ELSEVIER

Journal of Neuroscience Methods 142 (2005) 209–217

JOURNAL OF NEUROSCIENCE METHODS

www.elsevier.com/locate/jneumeth

New concepts of multiple tests and their use for evaluating high-dimensional EEG data

Claudia Hemmelmann[a,*], Manfred Horn[a], Thomas Süsse[a], Rüdiger Vollandt[a], Sabine Weiss[b,c]

[a] *Institute of Medical Statistics, Computer Sciences and Documentation, University of Jena, D-07740 Jena, Germany*
[b] *SFB 360, University of Bielefeld, Germany*
[c] *Brain Research Institute, Cognitive Neuroscience Group, Medical University of Vienna, Austria*

Received 30 April 2004; received in revised form 12 August 2004; accepted 18 August 2004

Abstract

Recently, new concepts of type I error control in multiple comparisons have been proposed, in addition to FWE and FDR control. We introduce these criteria and investigate in simulations how the powers of corresponding test procedures for multiple endpoints depend on various quantities such as number and correlation of endpoints, percentage of false hypotheses, etc. We applied the different multiple tests to EEG coherence data. We compared the memory encoding of subsequently recalled and not recalled nouns. The results show that subsequently recalled nouns elicited significantly higher coherence than not recalled ones.

Keywords: Coherence; EEG data; FDR control; FWE control; Multiple endpoints; Multiple test; Permutation test; Power of test

1. Introduction

Modern procedures of EEG analysis yield large sets of high-dimensional parameters, which have to be evaluated statistically. Let k denote the dimension of the observations. This means there are k components, which are also called multiple endpoints. In Hemmelmann et al. (2004) we dealt with so-called global tests or multivariate tests which provide one joint statement on all k endpoints. We now consider procedures that provide a statement for each endpoint. Many authors use an α-level test for each single component or endpoint of the observational vector, see e.g. Rappelsberger and Petsche (1988). However, this practice results in a large number of false positive statements (false discoveries, type I errors). There exist several techniques to cope with this general drawback in multiple comparisons. Corresponding multiple tests will be considered in the present paper.

Our paper has the following aims: (a) to introduce both traditional and recently proposed concepts of error control in multiple comparisons, (b) to investigate corresponding multiple test procedures regarding their dependence on the dimension k, the fraction of false hypotheses and the correlation structure of the data and compare the powers of different methods, and (c) to demonstrate the use of different multiple tests in problems of multiple comparisons of coherence values obtained from EEG data recorded during the memory encoding of subsequently recalled or not recalled abstract nouns (Weiss et al., 2000).

The techniques we discuss are not specific to EEG data; they are equally applicable to the large data in MEG and fMRI.

2. Methods

2.1. Multiple tests and type I error control

As explained in Section 1, our observations are vectors of dimension k. Assume we have to compare paired samples or two independent samples. Let $x = (x_1, \ldots, x_k)$ and $y = (y_1, \ldots, y_k)$ denote the corresponding random vec-

* Corresponding author. Tel.: +49 3641 9 33610; fax: +49 3641 9 33200.
E-mail address: hemmel@imsid.uni-jena.de (C. Hemmelmann).

0165-0270/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jneumeth.2004.08.008

Table 1
List of methods considered in this paper

Method	Abbreviation	Requirement
Bonferroni method	Bonf	
Step-down procedure of Holm (1979)	Holm	$P(V > 0) \leq \alpha$
Step-down procedure of Troendle (1995)	Troe	
Step-up procedure of Benjamini and Hochberg (1995)	BH	
Two-stage procedure of Benjamini et al. (2001)	BKY	$E(Q) \leq \alpha$
Step-down procedure of Benjamini and Liu (1999)	BL99	
Step-down procedure of Benjamini and Liu (2001)	BL01	
Step-down procedure A of Korn et al. (in press)	PrA_u	$P(V > u) \leq \alpha$ $(0 \leq u < k)$
Step-down procedure B of Korn et al. (in press)	PrB_γ	$P(Q > \gamma) \leq \alpha$ $(0 < \gamma < 1)$

tors and $(\mu_{x_1}, \ldots, \mu_{x_k})$ and $(\mu_{y_1}, \ldots, \mu_{y_k})$ the respective means. Then, the individual null hypotheses to be tested are $H_1 : \mu_{x_1} = \mu_{y_1}, \ldots, H_k : \mu_{x_k} = \mu_{y_k}$. Tests for $H_1, \ldots, H_k$ are called multiple tests.

It can happen that one of the k hypotheses, say H_i, is rejected though it is true. Such an event is called type I error or false discovery. Let R denote the random number of rejected hypotheses and V the random number of rejected true hypotheses, i.e. type I errors ($V \leq R \leq k$). An interesting quantity is the fraction V/R of falsely rejected hypotheses. As this is not defined for $R = 0$, we introduce a new random variable Q where $Q = V/R$ if $R > 0$ and $Q = 0$ if $R = 0$. In the literature, Q is called *false discovery proportion* whereas the expectation $E(Q)$ is called *false discovery rate* (FDR). Different concepts of controlling the proportion or the number of false discoveries have been proposed, together with corresponding methods which control these rates in multiple testing problems. In the next sections, we will present and compare the test procedures listed in Table 1 which satisfy four different criteria that are defined by the requirements given in the last column.

2.1.1. Control of the FWE

As already mentioned in Section 1, it is not advisable to use an α-level test for each of the k hypotheses, i.e. a test that rejects a true hypothesis with probability α, because in this case the expected number of false discoveries $E(V)$ may be rather large; in the worst case, it can be as high as $k\alpha$. And the probability FWE = $P(V > 0)$, i.e. the probability of committing at least one type I error may also be very large, especially when k is large. FWE is the abbreviation of the term *familywise error rate*. For multiple comparisons it has been long recommended to use test procedures that control the FWE, i.e. that guarantee that FWE $\leq \alpha$ no matter how many and which hypotheses are true. α is a prespecified small probability.

The simplest way to ensure that FWE $\leq \alpha$ is to test all individual hypotheses $H_1, \ldots, H_k$ at level α/k. This is the well-known Bonferroni method (Bonf). Assume we use some parametric or nonparametric test appropriate for $H_1, \ldots, H_k$, e.g. the t-test and obtain the p-values $p_1, \ldots, p_k$. Then, Bonf rejects H_i if $p_i \leq \alpha/k$.

Another simple method is the step-down procedure of Holm (1979) (Holm). Let $p_{(1)} \leq \cdots \leq p_{(k)}$ denote the ordered p_i and $H_{(1)}, \ldots, H_{(k)}$ the corresponding hypotheses. In the first step, $p_{(1)}$ is compared with α/k. If $p_{(1)} > \alpha/k$, none of the k hypotheses will be rejected and the procedure stops. If $p_{(1)} \leq \alpha/k$, $H_{(1)}$ is rejected. Then $p_{(2)}$ is compared with $\alpha/(k-1)$. If $p_{(2)} > \alpha/(k-1)$, $H_{(2)}, \ldots, H_{(k)}$ are accepted. Otherwise $H_{(2)}$ will be rejected, etc. Clearly, Holm rejects at least all hypotheses that are rejected by Bonf. This means, Holm is more powerful.

Bonf and Holm do not take into consideration the correlation between the endpoints. In contrast, the step-down method of Troendle (1995) (Troe) is adaptive to data correlations because it is a permutation method. Similar to Holm, it is based on the ordered p-values $p_{(1)} \leq \cdots \leq p_{(k)}$. Choose $B-1$ random permutations of the data vectors consistent with the experimental design. Denote the univariate p-values for the variables from the jth permutation by $p_1^j, \ldots, p_k^j$ for $j = 1, \ldots, B-1$, where p_1^j corresponds to $H_{(1)}$, p_2^j to $H_{(2)}$, etc. Therefore, $p_1^j, \ldots, p_k^j$ are not ordered. Let $p_{\min,1}^j = \min\{p_1^j, \ldots, p_k^j\}$ for $j = 1, \ldots, B-1$. In the first step, $p_{(1)}$ is compared with the α-quantile of the B p-values $p_{(1)}, p_{\min,1}^1, \ldots, p_{\min,1}^{B-1}$. If $p_{(1)}$ is larger than this α-quantile, none of the k hypotheses will be rejected and the procedure stops. If $p_{(1)}$ is smaller than or equal to this α-quantile, $H_{(1)}$ is rejected. Then omit all p-values corresponding to $H_{(1)}$, i.e. $p_{(1)}, p_1^1, \ldots, p_1^{B-1}$. Now, let $p_{\min,2}^j = \min\{p_2^j, \ldots, p_k^j\}$ for $j = 1, \ldots, B-1$. Then $p_{(2)}$ is compared with the α-quantile of the B p-values $p_{(2)}, p_{\min,2}^1, \ldots, p_{\min,2}^{B-1}$. If $p_{(2)}$ is larger than this α-quantile, $H_{(2)}, \ldots, H_{(k)}$ are accepted and the procedure stops. Otherwise $H_{(2)}$ will be rejected, etc.

As already mentioned in Troendle (1995), Troe is identical with the method of Westfall and Young (1993). We will see that the power of Troe is higher than the power of Holm in many cases.

Holm and Troe (and Bonf in Section 3.3) are the only FWE controlling methods that we consider in this paper. However, there exist many other ones. A step-up analogue of Holm was proposed by Hochberg (1988). It compares the p-values $p_{(i)}$ in the reverse order with the same critical bounds, i.e., in the first step $p_{(k)}$ with α, in the second step $p_{(k-1)}$ with

$\alpha/2, \ldots,$ in the last step $p_{(1)}$ with α/k. However, Hochberg's method is only valid under independence of the test statistics or at least under positive regression dependency which is a very general statement of a nonnegative correlation structure, see Sarkar (1998). In many-one comparisons by Horn and Dunnett (2004) the powers of Hochberg's method and Holm did not essentially differ. Thus, we did not include Hochberg's procedure in our investigations.

Many FWE controlling methods and techniques were described in the monographs of Hochberg and Tamhane (1987) and Westfall et al. (1999). A further interesting type of FWE controlling procedures was proposed by Läuter (1997), Kropf (2000) and Kropf et al. (2004). This approach uses a data driven order of the hypotheses.

2.1.2. Control of the FDR

The requirement FWE $\leq \alpha$ is equivalent to $P(V = 0) \geq 1 - \alpha$. This means one requires that with large probability no true hypothesis is rejected no matter how large k is. In problems with large k, this requirement appears to be too strict. Thus, Benjamini and Hochberg (1995) introduced a new criterion which requires FDR $= E(Q) = E(V/R \mid R > 0)\, P(R > 0) \leq \alpha$. For example, FDR ≤ 0.05 roughly means that on average no more than 5 of 100 significance statements are type I errors.

The control of the FDR in the evaluation of EEG data has already been proposed by Durka et al. (2004).

Benjamini and Hochberg (1995) were the first who proposed a test procedure (BH) that controls the FDR. Similar to Holm, it is based on the ordered p-values $p_{(1)} \leq \cdots \leq p_{(k)}$ obtained with some parametric or nonparametric test appropriate for $H_{(1)}, \ldots, H_{(k)}$. However, BH is a step-up procedure. In the first step, $p_{(k)}$ is compared with α. If $p_{(k)} \leq \alpha$, all hypotheses are rejected. If $p_{(k)} > \alpha$, $H_{(k)}$ cannot be rejected. Then $p_{(k-1)}$ is compared with $\alpha(k-1)/k$. If $p_{(k-1)} \leq \alpha(k-1)/k$, $H_{(k-1)}, \ldots, H_{(1)}$ are rejected. If $p_{(k-1)} > \alpha(k-1)/k$, $H_{(k-1)}$ cannot be rejected, etc.

Let m denote the unknown number of false and k–m the number of true hypotheses. Benjamini and Hochberg (1995) have shown that for their procedure FDR $\leq \alpha(k-m)/k$. Thus, FDR is smaller than α if $m > 0$ and decreasing with increasing m. If m were known one could increase the power of this step-up procedure by using $\alpha' = \alpha^* \cdot k/(k-m)$ instead of α. Based on this idea, Benjamini et al. (2001) developed a two stage procedure (BKY). In the first stage, BH is applied comparing $p_{(k)}, p_{(k-1)}, \ldots, p_{(1)}$ with the critical constants $\alpha' \cdot k/k, \alpha' \cdot (k-1)/k, \ldots, \alpha' \cdot 1/k$ where $\alpha' = \alpha/(1+\alpha)$. Let r_1 denote the number of hypotheses that would be rejected. r_1 is an estimate of m. Hence, we replace the denominator k of the critical constants by $k - r_1$ and repeat in the second stage BH comparing $p_{(k)}, p_{(k-1)}, p_{(k-2)}, \ldots, p_{(1)}$ with the critical constants $\alpha' \cdot k/(k-r_1), \alpha'(k-1)/(k-r_1), \ldots, \alpha' \cdot 1/(k-r_1)$.

Note that using BKY it is possible to reject a hypothesis with a p-value greater than α. In most cases, such an event is undesirable. Hence, we have modified the rule of rejection; we reject a hypothesis only if the p-value does not exceed α as well.

Both BH and BKY were derived under the assumption that the k test statistics are uncorrelated. In Benjamini and Yekutieli (2001) and Sarkar (2002) it was shown that the methods are also valid under the weaker assumption of positive regression dependency, similarly like the step-up procedure of Hochberg (1988) mentioned in Section 2.1.1. We will investigate by simulations whether FDR $\leq \alpha$ holds when the endpoints are correlated.

Step-down procedures that control the FDR have been proposed by Benjamini and Liu (1999, 2001) (BL99, BL01). BL99 requires independence of the test statistics or at least positive regression dependency whereas BL01 is valid also under dependency. BL99 compares the ordered p-values $p_{(i)}$ with the critical bounds $1 - [1 - \min\{1, \alpha k/(k-i+1)\}]^{1/(k-i-1)}$ and BL01 with the critical bounds $\min[1, \alpha k/(k-i+1)^2]$ ($i = 1, \ldots, k$). In Horn and Dunnett (2004) was shown that the power of BL01 is only slightly lower than that of BL99. Therefore, we applied only BL01 to the data in Section 3.3. Moreover, former simulations have shown that both methods are distinctly inferior to BH and BKY concerning their power, see Horn et al. (2003). Therefore, in this paper we executed no simulations for BL99 and BL01.

2.1.3. Control of the number V and relative number V/R of false discoveries

We remind that FWE control means that $P(V > 0) \leq \alpha$. With large k, it may be sufficient to require that $P(V > u) \leq \alpha$ for some prespecified integer u ($0 \leq u < k$). This means the strict requirement that no type I error occurs is lessened now requiring that no more than u type I errors occur. For example with $u = 2$ and $\alpha = 0.05$, we may require that $P(V > 2) \leq 0.05$ or equivalently $P(V \leq 2) \geq 0.95$, which means that 2 or less type I errors are accepted with probability 0.95. Korn et al. (in press) proposed a step-down procedure called Procedure A (PrA_u) which ensures that $P(V > u) \leq \alpha$ for some specified integer $u < k$. Computationally, PrA_u can be considered as an extension of Troe if $u > 0$. For $u = 0$, it is identical with Troe. In its first step, PrA_u automatically rejects $H_{(1)}, \ldots, H_{(u)}$. The further steps are more complex than with Troe. For details see Korn et al. (in press).

Now we remind that FDR control means that FDR $= E(Q) \leq \alpha$. However, this does not prevent that Q attains values much greater than α in single cases. For example, it can happen that BH at level $\alpha = 0.05$ rejects 100 hypotheses 20 of which are true hypotheses, so that $V/R = 20/100 = 0.2 > 0.05$. Therefore, Korn et al. (in press) proposed a step-down procedure called Procedure B (PrB_γ) that (asymptotically) guarantees that $P(Q > \gamma) \leq \alpha$ for some prespecified γ ($0 < \gamma < 1$). For example, with $\alpha = 0.05$, $\mathrm{PrB}_{0.1}$ guarantees that $Q > 0.1$ is possible only with a probability ≤ 0.05. PrB_γ is also an extension of Troe and uses nearly the same computational techniques as PrA_u. In dependence on the specified γ it automatically rejects some hypotheses in the different steps except in the first step (in contrast to PrA_u). For details see Korn et al. (in press).

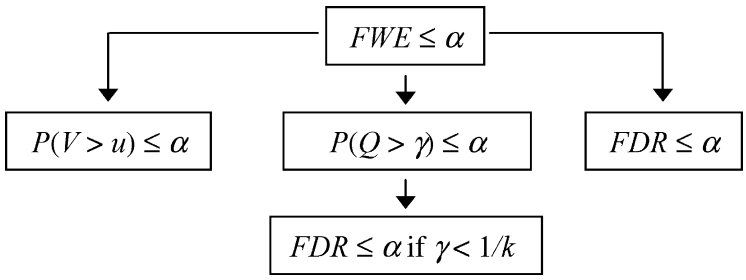

Fig. 1. Relations between control of different type I error rates.

Both, PrA_u and PrB_γ are based on the permutation principle and thus are adaptive to the correlation structure of the data.

2.1.4. Relations

The control of the FWE is the most stringent criterion. It implies the control of the FDR as well as of the number V and relative number V/R of false rejections. These implications are demonstrated in Fig. 1. In addition, we derived the relation FDR $\leq \gamma$ FWE + $(1 - \gamma)$ $P(Q > \gamma)$ for $0 < \gamma < 1$. The corresponding derivations are given in Appendix A.

2.2. Simulated data and real data (EEG data)

2.2.1. Simulated data

All procedures considered in this paper can be used for the paired samples case and the case of two independent samples. They all use the p-values for the different hypotheses. Therefore, it is not necessary to differentiate between the paired samples case and the case of two independent samples. Thus, we only simulated the paired samples case. Let $d_i = x_i - y_i$ denote the componentwise differences ($i = 1, \ldots, k$). Thereby in the paired samples case we have the random difference vectors $d = (d_1, \ldots, d_k)$ which have k-variate distributions. For these vectors, we generated samples from k-variate normal distributions for special configurations of means and correlation coefficients, and executed different multiple tests. The components of k-variate normally distributed vectors d had common variance 1, and the means $\mu_{x_i} - \mu_{y_i} = \mu_{d_i}$ ($i = 1, \ldots, k$) were chosen so that $\mu_{d_i} = \Delta$ ($i = 1, \ldots, m$) and $\mu_{d_i} = 0$ ($i = m + 1, \ldots, k$). This means we considered m false and $k - m$ true hypotheses, and the deviations of the false hypotheses were all into the same direction. The value of m was varied between 1 and k.

We denote the coefficients of correlation between d_i and d_j by ρ_{ij} ($1 \leq i \leq j \leq k$). We considered the cases $\rho_{ij} = 0.2$, i.e. constant low positive correlation, and $\rho_{ij} = 0.8$, i.e. constant high positive correlation ($i \neq j$). In most practical situations, the correlation coefficients ρ_{ij} do not have the same value and the same sign. Therefore, we also considered the following two types of correlation matrices Corr1 and Corr2. The matrix

$$\mathrm{Corr1} = \begin{pmatrix} 1 & \frac{k-1}{k} & \frac{k-2}{k} & \cdots & \frac{1}{k} \\ \frac{k-1}{k} & 1 & \frac{k-1}{k} & \cdots & \frac{2}{k} \\ \frac{k-2}{k} & \frac{k-1}{k} & 1 & \cdots & \frac{3}{k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \frac{1}{k} & \frac{2}{k} & \frac{3}{k} & \cdots & 1 \end{pmatrix}$$

may be typical for longitudinal observations, e.g. time series where neighboring observations have higher correlations than more distant observations. The matrix

$$\mathrm{Corr2} = \begin{pmatrix} R1 & R2 & R2 \\ R2 & R1 & R2 \\ R2 & R2 & R1 \end{pmatrix}$$

with

$$R1 = \begin{pmatrix} 1 & 2/3 & \cdots & 2/3 \\ 2/3 & 1 & \cdots & 2/3 \\ \vdots & \vdots & \cdots & \vdots \\ 2/3 & 2/3 & \cdots & 1 \end{pmatrix} \quad \text{and}$$

$$R2 = \begin{pmatrix} -1/3 & -1/3 & \cdots & -1/3 \\ -1/3 & -1/3 & \cdots & -1/3 \\ \vdots & \vdots & \cdots & \vdots \\ -1/3 & -1/3 & \cdots & -1/3 \end{pmatrix}$$

was used in order to investigate a case where both, positive and low negative correlations occur.

The number of repeated simulations for any configuration was 60.000 for most procedures, with the exception of the permutation methods Troe, PrA_u and PrB_γ where only 5.000 repetitions were done. The number of permutations in each permutation test was 1.000.

2.2.2. EEG data

A sample of 23 female German native speakers participated in the EEG experiment. They auditorily perceived two unrelated wordlists each containing 25 disyllabic abstract nouns. Participants had to memorize the nouns and immediately after the presentation of each list they were asked to recall the words previously encoded.

During word encoding EEG was recorded with 19 gold-cup electrodes according to the 10–20 system against the averaged signals ($A1 + A2$)/2 of both ear lobe electrodes. Filter settings were 0.3–35 Hz, sampling frequency was 256 Hz. According to the behavioral results EEG epochs of subsequently recalled and not recalled nouns were selected. One second epochs beginning with word onset were Fourier transformed and averaged cross-power spectra between all possible electrode pairs (171) were computed for each participant.

As it has been demonstrated, that particularly lower EEG frequencies are associated with dm-effects (differences due to memory performance; Fell et al., 2001; Klimesch et al., 1996; Weiss and Rappelsberger, 2000), adjacent spectral values were averaged to obtain broad band parameters for the following frequency bands: delta1 (1–2 Hz), delta (3–4 Hz), theta (5–7 Hz), alpha1 (8–10 Hz), alpha2 (11–12 Hz) and beta1 (13–18 Hz). The normalization of the cross-power spectra yielded 171 coherence values per frequency band, condition (recalled or not recalled) and participant. Coherence values were Fisher-*z*-transformed for the current statistical analysis. Further details of the experimental setup and methods of EEG analysis can be found in Weiss and Rappelsberger (2000) and Weiss et al. (2000).

In the present study our aim was to find out which of the 171 pairs of electrodes significantly differ in their means of coherence values for subsequently recalled versus not recalled nouns. For this task we needed a multiple test.

3. Results

3.1. Estimation of the FDR and P(Q > 0.1)

As mentioned in Section 2.1.2, BH and BKY control the FDR under the condition of independence of the test statistics or at least of positive regression dependency. In multiple endpoint problems, it is difficult to determine the correlation between the test statistics. However, it may be possible to estimate the correlation between the endpoints. (Of course, the correlation of the test statistics will be related to the correlation of the endpoints.) Thus, we investigated how the FDR of BH and BKY depends on the correlation of the endpoints. Here the FDR for the correlation structure Corr2 was most interesting, as there are negative correlation coefficients. Fig. 2 (left side) demonstrates for Corr2 that the FDR of BH decreases with increasing m whereas the FDR of BKY does not strongly change when m increases. The FDR of both methods is below the nominal level of 0.05. Similar results were obtained for $\rho = 0.2$, $\rho = 0.8$ and Corr1.

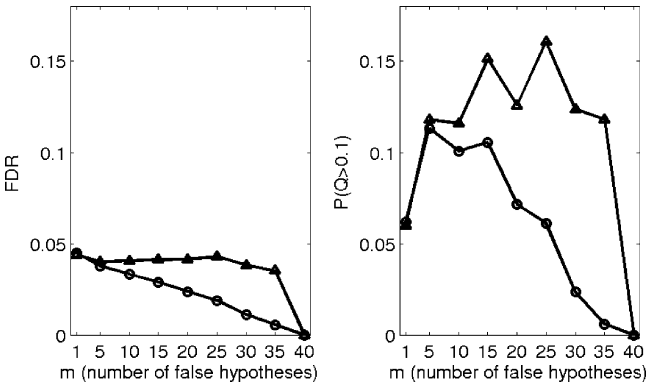

Fig. 2. Estimates of the FDR and of $P(Q > 0.1)$ for BH (circles) and BKY (triangles) under Corr2 ($n = 8$, $k = 40$, $\alpha = 0.05$, $\Delta = 1.5$).

As mentioned in Section 2.1.3, the requirement FDR $\leq \alpha$ cannot prevent that V/R attains large values. Therefore, we estimated the probability $P(Q > 0.1)$ for BKY and BH, see Fig. 2 (right side). We state that $P(Q > 0.1)$ for BKY and BH is smaller than 0.2.

3.2. Power comparisons

In multiple comparisons, there exist different concepts of power. We will use terms that originally were used in connection with pairwise multiple comparisons. The probability of rejecting at least one of the false hypotheses is called any-pair power, and the probability of rejecting all false hypotheses is called all-pairs power, see Ramsey (1978). If we consider a single false hypothesis, then the probability of rejecting it is called per-pair power, see Einot and Gabriel (1975). Kwong et al. (2002), Liu (1997) and Troendle (2000) preferred in their power comparisons the average power which is $E(R - V)/m$, i.e. the expected proportion of false hypotheses that were rejected. In our simulations, we considered equal differences $\mu_{x_i} - \mu_{y_i} = \Delta$ for all m false hypotheses. Then, the per-pair power of each false hypothesis has the same value, say p, so that $E(R - V) = mp$ and with it $E(R - V)/m = p$. This means that in our considerations the average power is identical with the per-pair power. We restricted our investigations to the per-pair power (average power) and all-pairs power as they seem to be most important in practice.

Our first task was to investigate how the power of our multiple test procedures depends on the fraction m/k of false hypotheses. We observed that the per-pair power of most method except PrA_2 increases with increasing m/k, see Figs. 3 and 4. However, the all-pairs power curves of most procedures are u-shaped, see Figs. 5 and 6.

Our second task was to investigate how the power of our methods depends on the correlation structure of the data, see Figs. 3–6. When $\rho_{ij} = \rho$ for $i \neq j$, i.e. when the correlation is the same for all pairs of components, the per-pair power of BH, BKY and PrA_2 decreases and that of Troe increases with increasing ρ, see Fig. 3, whereas the all-pairs power of all methods increases (only in some cases we state for large m/k a slight power decrease), see Fig. 5.

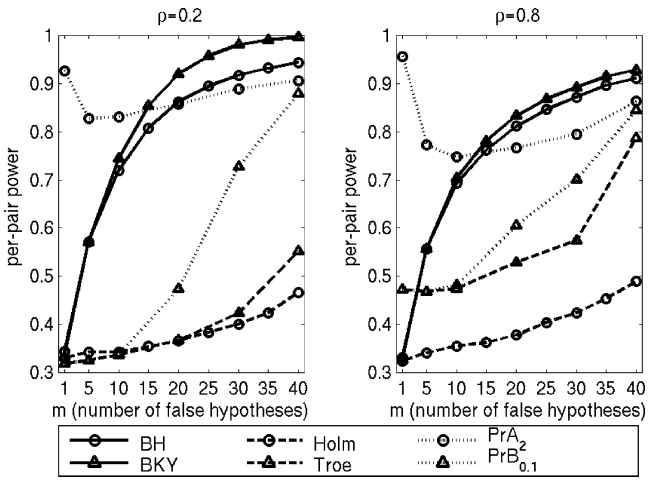

Fig. 3. Per-pair powers for $\rho = 0.2$ and $\rho = 0.8$ ($n = 8$, $k = 40$, $\alpha = 0.05$, $\Delta = 1.5$).

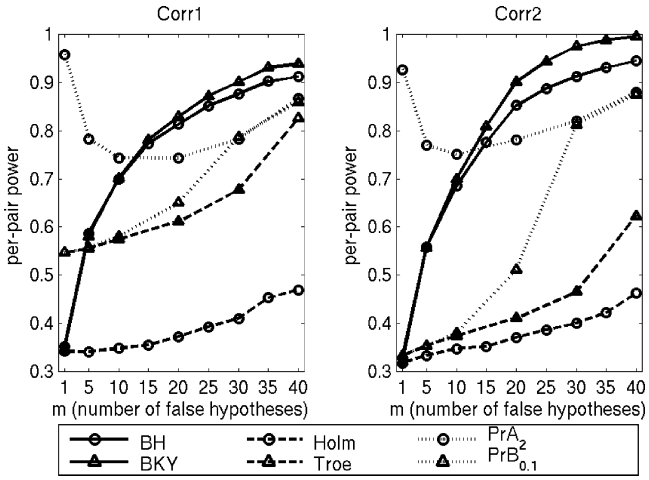

Fig. 4. Per-pair powers under Corr1 and Corr2 ($n = 8$, $k = 40$, $\alpha = 0.05$, $\Delta = 1.5$).

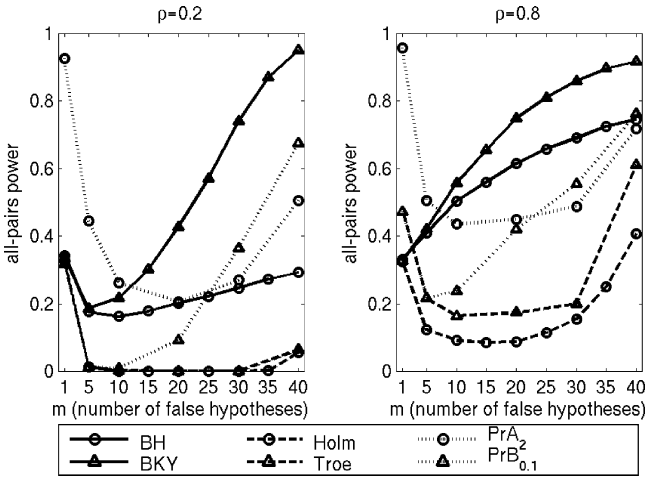

Fig. 5. All-pairs powers for $\rho = 0.2$ and $\rho = 0.8$ ($n = 8$, $k = 40$, $\alpha = 0.05$, $\Delta = 1.5$).

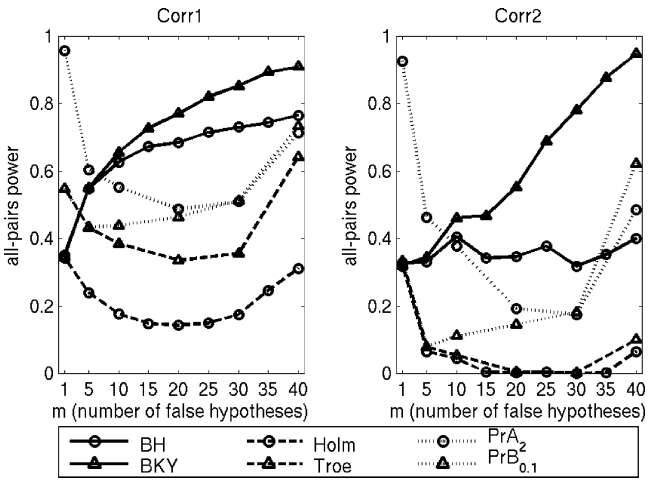

Fig. 6. All-pairs powers under Corr1 and Corr2 ($n = 8$, $k = 40$, $\alpha = 0.05$, $\Delta = 1.5$).

Our third task was to investigate how the power of our test procedures depends on the number k of hypotheses. Here the most important result is that the per-pair power of BKY and BH scarcely changes with increasing k, see Fig. 7. The all-pairs power of BKY and BH decreases moderately when the fraction of false hypotheses is small, and slightly when most hypotheses are false, see Fig. 8. As expected, the per-pair power and all-pairs power of the FWE controlling methods Holm and Troe decrease when k increases, see Figs. 7 and 8. The powers of Troe, PrA_u and PrB_γ were not calculated for $k > 40$ because of the immense computational effort. We expect that the per-pair power and all-pairs power of PrA_2, which are rather high for $m/k = 0.2$ and $k \leq 40$, will decrease very strongly with increasing k, so that they will be much lower for large k than the corresponding power values of BKY and BH. Figs. 7 and 8 are for $\rho = 0.8$. The results for $\rho = 0.2$ which are not shown here are very similar.

When we formally compare the different methods we state that PrA_2 has the highest per-pair power and all-pairs power

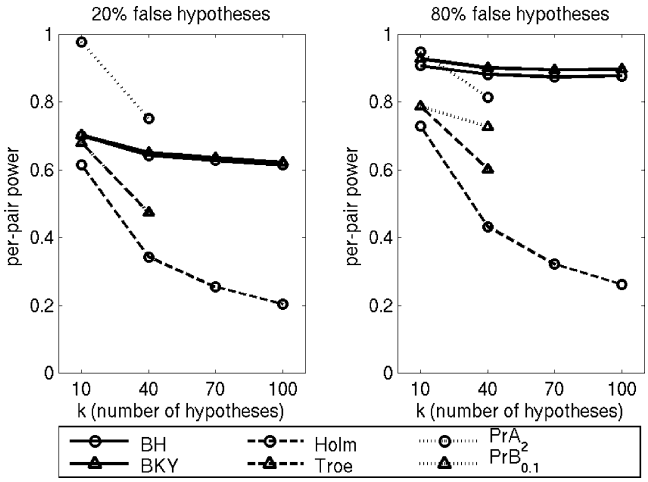

Fig. 7. Per-pair powers for $m/k = 0.2$ and $m/k = 0.8$ ($n = 8$, $\alpha = 0.05$, $\Delta = 1.5$, $\rho = 0.8$).

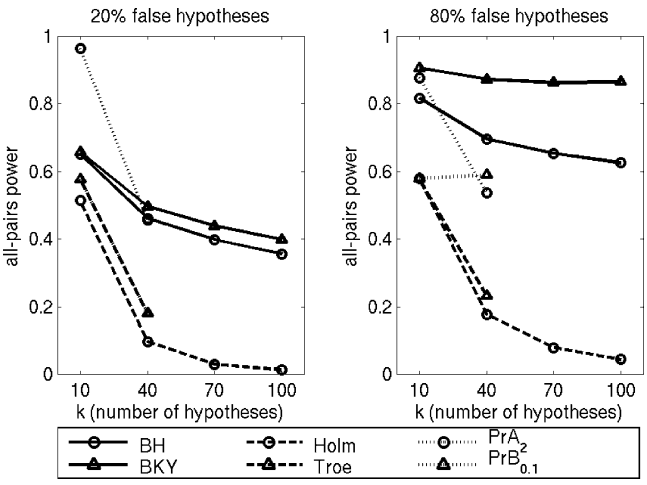

Fig. 8. All-pairs powers for $m/k = 0.2$ and $m/k = 0.8$ ($n = 8$, $\alpha = 0.05$, $\Delta = 1.5$, $\rho = 0.8$).

if $m/k \leq 1/4$ whereas BKY has the highest per-pair power and all-pairs power if $m/k > 1/4$. This applies for $\rho = 0.2$ and 0.8 as well as for Corr1 and Corr2, see Figs. 3–6. However, the per-pair power and all-pairs power of PrA_2 strongly decrease with increasing k, so that BKY becomes the most powerful method also for small fractions m/k.

3.3. Applications of multiple tests to EEG coherence data

The data we now evaluate come from the experiment described in Section 2.2.2. The number of subjects was 23. For each subject, a vector of 171 coherence values was obtained under two different conditions. This means we had to test $k = 171$ null hypotheses.

In all multiple tests, we used the paired t-test statistics. The number of significant mean differences at level $\alpha = 0.05$ with Bonf, Holm and Troe, PrA_1 and PrA_2, $PrB_{0.05}$, $PrB_{0.1}$, BL01, BH and BKY are given in Table 2. In the second column are also the number of significant results with the paired t-test which cannot be recommended as it is not a multiple test. Of course, this test provides more significant differences than the multiple tests. Among the multiple tests, most significant differences were found for BKY followed by BH. Among the FWE controlling methods, Troe is distinctly more powerful than Bonf and Holm for most frequency bands. It is also more powerful than the FDR controlling procedure BL01 which is considerably less powerful than BKY and BH.

4. Discussion

We have introduced four criteria for controlling type I errors (three of them may be new for most readers) and derived the relationships between them, see Fig. 1.

Control of the FWE means to require that no type I error occurs no matter how large the number of hypotheses and the number of rejected hypotheses is. With high-dimensional data, this criterion is too strict. Therefore, the other criteria were proposed.

In our opinion, the requirement $P(Q > \gamma) \leq \alpha$ provides the most reasonable criterion. Unfortunately, with our program the only corresponding procedure (PrB_γ) cannot be executed within an acceptable time of computation when $R > 30$. In this case, a compromise may be to stop the calculations when $R = 30$ and report the 30 corresponding (most significant) endpoints. Further research is needed to develop powerful procedures feasible for large k and R.

The requirement $P(V > u) \leq \alpha$ is a generalization of the FWE criterion. It is less strict, but it has similar disadvantages. In a practical application it is difficult to decide which number u is appropriate. Moreover, the only corresponding procedure (PrA_u) requires the same computational effort as PrB_γ.

As mentioned in Section 2.1.3, the requirement FDR = $E(Q) \leq \alpha$ does not prevent that Q attains large values. This is a general disadvantage of the FDR criterion. However, Fig. 2 shows that the probability $P(Q > 0.1)$ is relatively small for the FDR controlling methods BH and BKY. This means that these methods provide a good compromise as they do not strongly violate the requirement of the criterion we favor. Moreover, these two methods are computationally very simple and do not need a special computer program, in contrast to PrA_u and PrB_γ. In addition, our formal comparisons demonstrated that BKY has a relatively high power. Therefore, this procedure seems to be most recommendable. However, caution is needed when comparing methods that satisfy different crite-

Table 2
Number of significant coherence differences for different tests when comparing the processing of subsequently recalled and not recalled nouns for all frequency bands analyzed

	t-test	FWE ≤ 0.05			$P(V > u) \leq 0.05$		$P(Q > \gamma) \leq 0.05$		FDR ≤ 0.05		
		Bonf	Holm	Troe	PrA_1	PrA_2	$PrB_{0.05}$	$PrB_{0.1}$	BL01	BH	BKY
delta1	78	5	6	10	12	25	9	17	7	54	60
delta	64	1	1	7	16	21	7	7	1	44	47
theta	27	4	4	6	7	10	6	6	4	10	10
alpha1	14	0	0	0	1	2	0	0	0	0	0
alpha2	44	0	0	3	7	12	1	1	0	12	12
beta1	55	5	5	5	11	14	7	7	5	23	25

ria because one can expect that the strictest criterion leads to the lowest power.

Note that PrA_u, PrB_γ and Troe are permutation methods. Such methods have the advantage that they consider the correlation of data. This seems to be the reason why these methods are more powerful for $\rho = 0.8$ than for $\rho = 0.2$, see Figs. 3 and 5.

In order to demonstrate the properties of different multiple tests we applied them to EEG coherence data obtained while participants memorized abstract nouns subsequently recalled or not. The major result was that during the phase of word encoding, subsequently recalled nouns elicited higher EEG coherence than not recalled nouns at all electrode pairs showing significant differences. Thus those words which are likely to be recalled are associated with an increase of synchronized activity between various brain regions, in particular left hemispheric sites and between both hemispheres. All frequency bands analyzed demonstrated significantly higher coherence for recalled nouns with the exception of the alpha1 band (8–10 Hz), which did not show any significant differences. The latter finding agrees well with the assumption that alpha1 predominantly reflects sensory processing, or general attentional processes (Klimesch et al., 1996; Weiss and Rappelsberger, 2000), and does not reflect differences in memory encoding. In contrast, coherence in the other frequency bands differed considerably for recalled and not recalled abstract nouns. A similar finding to our study was reported by Fell et al. (2003) studying intracortical recordings of the temporal lobe during memory encoding of words. In addition, Besthorn et al. (1994) found that patients with Alzheimer's disease, who suffer from a major disturbance of memory functions, exhibited lower coherence than healthy controls in the theta, alpha and beta frequency bands. In our opinion, the higher synchronization for recalled nouns in various frequency bands is associated with the participation of different cognitive operations such as short-time as well as long-time memory, attention to internal thinking, encoding and storage of episodic information and semantic associations occurring during the task, the latter probably being increased during the encoding phase and therefore leading to an improved ability to recall the nouns at a later time (Weiss and Rappelsberger, 2000). Multiple tests demonstrate that during the encoding phase subsequently recalled nouns are embedded within a more complicated network of interactions between various recording sites than not recalled nouns.

Acknowledgement

This work was supported by Interdisziplinäres Zentrum für Klinische Forschung Jena, Projekt 1.8, the Austrian Science Foundation ("Herta Firnberg"-project T127) and the German Science Foundation (SFB 360). We gratefully acknowledge the helpful comments of the two referees.

Appendix A

Here are the derivations for the statements given in Fig. 1. We have

$$\begin{aligned}\mathrm{FDR} &= E\left(\frac{V}{R}|R>0\right)P(R>0)\\ &= \left\{0\cdot P\left(\frac{V}{R}=0|R>0\right)\right.\\ &\qquad \left.+\sum_{w>0} wP\left(\frac{V}{R}=w|R>0\right)\right\}P(R>0)\\ &\le P\left(\frac{V}{R}>0|R>0\right)P(R>0)\\ &= P(V>0,R>0)=P(V>0)=\mathrm{FWE}.\end{aligned}$$

This means FWE $\le \alpha$ implies FDR $\le \alpha$. If all hypotheses are true we have $V = R$. Then FDR $= E(V/R|R>0)P(R>0) = P(R>0) =$ FWE. Thus, in this case FWE and FDR control are equivalent. (If FWE $\le \alpha$ when all hypotheses are true then the FWE control is called *weak*.)

As $P(V>u) \le P(V>0)$ for $u \ge 0$, FWE $\le \alpha$ also implies $P(V>u) \le \alpha$.

As $P(Q>\gamma) = P(V>\gamma R) \le P(V>0)$, FWE $\le \alpha$ also implies $P(Q>\gamma) \le \alpha$.

Furthermore, we have

$$\begin{aligned}\mathrm{FDR} &= \left\{0\cdot P\left(\frac{V}{R}=0|R>0\right)\right.\\ &\qquad +\sum_{0<w\le\gamma} wP\left(\frac{V}{R}=w|R>0\right)\\ &\qquad \left.+\sum_{w>\gamma} wP\left(\frac{V}{R}=w|R>0\right)\right\}P(R>0)\\ &\le \left\{\gamma P\left(0<\frac{V}{R}\le\gamma|R>0\right)\right.\\ &\qquad \left.+P\left(\frac{V}{R}>\gamma|R>0\right)\right\}P(R>0)=B.\end{aligned}$$

If $\gamma < 1/k$, we have $P(0 < V/R \le \gamma \mid R>0) = 0$ because V/R cannot take positive values smaller than $1/k$. We then obtain $B = P(V/R > \gamma \mid R>0)P(R>0) = P(Q>\gamma)$. Thus, $P(Q>\gamma) \le \alpha$ implies FDR $\le \alpha$, when $\gamma < 1/k$.

If $0 < \gamma < 1$, we have

$$\begin{aligned}B &= \left\{\gamma\left(1-P\left(\frac{V}{R}=0|R>0\right)-P\left(\frac{V}{R}>\gamma|R>0\right)\right)\right.\\ &\qquad \left.+P\left(\frac{V}{R}>\gamma|R>0\right)\right\}P(R>0)\\ &= \gamma P(R>0)-\gamma\{P(V=0)-P(R=0)\}\end{aligned}$$

$$+(1-\gamma)P\left(\frac{V}{R}>\gamma|R>0\right)P(R>0)$$

$$=\gamma-\gamma\{1-P(V>0)\}$$

$$+(1-\gamma)P\left(\frac{V}{R}>\gamma|R>0\right)P(R>0)$$

$$=\gamma\mathrm{FWE}+(1-\gamma)P(Q>\gamma),$$

so that $\mathrm{FDR}\leq\gamma\mathrm{FWE}+(1-\gamma)P(Q>\gamma)$.

References

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B 1995;57:289–300.

Benjamini Y, Liu W. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. J Stat Planning Inference 1999;82:163–70.

Benjamini Y, Liu W. A distribution-free multiple-test procedure that controls the false discovery rate. Technical Report. Tel Aviv University, 2001.

Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. The Ann Stat 2001;29:1165–88.

Benjamini Y, Krieger A, Yekutieli D. Two staged linear step up FDR controlling procedure. Technical Report. Tel Aviv University, 2001.

Besthorn C, Forstl H, Geiger-Kabisch C, Sattel H, Gasser T, Schreiter-Gasser U. EEG coherence in Alzheimer disease. Electroencephalogr Clin Neurophysiol 1994;90:242–5.

Durka PJ, Zygierewicz J, Klekowicz H, Ginter J, Blinowska KJ. On the statistical significance of event-related EEG desynchonization and synchronization in the time–frequency plane. IEEE Trans Biomed Eng 2004;51:1167–75.

Einot I, Gabriel KR. A study of the powers of several methods of multiple comparisons. J Am Stat Assoc 1975;70:574–83.

Fell J, Klaver P, Lehnertz K, Grunwald T, Schaller C, Elger CE, et al. Human memory formation is accompanied by rhinal–hippocampal coupling and decoupling. Nat Neurosci 2001;4:1259–63.

Fell J, Klaver P, Elfadil H, Schaller C, Elger CE, Fernández G. Rhinal-hippocampal theta coherence during declarative memory formation: interaction with gamma synchronization? Eur J Neurosci 2003;17:1082–8.

Hemmelmann C, Horn M, Reiterer S, Schack B, Süsse T, Weiss S. Multivariate tests for the evaluation of high-dimensional EEG data. J Neurosci Methods 2004;139(1):111–20.

Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75:800–2.

Hochberg Y, Tamhane AC. Multiple comparison procedures. New York: Wiley; 1987.

Holm S. A simple sequentially rejective multiple testing procedure. Scand J Stat 1979;6:65–70.

Horn M, Dunnett CW. Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. In: Benjamini Y, Sarkar SK, Bretz F, editors. Recent developments in multiple comparison procedures, vol. 47. IMS Lecture Notes Monograph Series; 2004.

Horn M, Hemmelmann C, Süsse T. A comparative study of test procedures for multiple endpoints concerning FDR, FDP FWE and power. Paper for Meeting of the German MCP Group, Magdeburg, 2003.

Klimesch W, Schimke H, Doppelmayr M, Ripper B, Schwaiger J, Pfurtscheller G. Event-related desynchronisation (ERD) and the Dm effect: does alpha desynchronization during encoding predict later recall performance? Int J Psychophysiol 1996;24:47–60.

Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. J Stat Planning Inference, in press.

Kropf S. Hochdimensionale multivariate Verfahren in der medizinischen Statistik. Aachen: Shaker Verlag; 2000.

Kropf S, Läuter J, Eszlinger M, Krohn K, Paschke R. Nonparametric multiple test procedures with data-driven order of hypotheses and with weighted hypotheses. J Stat Planning Inference 2004;125:31–47.

Kwong KS, Holland B, Cheung SH. A modified Benjamini–Hochberg multiple comparisons procedure for controlling the false discovery rate. J Stat Planning Inference 2002;104:351–62.

Läuter J. Stable multivariate tests for analyzing repeated measurements. In: Vollmar J, et al., editors. Proceedings of the Eighth DIA Annual European Workshop on Statistical Methodology in Clinical Research and Development, 1997.

Liu W. On step-up tests for comparing several treatments. Stat Sinica 1997;7:957–72.

Ramsey PH. Power differences between pairwise multiple comparisons, comments. J Am Stat Assoc 1978;73:479–87.

Rappelsberger P, Petsche H. Probability mapping: power and coherence analyses of cognitive processes. Brain Topogr 1988;1:46–54.

Sarkar SK. Some probability inequalities for ordered MTP_2 random variables: a proof of Simes' conjecture. Ann Stat 1998;26:494–502.

Sarkar SK. Some results on false discovery rate in stepwise multiple testing procedures. Ann Stat 2002;30:239–57.

Troendle JF. A stepwise resampling method of multiple testing. J Am Stat Assoc 1995;90:370–8.

Troendle JF. Stepwise normal theory multiple test procedures controlling the false discovery rate. J Stat Planning Inference 2000;84:139–58.

Weiss S, Rappelsberger P. Long-range EEG synchronization during word encoding correlates with successful memory performance. Cogn Brain Res 2000;9:299–312.

Weiss S, Müller HM, Rappelsberger P. Theta synchronisation predicts efficient memory encoding of concrete and abstract nouns. NeuroReport 2000;11:2357–61.

Westfall PH, Young SS. Resampling-based multiple testing: Examples and methods for p-value adjustment. New York: Wiley; 1993.

Westfall PH, Tobias RD, Rom D, Wolfinger RD, Hochberg Y. Multiple comparisons and multiple tests using the SAS system. Cary, NC: SAS Institute Inc.; 1999.