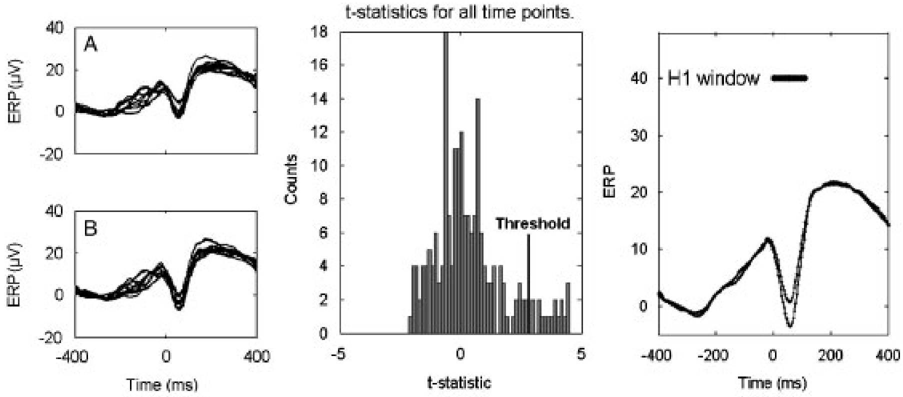

Figure 1. Procedure for the statistical analysis of ERPs in the whole time domain (single channel). Left: The population of ERP waveforms corresponding to two different experimental conditions A and B (in this case, these are simulated ERPs for 16 subjects, following the Scenario II, explained in the Methods) Central: Histogram of the statistics computed (in this case t-tests), from which (or from some theoretical null distribution) a threshold value can be determined. Right: Those statistics higher than the threshold represent rejections of the null hypothesis (alternative, H1), thus showing which time points (black points) present statistically significant differences between ERPs of both conditions.

The generic procedure for finding statistically significant differences between ERPs consists of two main steps: first, computing a measure or statistic of the 'difference' between waveforms for every space/time point and second, determining a threshold for the selection of the relevant statistics, i.e. rejection or acceptance of the null hypothesis. Figure 1 illustrates these steps in the comparison of two simulated ERPs.

The simplest but still commonest statistic in the ERP analysis is the univariate *t*-Student for each time point and/or derivation [2, 4, 5]. Other approaches include combining univariate *t*-tests with general linear models [6, 7] or using multivariate parametric analysis [8].

Theoretically, parametric methods are not suitable since time points are known to be highly correlated and there is usually a small number of subjects. However, the appropriateness of the use of massive univariate *t*-tests and the treatment of the temporal dimension was deeply discussed in Kiebel and Friston [7], concluding that this approach is a useful strategy. In their work, they chose to use time as a factor in order to make inferences on temporally extended effects. At the same time, they recognized the validity of the treatment of time as an independent dimension and deal with the temporal correlations in a separate step [7]. In this sense, the massive univariate *t* statistics allows the identification of channels and/or time points where relevant differences between the experimental conditions appear; avoiding the *a priori* selection of time windows, specific latencies and channels (results are considered *post-hoc*). The statistical challenge of this procedure, as some authors have pointed out [2, 9], is the increasing risk of type I errors due to the high number of hypotheses tested simultaneously. The type I error is defined as the rejection of a hypothesis that is a true null hypothesis, which is also called a false-positive finding.

The simultaneous analysis of large sets of hypotheses is known in the statistical literature as the problem of multiple comparisons and has been a topic of intense research in the last decade [10]. This issue has been faced by the use of techniques that control the ratio of false-positive findings inside the set of alternative hypotheses. Some works have proposed the use of the Bonferroni correction [11], the Random Field Theory [7, 8, 12–14], or the use of computer-intensive methods based on the permutation test [4, 5, 15–25].

In this work we intend to introduce in the field of ERP analysis, the use of methods devised to control the ratio of false-positive findings, based on the false discovery rate (*FDR*) statistic. There are two basic or principal FDR methods. A first one proposed by Benjamini and Hochberg [26] consists of a procedure that controls the global proportion of false-positive findings. A conceptually different approach, named (local-fdr), was recently developed by Efron [27] and is based on estimating the probability density function of the FDR directly from the actual data.

Based on particular improvements to these approaches, several new techniques have been developed. In a similar spirit of the Benjamini-Hochberg method, Ge [28] designed a Holm-type procedure with the objective of increasing power. With the same purpose Langers [29] exploited the spatial nature of neuroimaging data and Finos [30] used data-driven weights for estimating the global *FDR* statistic. Related with the local-fdr algorithm, Ploner [31] proposed a multidimensional extension of the method while Schwartzman [32] expanded the approach for exponential distributions. Based on earlier information, Robin [33] developed a semi-supervised approach to estimate the local *FDR* statistics. For modeling the null distribution required in the local-fdr estimation, Xie [34] used a permutation approach and applied it to microarray data.

This family of methods has been explored in the field of neuroimaging [35–37] and time–frequency analysis [38], as well as in the analysis of microarray data [34], but is unfamiliar for the ERP community. Therefore, despite the wide variety of improvements reported in the FDR literature, this article will focus on applying the two principal FDR methods in their initial form to the analysis of ERPs. We aim at providing an initial validation of the usefulness of the application of FDR methods in the ERP field.

In particular, the main goals of this article are the introduction of both FDR methods and the comparison of their performance with permutation test in the analysis of simulated and real ERP data. This is first explored in single channel problems and then extended to multichannel problems, which is a typical example of large-scale testing. We will also study the differences between the empirical null distributions obtained by local-fdr and permutation test with the theoretical null distribution of the *t* statistic.

Methods

Notation and terminology

Consider the ERP data sets obtained in two different experimental conditions (A and B), for N_s subjects, which are recorded in N_d derivations and N_t time points. Then, searching for differences between both ERPs in the whole spatio-temporal domain leads to the simultaneous testing of $m = N_d N_t$ hypotheses, denoted by H_{ij}, where *i* and *j* refer to the spatial and temporal dimensions, respectively. Each hypothesis H_{ij} is tested by the statistic z_{ij}, which in this work will correspond to the paired *t* statistic between the experimental conditions, although for the methods studied in this article other statistics can be used. Assigning $H_{ij} = 0$ implies the acceptance of the null hypothesis for the *ij*-th test while $H_{ij} = 1$ implies rejection of the null hypothesis (acceptance of the alternative) for the *ij*-th test. The Student's *t* probability distribution with $N_s - 1$ degrees of freedom will be used as the theoretical null for comparative purposes.

Permutation test ('$T_{\max}$')

Permutation test is based on an intuitive logic. Consider the comparison of ERP data sets between experimental conditions A and B, focusing on a particular spatio-temporal hypothesis H_{ij}. Then, if there is no actual difference between conditions, the labeling of the ERP data is arbitrary. It means that the same data would have arisen whatever the experimental condition is [15, 19, 21].

To obtain the distribution of the test statistic z_{ij} at a particular derivation and time under the null hypothesis, the method generates data sets by randomly permuting the condition's labels. Using each permuted data set the test statistic $\tilde{z}_{ij}$ under the null hypothesis is computed. Repeating the process a large number of times (N_p), a set of test statistics is obtained; whose histogram defines the empirical permutation distribution of the null hypothesis. Then, the null hypothesis is rejected at a significance level α if the actual statistic z_{ij} is greater than the $1-\alpha$ percentile of the empirical permutation distribution [17, 20, 22].

To this point we have considered the permutation test for a particular derivation and a particular time. The application of the method for the entire spatio-temporal domain implies the simultaneous testing of all H_{ij} hypotheses. In this case we expect that a fraction of the null hypotheses is considered as alternative ($H_{ij} = 1$) only by chance (false positives). The control of these false rejections of the null hypothesis must be taken into account in the permutation mechanics by using an appropriate maximal statistic. If we test the omnibus hypotheses ($H_{ij} = 0, \forall i, j$), the maximum of the $\tilde{z}_{ij}$ in the whole spatio-temporal domain is a suitable statistic to control the multiplicity of tests [20, 23]. Then, to reject the omnibus hypothesis at a significance level α, the actual maximal statistic in the whole spatio-temporal domain must exceed the $1-\alpha$ percentile of the empirical permutation distribution of the maximum. More relevant, we can assign a particular hypothesis $H_{ij} = 1$ at the level α, if the actual statistic z_{ij} is largest than the $1-\alpha$ percentile of the distribution of the maximum. This procedure controls the Family-wise Error Rate (FWER), which is the probability of having at least one false positive among the whole set of hypotheses considered as an alternative at the desired level α [17]. In this work we will use $\alpha = 0.05$, which means that across 100 repetitions of the same experiment we expect that in 5 of them at least one false-positive finding exists.

The detailed procedure used here is as follows. First, generate N_p permuted data sets, keeping the same spatio-temporal order in the ERP time series. For each permuted data set compute the permuted test statistic matrix $\tilde{Z}^l (l = 1, 2, \ldots, N_p)$ (the element $\tilde{z}_{ij}^l$ reflects the comparison between experimental conditions at derivation *i* and time *j*, for the *l*-th permuted data set). Then, for each test statistic matrix $\tilde{Z}^l$, extract the maximum $\tilde{T}^l = \max(\tilde{Z}^l)$. The empirical distribution of the maximum is formed by $\tilde{T}^l$ for all *l*. The $1-\alpha$ percentile of this distribution defines the threshold T_α at a significance level α. Then, assign $H_{ij} = 1$ to those individual hypotheses whose actual statistics z_{ij} exceed T_α.

False Discovery Rate methods

Two different *FDR* statistics have been stated and sometimes confused in the literature: the global false discovery rate (FDR) [26] and the local false discovery rate (*fdr*) [27]. They are both related with the number of false positives in a subset of rejected hypotheses. Consider a problem in which $m = N_d N_t$ hypotheses H_{ij} $(i = 1, \ldots, N_d; j = 1, \ldots, N_t)$ are tested simultaneously. Then the set of rejected hypotheses is defined as $H_{ij} : |z_{ij}| > z$, for a predefined cutoff *z*. For simplicity, in the following we omit the absolute value and use the notation for the right tail, since the generalization to the two-tailed case is straightforward. The *FDR* is the probability of $H_{ij} = 0$ given that z_{ij} fulfills the condition to be included in the set of rejected hypothesis:

$$FDR(z) = \Pr(H_{ij} = 0 | z_{ij} > z) \quad (1)$$

The *FDR* and the FWER are different measures of type I error. They are equivalent only in the case when all tests are true null [23]. The *fdr*, in turn, refers to the probability of $H_{ij} = 0$ given a particular value *z* for the statistic z_{ij}:

$$fdr(z) = \Pr(H_{ij} = 0 | z_{ij} = z) \quad (2)$$

Both measures are confused since *fdr* is often interpreted as the expected proportion of false-positives discoveries in the selected subset. The correct interpretation is that *fdr* represents the probability density of false positives in the whole domain of z_{ij}, while

FDR is the expected value of the *fdr* in the tail of the mixture of null and non-null hypotheses distribution $f(z)$. In practice, the *FDR* statistic can be estimated as the average of the *fdr* in the region $z_{ij}>z$ [27]:

$$FDR(z)=E(fdr(z_{ij})|z_{ij}>z) \tag{3}$$

Benjamini and Hochberg [26] proposed a first approach to estimate the *FDR* (hereinafter 'BH' method) based on the uniformity of the p-values under the null hypothesis. In practice, the method ranks the m p-values from the smallest to the largest to obtain a sequence $\{P_k\}$ and classifies as $H_{ij}=1$ those hypotheses having p-values lower than P_l, where l is given by

$$l=\arg\max_k\left(P_k<\frac{k}{ma}q\right),\quad a=\sum_{k=1}^{m}\frac{1}{k} \tag{4}$$

where q is the upper bound for the *FDR* chosen by the researcher. This procedure controls that $FDR<p_0q$, where p_0 is the proportion of actual null hypotheses in the set of statistics. Since p_0 is not known, this means that q is an upper bound for the *FDR*, corresponding to the worst possible case where all hypotheses are true null ($p_0=1$).

The constant a is a correction applied for dependent tests [39], which can be ignored ($a=1$) if tests are considered independent. The equation for a reflects the strongest possible dependence among p-values and increases monotonically with m, which makes the BH method very conservative when a large set of hypotheses is tested. Therefore, this correction is frequently ignored in the analysis of high-dimensional data, where certain degree of dependence among p-values is always present [23]. Here we will report the results for both, the dependent ('BHd') and independent ('BH') variants of the method.

The BH method is parameter-free and the result only depends on the lower tail of the histogram of p-values, consequently being robust to variations in this histogram for large p-values. Unfortunately, the level of conservativeness increases with the number of hypotheses (m), leading to a reduction in statistical power in high-dimensional problems [40, 41] such as multichannel testing.

More recently, Efron proposed an algorithm named local-fdr to estimate the *fdr* statistic [27]. This is based on modeling the histogram of the actual tests z_{ij}. This histogram is modeled as a 'two class' bayes mixture:

$$f(z_{ij})=p_1f_1(z_{ij})+p_0f_0(z_{ij}) \tag{5}$$

The null $H_{ij}=0$ class occurs with prior probability p_0 and has density $f_0(z_{ij})$ while the alternative $H_{ij}=1$ class occurs with prior probability $p_1=1-p_0$ and has density $f_1(z_{ij})$. The *fdr* can be defined in terms of these densities [27]:

$$fdr(z)=\Pr(H_{ij}=0|z_{ij}=z)=\frac{p_0f_0(z)}{f(z)} \tag{6}$$

Estimation of $fdr(z)$ is achieved by modeling separately $p_0f_0(z)$ and $f(z)$. The mixture density $f(z)$ is estimated using a parametric exponential function or a natural spline basis fitted to the histogram of z_{ij}. The estimation of $p_0f_0(z)$ becomes a critical issue in the procedure and is achieved by using the central matching algorithm. This algorithm assumes that $f_0(z)$ is predominant in the data ($p_0\gg p_1$) around $z=0$ and fits a second degree exponential model to the histogram of z_{ij} in this neighborhood. Then, $f_0(z)$ is expanded to the whole range of z for computing the quotient in equation (6). Since the accuracy of the method depends on these fitting procedures, this method is not affected by increasing the dimensionality, on the contrary, the higher the number of tests, the better the histogram modeling.

In practice, the selection of the threshold for classifying relevant and null hypotheses in both FDR methods is achieved by setting a value for q (known in the FDR literature as q-value [42]). As stated above, q is an upper bound of the *FDR*, thus the methods control this statistic. In the case of the local-fdr algorithm, the *FDR* is computed for each hypothesis (equation (3)) and those with *FDR* below the predefined q-value are classified as 'relevant' or alternative hypotheses. An accepted standard in neuroimaging studies for the q-value cutoff is 0.05 [29, 35, 43]. It means that across repetitions of the same experiment, we expect an average of 5 per cent false-positive findings.

Univariate parametric t-test ('Univ')

For comparison purposes, we report the analyses of the simulated ERP data (to be explained in the following subsections) with univariate parametric t-tests. This consists in computing the t statistics for each time point and/or derivation and finding the threshold as the usual 95 percentile (p-value<0.05) of the theoretical t- Student distribution with N_s-1 degrees of freedom. Those hypotheses whose statistics have values above the threshold are assigned as the alternative.

Simulation of single channel ERP data

In order to compare the performance of the methods described above, an ERP data for a single derivation was simulated. Individual trials were obtained using the model proposed by Yeung and colleagues, which shows high correspondence with real auditory ERP data [44]. In this model, the ERP signal is obtained by summing two sinusoidal components. The first one corresponds to the negative peak of a 5 Hz sinusoid, with latency 60 ms (N60 peak). The second one corresponds to a slower sinusoid of 1 Hz with positive peak at latency 200 ms. The time window runs between −400 and 400 ms, with a sampling period of 4 ms (250 Hz), for a total of 200 time points.

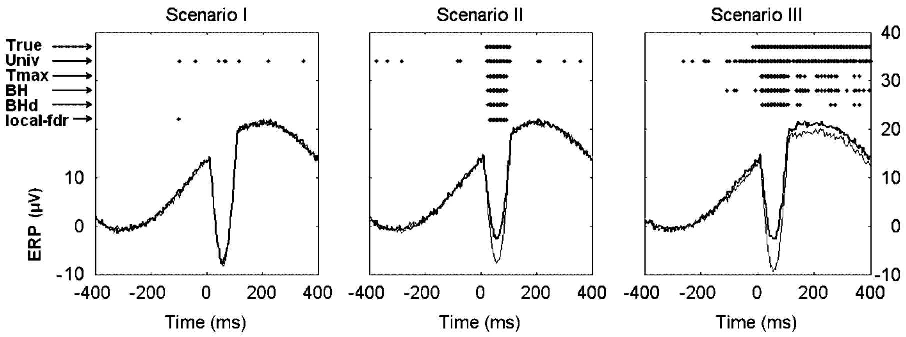

Figure 2. Grand average of simulated ERPs for conditions A (thick line) and B (thin line) and results obtained for a typical run of the single-channel simulation experiment. Left panel corresponds to the Scenario I (null data set). Central panel corresponds to the Scenario II (clear difference in the amplitude of N60 peak) and right panel to the Scenario III (clear difference in the amplitude of N60 peak and small difference in the amplitude of the second peak around 200 ms). In the upper part of each panel, asterisks represent those time points showing significant differences as assessed by each method (shown to the left). The upmost level corresponds to those time points with true significant differences according to the effect size (*Cohen's d* statistic$\geqslant$0.5).

Two different sets of data (experimental conditions) can be simulated by modifying the first and/or the second sinusoidal components. Inter-subject variability can then be added with the objective of obtaining realistic signal-to-noise ratios. For this purpose, a linear regression of each individual ERP versus the corresponding grand mean ERP was performed in all derivations using a real data set (described in the next section). The residuals of this regression correspond to the fraction of the individual potential that is unrelated with the grand mean ERP and was used to assemble a pure noise (zero mean) database. Since the noisy component of a real ERP signal have strong temporal correlations, with this procedure we avoid the use of white noise which would lead to an unrealistic model of ERP data. Finally, the individual potentials are simulated by combining an ERP signal obtained by the Yeung model and a randomly selected noise signal from the database, such that the peak signal-to-noise ratio reaches a predefined value, set to 20 db in all simulations. Baseline correction was performed by subtracting the average of the signal in a pre-stimulus window [-400 ms, -200 ms] to each point of the whole ERP.

Three simulated scenarios were prepared for each of 16 subjects. In all of them, two ERPs for experimental conditions A and B were constructed. Each scenario is defined in the following manner:

Scenario I. Null data set scenario: no differences between experimental conditions are imposed. The amplitudes of the first and second ERP components are the same in both experimental conditions, around $-8\,\mu$V (Figure 2, left).

Scenario II. An evident difference between the amplitudes of the first ERP component (N60 peak) in each condition is imposed. The amplitude of the N60 peak in condition A is around $-3\,\mu$V and in condition B around $-8\,\mu$V. The second ERP component (slow wave of 1 Hz) is the same in both conditions, around 21 μV (Figure 2, center).

Scenario III. An evident difference between the amplitudes of the first ERP component (N60 peak) is imposed, combined with small but systematic differences for the whole time domain, produced by a difference in the amplitudes of the second ERP component (slow wave of 1 Hz). The amplitude of the N60 peak in condition A is around $-3\,\mu$V and in condition B around $-8\,\mu$V, while the amplitude of the second component in condition A is around 21 μV and in condition B around 19 μV (Figure 2, right).

As discussed in the previous sections, the permutation test controls the FWER while the FDR methods controls the *FDR* statistic. Both magnitudes are different except in the case that all hypotheses are true null. Therefore, the null data set simulated in Scenario I is introduced with the objective of making the methods comparable on the basis of the same statistics. On the other hand, the data set simulated in Scenario III allows the study of those cases in which moderate differences prevail in all the time window and larger differences occurs in the N60 peak region, leading to a large proportion of non-null hypotheses. From the mathematical viewpoint, this situation violates the assumption of FDR methods about the predominance of null hypotheses in the data.

For comparative purposes, a precise definition of the time window of true differences between experimental conditions is needed. Since this is not straightforwardly achieved with the simulation procedure described above, the window of true differences is defined by those time points having a substantial effect size. Using the *Cohen's d* statistic [45] as a measure of the effect size, the time points belonging to the window of true differences were selected as those with a *Cohen's d* statistic higher than or equal to 0.5. Then, we estimate the average power and *FDR* values given by each method by applying them to 1000 runs of each scenario. The power in each run is computed as the percentage of true alternative hypotheses detected by the method. The permutation test used $Np=1000$ permuted data sets for every run of the experiment.

Simulation of multichannel ERP data

The multichannel simulation consisted of 19 derivations: 14 of them with signals constructed following the Scenario I and the other 5 with ERPs obtained following the Scenario II. The remaining parameters, (i.e. sampling frequency, time window and number of subjects) are identical to those used in single channel simulations. Noise signals for each derivation were independently and randomly selected from the noise database (see previous section) and added to the simulated ERPs, such that the signal-to-noise ratio reaches 20 db. Using this approach, the spatio-temporal ERP data for each subject is composed of two matrices of 19 derivations x 200 time points, one for each condition. Assuming that a *t*-statistic is appropriate for measuring differences between conditions, the problem results in the thresholding of a *t*-test matrix of 3800 hypotheses (see Figure 3, top left panel). The

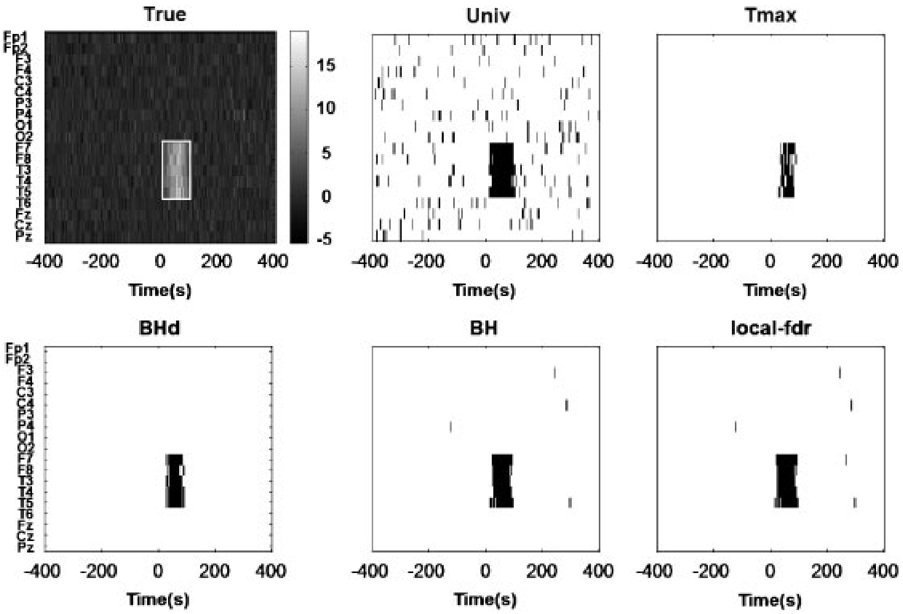

Figure 3. Spatio-temporal maps obtained from the statistical analysis of a typical simulation of multichannel ERPs. Top left: Computed *t*-tests in a grayscale. The white rectangle demarks those space-time points where simulated differences between conditions have a high effect size (*Cohen's d* statistic $\geqslant 0.5$), which are assumed as the true significant tests. In the maps given by each method the significant space/time points are plotted in black. The channel's labels are shown in the *y* axes of the left panels.

increased dimensionality implies a higher risk of type I error in comparison with the 200 simultaneous hypotheses of the single channel problem. Again, the spatio-temporal window with true differences between conditions is defined by those points where the *Cohen's d* statistic is higher than or equal to 0.5. This window is marked in the top left panel of Figure 3 with a white rectangle. We estimate the average power and *FDR* values given by each method by applying them to 1000 runs of simulated multichannel data. The permutation test used $Np = 500$ permuted data sets for every run of the experiment.

Real ERP data

To evaluate the performance of the methods in the analysis of real data, we selected an experiment of face processing. The experimental design consists of an oddball paradigm where the subject is presented with unfamiliar faces as standards and familiar faces as target stimuli. Familiar faces convey different types of information, unlocking memories related to emotional significance. The processing of this information is crucial to adaptability and social behavior [46]. In this experiment, the infrequent targets were classified as positive, negative or neutral depending to the affective valence of the stimuli. The valences of these faces were measured by means of a liker-type scale. ERPs were recorded for each experimental condition, showing a main peak about 300–500 ms after the presentation of the stimulus, which is known as the P300 component. For the methodological objective of this paper, we focused on the comparison between ERP potentials corresponding to positive and neutral faces (Positive and Neutral conditions) for all derivations and all time points, since there are strong evidences about the enhanced amplitude of the P300 for positive vs neutral faces [47]. The electrophysiological responses were recorded in 16 normal subjects using a MEDICID 5 system (Neuronic SA) from 19 channels according to the 10–20 system using disk derivations (Ag/AgCl) referenced to an electrode placed on the nose. Other details about this data can be found in [48].

Results

Single channel simulation

Table I shows the results for the three scenarios studied in the single channel simulation. For the null data set (Scenario I) the comparison among the methods should be done through the analysis of the FWER, since in this case the permutation test and the FDR methods control the same statistics, i.e. $FDR = \text{FWER}$. In this scenario the power is zero by definition, since there cannot be any type II error, i.e. none of the hypotheses can be mistakenly classified as null since all of them are truly null hypotheses. For Scenario II and III, we report the average *FDR* and power across 1000 repetitions.

In Scenario I, the values of FWER given by the permutation test and BH method were close to the theoretically expected value of 0.05, given that we used $\alpha = 0.05$ for $T_{\max}$ and $q = 0.05$ for FDR methods. The univariate test gave at least one false discovery in every repetition (FWER = 1), which means that these methods are usually too permissive. On the contrary, the dependent variant of the BH method behaves more conservative (lowest FWER) and the local-fdr algorithm reaches an intermediate performance. Scenario II represents the situation for which FDR methods are designed to be used. The BH method (assuming independent tests) gives the best balance between type I error and type II error, followed by the local-fdr with reasonable *FDR* (lower than 5 per cent)

Table I. Statistical analysis of single channel simulated ERPs.

	Scenario I	Scenario II		Scenario III	
Method	FWER=*FDR*	*FDR*	Power (per cent)	*FDR*	Power (per cent)
Univ ($p<0.05$)	1.000	0.303	97.3	0.144	88.8
$T_{\max}$ ($\alpha=0.05$)	0.046*	0.003	79.1	0.005	33.1
BH ($q=0.05$)	0.042*	0.053*	91.3*	0.097*	80.8*
BHd ($q=0.05$)	0.002	0.008	84.4	0.024	51.9
Local-fdr ($q=0.05$)	0.073	0.026*	88.1*	0.001	1.0

For the null data set (Scenario I) the FWER statistic coincides with the *FDR*. The average values of *FDR* and power across 1000 runs of each experiment are reported. Asterisks highlight the best results.

Table II. Statistical analysis of multichannel simulated ERPs.

Method	*FDR*	Power (per cent)
Univ ($p<0.05$)	0.647	95.3
$T_{\max}$ ($\alpha=0.05$)	0.001	49.0
BH ($q=0.05$)	0.050*	79.5*
BHd ($q=0.05$)	0.005	64.6
Local-fdr ($q=0.05$)	0.045*	78.8*

The average values of *FDR* and power across 1000 runs of the experiment are reported. Asterisks highlight the best results. The average *FDR* and power for BH and local-fdr methods are equivalent, since their variances were 0.02 and 3 per cent, respectively.

while keeping the power near 90 per cent. The univariate test is the most powerful but at the cost of an unacceptable rate of false discoveries of around 30 per cent of the hypotheses. On the other edge, the permutation test is the most conservative, giving the lowest *FDR* at a cost of a power lower than 80 per cent. In Scenario III, we obtained the same picture: the univariate test is the most permissive and the $T_{\max}$ the most conservative, while the BH method gives the best balance between *FDR* and power. The main difference appears in the performance of the local-fdr method, which is very bad in this last scenario. This is explained by the fact that in this scenario the number of null hypotheses is not predominant, so it is more difficult to estimate the null distribution from the histogram of statistics.

Figure 2 shows a typical run for the three scenarios studied in the single channel simulation. The marked points locate those time points with significant difference between experimental conditions as assessed by each method. This figure shows a particular example of the general behavior offered in Table I.

Multichannel simulation

Table II shows the average *FDR* and power given by the application of each method to 1000 multichannel simulated ERP data sets. Similar to the single channel case, the univariate test was the most powerful method, but at the cost of the highest rate of type I error. Also, the permutation test was the most conservative with very low *FDR* but offering a very poor power. In this case, the BH method, assuming independent tests, and the local-fdr offer the best balance between *FDR* and power. However, both variants of the BH method show lower power and higher *FDR* in this case with respect to the single channel simulation (Scenario II), which is a known consequence of the increased dimensionality of the problem. At the same time, this fact is beneficial for the estimation of empirical probability densities carried out in the local-fdr algorithm, thus leading to a better performance of this method for high-dimensional problems.

Figure 3 shows the spatio-temporal binary maps of the significant differences between experimental conditions as assessed by each method for a typical multichannel simulation. The top left panel of Figure 3 presents the spatio-temporal map of computed t-tests, where a white rectangle demarks those space/time points with true significant differences (as assessed by those points with *Cohen's d* statistic $\geqslant 0.5$). This example shows the conservative nature of $T_{\max}$ and BHd, while the univariate test is too permissive. BH and local-fdr correctly locate most of the true significant points, although there are around 5 per cent of false-positive discoveries out of the true window.

Real data analysis

The ERPs associated with 'neutral' and 'positive' experimental conditions showed the occurrence of a positive peak, maximal over centro-parietal sites, around 500 ms, that is known as the P300 component. This component has been shown to have enhanced amplitudes in response to positive faces as compared with neutral faces, which is easily seen in the ERPs in derivation Pz, shown in the Figure 4, top left panel. In our analysis, the differences in the amplitude of the P300 are detected with dissimilar

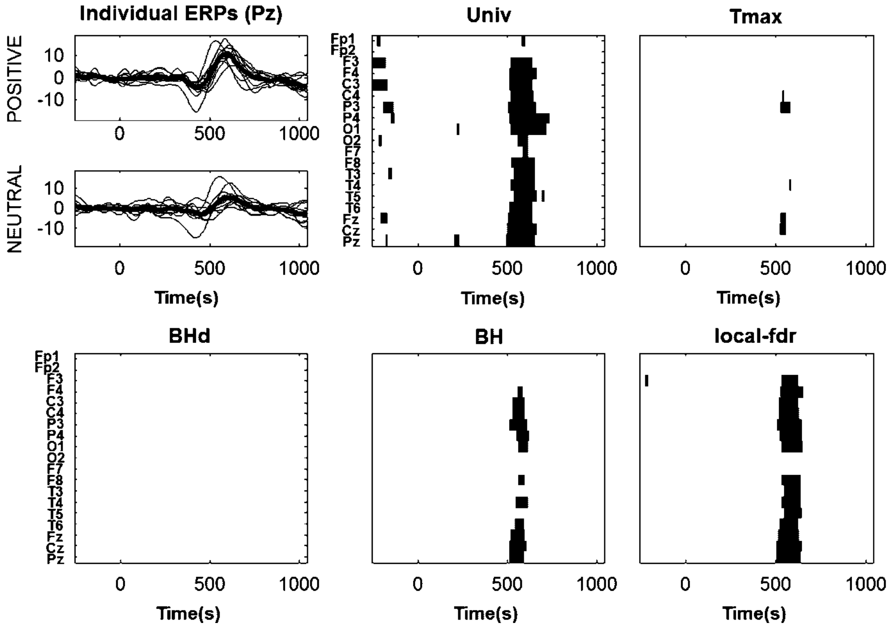

Figure 4. Spatio-temporal maps obtained from the statistical analysis of the real ERP data set. Top left: Individual ERPs for experimental conditions corresponding to the presentation of neutral faces and faces with affective valence (positive). In the maps given by each method the significant space/time points are plotted in black. The channel's labels are shown in the *y* axes of the left panels. The BHd method did not show any significant point.

spatio-temporal location by each method, as shown in Figure 4. In accordance with the results obtained for the simulated data sets, the univariate *t*-tests (requiring $p<0.05$) showed significant differences for almost all derivations around 500 ms but also for other latencies, mostly in the pre-stimulus period, which are clearly not related with the experiment and suggests a large number of false positives. Thus the method is not suitable for the analysis of the data. The permutation test and the BH method assuming dependent tests (BHd) are too restrictive, the former showing only a few derivations with significant differences (not including Pz). The latter did not show any significant point, which agrees with its behavior in the multichannel data where the increased dimensionality affected the power of the method. Finally, the BH method for independent tests and the local-fdr offer more reasonable detection of significant points, covering a smaller subset of derivations and a thinner latencies interval.

An insightful comparison of the thresholds selected by the methods in this specific run of the experiment can be done in terms of the different null distribution used by each of them. Figure 5 (top left panel) shows the histogram of the computed *t*-statistics and the empirical probability densities (mixture and null) estimated by the local-fdr method. The estimated threshold (dashed line) clearly detects the right flat tail of the empirical distribution as formed by relevant statistics. The threshold was selected as the value corresponding to the estimated global *FDR* of 0.05 (q-value). This is shown in detail in the bottom left panel of Figure 5, where both global and local *FDR* statistics, estimated in the local-fdr procedure, are plotted. Note that this threshold corresponds to a value of *local fdr* of about 0.2, stating the conceptual difference between both statistics.

The top right panel of figure 5 shows the histograms of the statistics under the null hypothesis $\tilde{z}_{ij}$ and of their maxima as obtained by the permutation test (see section 'Permutation test' above). The threshold (dashed line) is selected as the 95 percentile of the distribution of the maximum. Since the BH methods and the univariate test use the *t*-Student distribution as the theoretical null, this distribution is shown in the bottom right panel of Figure 5 with a thin solid line. The thresholds selected by these methods are also shown (dashed lines). The BH method assuming dependent tests was extremely conservative, to the point that no test was classified as significant (threshold equals 7.40, not shown). Among the other methods, the permutation test offered the most restrictive (highest) threshold, while the univariate test is the most permissive. Bottom right panel of Figure 5 also shows the empirical null probability densities estimated by the permutation test (dotted line) and local-fdr (dot-dashed line). The former is almost identical to the theoretical null and the latter is wider and with negative mode.

Discussion

The comparison of the performance of the different methods in simulated and real data uncovers several issues that we will discuss separately. First, the methods offer different theoretical advantages and disadvantages. The univariate test is not appropriate for the analysis of ERP data since, despite showing the highest power, it gives an unacceptable number of type I errors. This was expected since this method does not cope with the problem of multiple comparisons, while the permutation test and the FDR methods are designed for this purpose. The permutation test has been claimed to have several advantages over parametric methods for the ERP analysis: the tests are distribution free, no assumptions of an underlying correlation structure are required, and they provide exact *p*-values for any number of subjects, time points and channels [5]. Some of these advantages are shared

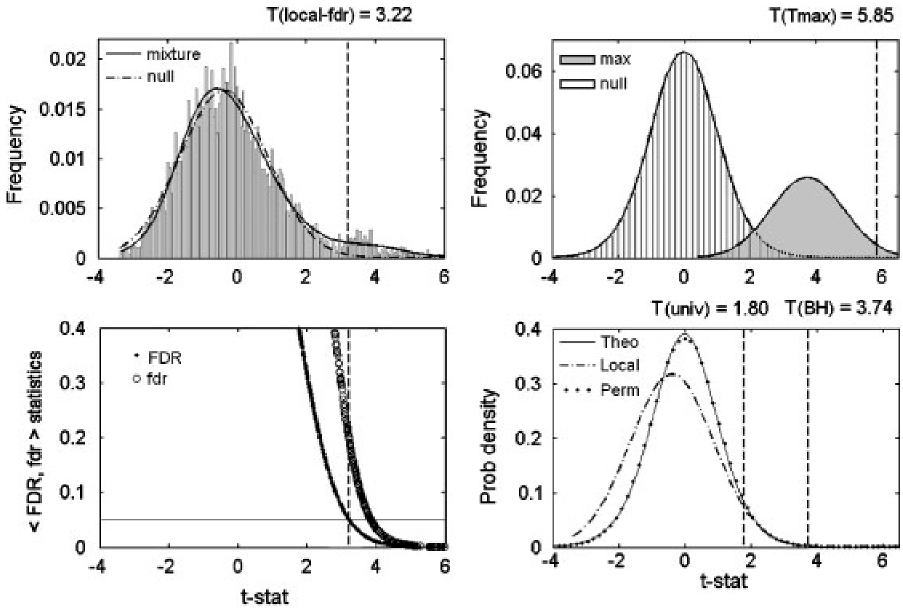

Figure 5. Empirical distributions and thresholds for the analysis of the real data set. Top left: Histogram of all *t*-tests, estimated mixture density (solid line) and null density (dot-dashed line). The vertical dashed line represents the threshold selected by the procedure. Bottom left: Plots of the local and global *FDR* statistics (circles and dots, respectively) around the selected threshold, as obtained with the local-fdr algorithm. This corresponds to a required *FDR* of 0.05, but to a local *FDR* of 0.2. Top right: histograms of all spatio-temporal *t*-statistics (white) and their maximum value (grey) for each permuted data set, as obtained by the permutation test. The threshold, selected as the 95 percentile of the maximum distribution ($\alpha = 0.05$) is shown as a vertical dashed line. Bottom right: Theoretical null (thin solid line) and empirical null distributions given by permutation test (dotted line) and local-fdr (dot-dashed line). The theoretical null (*t*-student distribution with 15 degrees of freedom) is used by the univariate test and BH methods for selecting the corresponding thresholds (vertical dashed lines). The BH method for dependent tests (BHd) did not show any significant test, estimating a very high threshold of 7.40 (not shown).

by the local-fdr procedure, which does not require any assumption on the distribution of the statistics or on the correlation structure among them. However, the local-fdr also implies fitting histograms, making it sensitive to the choice of fitting methods, given that in real neuroscience data the histogram of the tests can be very noisy.

Second, the methods showed differences in performance considering the control of type I and type II errors. Here, the most important issue is that the permutation test and the FDR methods control different measures of the type I error, namely, the FWER and the *FDR* statistics, respectively. Therefore, the permutation test exerts a stronger control of the type I error, which leads it to be more conservative. However, the analysis of the simulated null data set, allows a comparison of the methods since in this case the FWER coincides with the *FDR* statistic (see Table I). In this case, the permutation test and the BH method for independent tests controlled the FWER at almost exactly the theoretical value required, demonstrating the reliability of the method for this purpose. The local-fdr and BH method for dependent tests over and under controlled the type I error, respectively.

In the analysis of simulated non-null data, the BH method (independent tests) showed the best balance in the control of the type I and type II errors, as measured by the *FDR* (with expected value of 0.05) and the power. This method assumes a theoretical distribution for the tests, thus the fact that the statistics computed in the simulations do conform to the assumed distribution (*t*-Student) might play a role in this good overall performance. Although the lower *FDR* given by the permutation test was expected (since controlling the FWER is stricter than controlling the *FDR*), this method showed a relevant loss in power. As the maximum statistic used in the current permutation test is too severe, one might use other maximal statistic (e.g. 99 or 95 percentile of the tests) or a variant of the method such as the step-down procedure proposed by Holmes [19]. This method consists in repetitions of the permutation procedure, excluding from the sampling distribution, at each step, those hypotheses rejected in the previous step. The rationale behind this method is to protect the permutation distribution from the influence of relabellings that are close to the observed data, whose maximal statistics are dominated by true alternative hypotheses. However, Holmes showed that this procedure does not provide relevant improvements in the analysis of PET data, with the drawback of a higher computational effort [19]. Here, the step-down procedure was tested in the multichannel simulation (data not shown) obtaining no differences with the single step permutation test. Therefore, we might say that in the spatio-temporal analysis of ERP data, the step-down procedure does not provide additional advantages with respect to the permutation test used here. Future ERP experiments should be designed to explore less stringent permutation methods.

The local-fdr method showed a good balance between *FDR* and power only in the multichannel simulation where the histograms can be better modeled (see Table II). A very important requirement for the application of the local-fdr is the prevalence of null hypotheses in the whole set to be tested. Scenario III showed that when this condition is violated, the local-fdr method undergoes unreliable estimates due to badly fitted probability densities and we do not recommend its use. Other differences between local and global FDR methods will be discussed later on.

It is important to stress that the methods compared here do not use the same criteria or parameter for selecting the thresholds. The univariate test selects the threshold corresponding to a particular *p*-value ($1-p$ percentile of the theoretical null distribution), the permutation test uses α ($1-\alpha$ percentile of the empirical distribution of the maximum of permuted statistics) and the FDR methods use a predefined q-value, which corresponds to an upper bound of the *FDR* statistic. In this work, we used a p-value of 0.05 for the univariate test, which means that the probability of classifying a hypothesis as significant under the null hypothesis is lower than 5 percent. In the permutation procedure we set $\alpha=0.05$ (controls the FWER at level of 5 per cent), meaning that across 100 repetitions of the same experiment we expect that 5 of them show at least one false-positive finding. Setting $q=0.05$ in the FDR methods means that across repetitions of the same experiment the proportion of false discoveries is, on an average, 5 per cent. Despite this, we consider the comparisons performed in this work to be valid, since our interest is to study their performance in the practical analysis of ERP data, in which all these methods would be applied with the same paradigm followed here.

Third, the methods are based on different null distributions. The univariate test and the BH method are based on the theoretical distribution of the statistics (in our case the t-Student distribution). The permutation test finds an empirical null distribution based on the permuted statistics and the local-fdr estimates an empirical null distribution from the modeling of the histogram of actual statistics (after a z-scaling defined in [27]) in a region around its maximum peak. On the one hand, the use of the empirical null estimated by local-fdr, against the theoretical null centered at zero, is based on the attempt of discovering those 'interesting' hypothesis instead of those hypotheses 'significantly different from 0'. The term 'interesting', introduced by Efron [27], refers to those extreme tests that are distant from the principal tendency of the data, regardless if this principal tendency matches the theoretical null with zero mean. As a consequence, it is possible to find situations where the statistical significance is not equivalent to the discovery of a relevant finding. In neuroscience, the correct assessment of the proper null hypothesis is a critical issue. For instance, in cognitive processes, where the 'interesting' features are known to be sparse and spatially or temporally well located (such as ERP, or the determination of regions of interest in functional neuroimaging), it is appropriate to consider most of the hypotheses as 'non-interesting' and make inference based on the observed empirical null.

On the other hand, the empirical null distribution used by the permutation test receives the influence of temporal and spatial correlations. In our real data, the empirical null distribution using permutation test was nearly identical to the theoretical null. However, the empirical null estimated by local-fdr is wider than the theoretical null and the mode is slightly moved to the negative values (see Figures 5, bottom right panel). Efron exposed how the presence of unobserved covariates is likely to widen the observed empirical null distribution and the impossibility of permutations test to detect this effect [27]. Unraveling these covariates in ERP data sets presents an interesting challenge for future works. These findings suggest being cautious in the use of the theoretical null to analyze spatio-temporal correlated tests.

It is also important to mention that the computational complexity of the permutation test is considerably higher than that of the FDR methods and depends on the number of permuted data sets used for obtaining the empirical null. In our analyses of multichannel data, for instance, permutation test took about 15 s on an average, when using 500 permuted data sets. The local-fdr method took about 0.5 s and the BH method around 1 ms. This means that FDR methods are preferable in those cases in which they perform similar to the permutation test.

Finally, our study revealed some important differences between the two FDR methods introduced here. The most important difference being the dependence of the methods with the dimensionality of the problem analyzed. We found that the recently proposed local-fdr is not the most appropriate method for the analysis of single channel ERP data. However, it offered a good balance between type I error and power in high-dimensional (multichannel) problems (see Tables I and II). The BH method suffered a severe reduction in power in multichannel situations. This effect is worse if the correlations between tests are taken into account, which is a consequence of the very stringent correction imposed in the BHd algorithm (see a in equation (4)) and explains why the version of the BH method for independent tests is more common than the corrected version in neuroimaging studies [23, 35].

The reason of these dissimilar results is that in single channel problems (200 hypotheses) the risk of type I error is much smaller than in multichannel problems (4940 hypotheses). On one hand, the BH method follows the logic of finding the largest subset of alternative hypotheses, thus it is affected when the number of tests increases, obtaining conservative estimates and suffering a clear loss of power directly proportional to the number of hypotheses tested [41]. Given that this method assumes a uniform distribution of p-values, it is not sensitive to large changes in the histogram of the tests, which allows it to perform better than the local-fdr in low-dimensional (single-channel) problems. On the other hand, the local-fdr algorithm provides good estimations in multichannel problems since testing more hypotheses leads to more reliable fitting of the densities and to a higher probability of the histogram to be dominated by null hypotheses.

In this sense, a disadvantage of the local-fdr is that it strongly depends on the selection of the null distribution. Several transformations of the tests and *ad hoc* selection of the null region are used in the original algorithm to ensure a proper estimation of the null density, although other variants of this method have been developed to cope with this problem [32–34]. However, an advantage of the local-fdr over the BH method is that it does not require the assumption of a particular theoretical distribution for the tests, which is important in the analysis of non-gaussian data or when the statistics do not conform to any of the well-known distributions (t-Student, chi-squared, F-statistic). Other variants of the local-fdr, perhaps specifically developed for the case of ERP analysis, should be taken into account for future studies on the usefulness of this method in the field.

A final relevant issue concerns the implications of the differences between the global and local *FDR* statistics. The mathematical relationship between these statistics (equation (3)) implies that the global FDR results in a lower bound of the local FDR. This can be practically seen in the bottom left panel of Figure 5, where the curve of the *FDR* statistic is always below that of the local *FDR* statistic. This means that selecting a threshold that corresponds to a predefined cutoff value of the local FDR (as proposed originally by Efron [27]) guarantees a corresponding value of the global FDR not higher than the predefined cutoff. In the real data example presented here, the threshold obtained requiring a q-value of 0.05 (global FDR) corresponds to a local FDR of 0.2.

If we select the threshold corresponding to a local FDR of 0.05, the global *FDR* statistic will yield a value of 0.015, i.e. we would be expecting a much lower ratio of false discoveries.

Conclusions

In this article we have introduced two FDR techniques that control the proportion of false positives findings in the field of ERP data analysis: the BH method and the local-fdr procedure. Using simulated and real data we compared their performance with that of the permutation test and the univariate (uncorrected) threshold. We found that the exploration of ERP data sets in the whole spatio-temporal domain is an advantageous strategy that can be managed by the statistical tools used here to cope with the problem of multiple comparisons: the permutation test and the FDR methods. In general, this would avoid the *ad hoc* selection of channels and/or time windows of interest when searching for differences between ERPs of two experimental conditions, allowing exploratory studies when there is not previous information about where the interesting ERP components should appear.

The comparison showed that the permutation test exerts a stronger control of the type I error than the FDR methods, but at the cost of a lower power. The computational complexity of this method is also higher than that of the FDR methods and the estimation usually takes 10 times longer. Therefore, the FDR methods introduced here showed to be a good alternative to deal with the multiple comparisons in the analysis of ERP data. They presented important differences that should be taken into account in practical applications: (a) the BH method assuming independent tests gave the best overall performance in terms of offering a balance between type I and type II errors. The BH method assuming dependent tests is too conservative in most of the cases and both BH variants suffer a decrease in power when the dimensionality of the problem increase; (b) the local-fdr method is preferable for high-dimensional (multichannel) problems where most of the tests to be compared conforms to the empirical null hypothesis, since it performs similar to the BH method but does not require the assumption of a theoretical distribution for the tests. The simple implementation of this method allows the development of diverse variants that help to make good estimations of the empirical null and the mixture densities, which is crucial for its good performance.

Acknowledgements

The authors acknowledge Ileana Quiñones and María A. Bobes, from the Cognitive Neuroscience Department of the Cuban Neuroscience Center, for providing the real data used in this paper and for their help in the description of the data and revision of the manuscript.

References

1. Fisher RA. *The Design of Experiments* (1st edn). Oliver and Boyd: Edinburgh, 1935.
2. Handy T. *Event-Related Potentials*. The MIT Press: Cambridge, 2004.
3. Goodman S. Toward evidence-based medical statistics. 1: the p value fallacy. *Annals of Internal Medicine* 1999; **130**(12):995–1004.
4. Bobes MA, Lopera F, Díaz-Comas L, Galan L, Carbonell F, Bringas ML, Valdés-Sosa M. Brain potentials reflect residual face processing in a case of prosopagnosia. *Cognitive Neurophysiology* 2004; **21**(7):691–718. DOI: 10.1080/0264329034200025.
5. Bobes MA, Quiñones I, Perez J, Leon I, Valdes-Sosa M. Brain potentials reflect access to visual and emotional memories for faces. *Biological Psychology* 2007; **75**:146–153. DOI: 10.1016/j.biopsycho.2007.01.006.
6. Jennings JR, Cohen MJ, Ruchkin DS, Fridlund AJ. Editorial policy on analysis of variance with repeated measures. *Psychophysiology* 1987; **24**(4):474–478. DOI: 10.1111/j.1469-8986.1987.tb00320.x.
7. Kiebel SJ, Friston K. Statistical parametric mapping for event-related potentials: 1 generic considerations. *Neuroimage* 2004; **22**:492–502. DOI: 10.1016/j.neuroimage.2004.02.012.
8. Carbonell F, Galan L, Valdes P, Worsley K, Biscay RJ, Diaz-Comas L, Bobes MA, Parra M. Random field-union intersection tests for EEG/MEG imaging. *NeuroImage* 2004; **22**:268–276. DOI: 10.1016/j.neuroimage.2004.01.020.
9. Picton T, Bentin S, Berg P, Donchin E, Hillyard S, Johnson R, Miller G, Ritter W, Ruchkin D, Rugg M, Taylor M. Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *Psychophysiology* 2000; **37**(2):127–152. DOI: 10.1111/1469-8986.3720127.
10. Goodman S. Multiple comparison, explained. *American Journal of Epidemiology* 1998; **147**:807–812. DOI: 10.1080/02643290342000258.
11. Yandell BS. *Practical Data Analysis for Designed Experiments* (1st edn). Chapman & Hall: London, 1997.
12. Worsley KJ, Marett S, Neelin P, Vandal A, Friston K, Evans A. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping* 1996; **4**:58–73. DOI: 10.1002/(SICI)1097-0193(1996)4.
13. Bosch-Bayard JJ, Valdés-Sosa PA, Virues-Alba T, Aubert-Vázquez E, John E, Harmony T, Riera-Díaz J, Trujillo-Barreto N. 3D statistical parametric mapping of EEG source spectra by means of variable resolution electromagnetic tomography (VARETA). *Clinical Electroencephalography* 2001; **32**:47–61.
14. Kilner JM, Kiebel SJ, Friston KJ. Applications of random field theory to electrophysiology. *Neuroscience Letters* 2005; **374**:174–178. DOI: 10.1016/j.neulet.2004.10.052.
15. Edgington ES. *Statistical Inference*: *The Distribution Free Approach*. McGraw-Hill: New York, 1969.
16. Raz J. Analysis of repeated measurements using nonparametric smoothers and randomization tests. *Biometrics* 1989; **45**:851–871.
17. Westfall PH, Young SS. *Resampling-Based Multiple Testing*: *Examples and Methods for p-Value Adjustment*. Wiley: New York, 1993.

18. Blair RC, Karniski W. An alternative method for significance testing of waveform difference potential. *Psychophysiology* 1993; **30**:518–524. DOI: 10.1111/j.1469-8986.1993.tb02075.x.
19. Holmes AP, Blair RC, Watson JDG, Ford I. Non-parametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow and Metabolism* 1996; **16**:7–22. DOI: 10.1097/00004647-199601000-00002.
20. Galán L, Biscay R, Rodríguez JL, Ábalo C, Rodríguez R. Testing topographic differences between event related brain potentials by using non-parametric combinations of permutation tests. *Electroencephalography and Clinical Neurophysiology* 1997; **102**:240–247. DOI: 10.1016/S0013-4694(96)95155-3.
21. Manly BFJ. *Randomization, Bootstrap, and Monte-Carlo Methods in Biology*. Chapman & Hall: London, 1997.
22. Pesarin F. *Multivariate Permutation Tests*. Wiley: New York, 2001.
23. Nichols T, Holmes A. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping* 2001; **15**(1):1–25. DOI: 10.1002/hbm.1058.
24. Bobes MA, Lopera F. Covert matching of unfamiliar faces in a case of prosopagnosia: an erp study. *Cortex* 2003; **39**(1):41–56. DOI: 10.1080/02643290342000258.
25. Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* 2007; **164**(1):177–190. DOI: 10.1016/j.jneumeth.2007.03.024.
26. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 1995; **57**(1):289–300.
27. Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of American Statistical Association* 2004; **99**:96–104. DOI: 10.1198/016214504000000089.
28. Ge Y, Sealfon SC, Tseng CH, Speed TP. A Holm-type procedure controlling the false discovery rate. *Statistics and Probability Letters* 2007; **77**(12):1756–1762. DOI: 10.1016/j.spl.2007.04.019.
29. Langers D, Jansen J, Backes W. Enhanced signal detection in neuroimaging by means of regional control of the global false discovery rate. *NeuroImage* 2007; **38**(1):43–56. DOI: 10.1016/j.neuroimage.2007.07.031.
30. Finos L, Salmaso L. FDR- and FWE-controlling methods using data-driven weights. *Journal of Statistical Planning and Inference* 2007; **137**(12):3859–3870. DOI: 10.1016/j.jspi.2007.04.004.
31. Ploner A, Calza S, Gusnanto A, Pawitan Y. Multidimensional local false discovery rate for microarray studies. *Bioinformatics* 2006; **22**(5):556–565. DOI: 10.1093/bioinformatics/btk013.
32. Schwartzman A. Empirical null and false discovery rate inference for exponential families. *Annals of Applied Statistics* 2008; **2**(4):1332–1359. DOI: 10.1214/08-AOAS184.
33. Robin S, Avner BH, Daudin JJ, Pierre L. A semi-parametric approach for mixture models: application to local false discovery rate estimation. *Computational Statistics and Data Analysis* 2007; **51**(12):5483–5493. DOI: 10.1016/j.csda.2007.02.028.
34. Xie Y, Pan W, Khodursky AB. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics* 2005; **21**(23):4280–4288. DOI: 10.1093/bioinformatics/bti685.
35. Genovese C, Lazar N, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 2002; **15**(N 1053-8119):870–878. DOI: 10.1006/nimg.2001.1037.
36. Sánchez-Bornot J, Martínez-Montes E, Lage-Castellanos A, Vega M, Valdés P. Uncovering sparse brain effective connectivity: a voxel-based approach using penalized regression. *Statistica Sinica* 2008; **18**(4):1501–1518.
37. Marchini J, Presanis A. Comparing methods of analyzing fMRI statistical parametric maps. *NeuroImage* 2004; **22**:1203–1213. DOI: 10.1016/j.neuroimage.2004.03.030.
38. Martínez-Montes E, Cuspineda E, El-Deredy W, Sánchez-Bornot J, Lage-Castellanos A, Valdés P. Exploring event-related brain dynamics with tests on complex valued time-frequency representations. *Statistics in Medicine* 2007; **27**(15):2922–2947. DOI: 10.1002/sim.3132.
39. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 2001; **29**(4):1165–1188. DOI: 10.1214/aos/1013699998.
40. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B* 2002; **64**:479–498. DOI: 10.1111/1467-9868.00346.
41. Qian HR, Huang S. Comparison of false discovery rate methods in identifying genes with differential expression. *Genomics* 2005; **86**(4):495–503. DOI: 10.1016/j.ygeno.2005.06.007.
42. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* 2003; **31**(6):2013–2035. DOI: 10.1214/aos/1074290335.
43. Chumbley RJ, Friston KJ. False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage* 2009; **44**(1):62–70. DOI: 10.1016/j.neuroimage.2008.05.021.
44. Yeung N, Bogacz R, Holroyd C, Cohen J. Detection of synchronized oscillations in the electroencephalogram: an evaluation of methods. *Psychophysiology* 2004; **41**:822–832. DOI: 10.1111/j.1469-8986.2004.00239.x.
45. Cohen J. *Statistical Power Analysis for the Behavioral Sciences* (2nd edn). Lawrence Erlbaum: NJ, 1988.
46. Delplanque S, Silvert L, Hot P, Rigoulot S, Sequeira H. Arousal and valence effects on event-related P3a and P3b during emotional categorization. *International Journal of Psychophysiology* 2006; **60**(3):315–322. DOI: 10.1016/j.ijpsycho.2005.06.006.
47. Lewis PA, Critchley HD, Rotshtein P, Dolan RJ. Neural correlates of processing valence and arousal in affective words. *Cerebral Cortex* 2007; **17**(3):742–748. DOI: 10.1093/cercor/bhk024.
48. Fernández A. Factores que modulan el componente positivo temprano frontal asociados a los estímulos de alto valor emocional. *Master Thesis*, Cuban Neuroscience Center, Havana, 2009.