8th International IEEE EMBS Conference on Neural Engineering
Shanghai, China, May 25 - 28 , 2017

Improving sensitivity of cluster-based permutation test for EEG/MEG data

Gan Huang and Zhiguo Zhang

***Abstract*—To solve multiple comparisons problems in EEG/MEG analyses, cluster-based permutation test is possibly the most powerful approach, while it also inherits the advantage of well-controlled family-wise error rate from point-level permutation test. Because the cluster-level statistics used accumulate statistical power of all points in a cluster, cluster-based permutation test has a much higher sensitivity for widespread clusters. In this study, we demonstrate that, when the threshold for cluster inclusion is inappropriately set, the existence of larger clusters lowers the sensitivity for detecting the presence of smaller clusters, because the influence of large clusters on permutation distribution is overlooked in previous studies. Further, we demonstrated that increasing the threshold for cluster inclusion can efficiently solve this problem and then proposed a new guideline for threshold selection in the cluster-based permutation test. Results on simulated data and real data show the proposed guideline can greatly improve the sensitivity of cluster-based permutation test for detecting small clusters while retaining the same family-wise error rate.**

I. Introduction

Point-wise comparison, which compares EEG/MEG data at massive time, spatial and/or frequency points, is gaining popularity for identifying significant effects of EEG/MEG experiments. Conventionally, researchers compare the peak amplitude or mean value of EEG/MEG data in a prespecified interval/region to examine the significant effect [1]. Compared with the conventional method, the point-wise comparison has several advantages. First, it does not depend on researchers' prior knowledge for interval selection; hence it is less possible to miss any unexpected effects outside the interval/region of interest. Second, the point-wise comparison could provide more accurate and more detailed information about the time, spatial and/or frequency points with significant effects. However, these advantages are often obtained at some cost, like much heavier computational load. More seriously, point-wise comparison will inevitably induce the Multiple Comparisons Problem (MCP). For example, for a single point test with α=0.05, there is a 5% chance of incorrectly rejecting the null hypothesis if the null hypothesis is true (i.e. Type I error). However, since the probability of at least one incorrect rejection, noted as Family-Wise Error Rate (FWER), is $1-(1-\alpha)^m$, FWER will be close to 1 when the number of points m is very large. Typically, the problem of FWER control in MCP would be more serious for joint domain analysis, like time-spatial, time-frequency or even time-frequency-spatial domain analysis, since the number of comparisons is extremely large.

Resrach supported by National Natural Science Foundation of China，No. 61640002 and Research Project of State Key Laboratory of Mechanical System and Vibration MSV201710.

Gan Huang is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China 21006, (corresponding author e-mail: huanggan1982@gmail.com).

Zhiguo Zhang is with the Health Science Center, Shenzhen University, Shenzhen, China 518060, (e-mail: zgzhang@szu.edu.cn).

Several methods could effectively control the FWER. Bonferroni correction controls the FWER strictly by setting α level to α/m for m comparisons [2]. However, Bonferroni correction comes at the cost of increasing the probability of false negatives (i.e., Type II error), and consequently reducing statistical power. In addition, in most cases of EEG/MEG analyses, the data are correlated with their neighbors, which does not agree with the independent assumption in Bonferroni correction. To alleviate the problem of statistical power, False Discovery Rate (FDR) [3, 4] is developed with a weak control of FWER (i.e., FWER is only guaranteed if all the null hypothesis is true). On the other hand, permutation test [5, 6] can automatically adapt to the degree of the correlation among data points with a statistic t_{max} (the most extreme positive or negative t-value). Unlike parametric tests, such as t-test, permutation test does not make specific assumptions about the shape of the population distribution. Depending on the definition of t_{max}, permutation test can be performed on point-level or cluster-level.

Point-level permutation test [5] works for single point as follows. 1). Repeat the permutation process for all possible combination. If the sample size is large, randomly permute labels for a larger number of times and calculate the t-value for each permutation. 2). Generate the distribution of t-value under the null hypothesis. With the distribution, calculated the threshold corresponding to a certain α level (e.g. 5%). The points with their t-value above the threshold are identified to be statistically significant. For multiple comparison, point-level permutation test uses the statistic t_{max} (the most extreme positive or negative t-value) to generate the permutation distribution. The distribution of t_{max} adaptively reflects the degree of the correlation among data points [8]. Similar to Bonferroni correction, this nonparametric method provides a strong FWER control (i.e. under any mixture of false and true null hypothesis) for MCP [10]. However, when the number of tests is extremely large, the permutation test will become conservative.

Cluster-based permutation test uses the cluster-level statistic t_{max} instead of the point level statistic t_{max}, so that it can drastically increase the sensitivity of the statistical test while strictly controls the FWER [6]. As a weak FWER control method, cluster-based permutation test provides a higher sensitivity than FDR [7]. In fact, cluster-based permutation test provides a cluster-level FWER control, which is different from FDR.

Although cluster-based permutation test is possibly the most powerful procedure for broad effects detection, it is shown to have lower sensitivity for smaller clusters so that some meaningful small EEG/MEG effects might be overlooked [8]. The sensitivity of cluster-based permutation test depends on the thresholds for cluster inclusion. Generally, strong and localized effects can be detected at higher thresholds for cluster inclusion, while weak and widespread effects prefer lower thresholds for cluster inclusion [9]. However, there is no guideline for the threshold selection in literature using cluster-based permutation test [6].

In this work, we investigate in-depth the relationship between the threshold selection and sensitivity (especially for smaller effects) in cluster-based permutation test and find that the lower sensitivity of

978-1-5090-4603-4/17/$31.00 ©2017 IEEE

cluster-based permutation test for small clusters is due to the overlooked influence of large clusters on permutation distribution. Further, we introduce a new guideline for threshold selection in the cluster-based permutation test and validate the guideline using synthetic data and real EEG data.

The rest of the paper is organized as follows. Cluster-based permutation test is firstly introduced in Section II. Then, with a simple simulation model, the problem of the sensitivity for smaller effects is illustrated in Section III. To fix this problem, a guideline for the selection of threshold for cluster inclusion is proposed in Section IV. Both the simulation and real data is used to show the effectiveness and robustness of the guideline. The conclusion is arranged in Section V.

II. Cluster-based permutation test

Cluster-based permutation test provides an alternative way to the solve MCP, which could control the FWER, but not so conservative as Bonferroni-Correction. This procedure mainly includes the following two steps:

1. Calculate the cluster-level statistic: Calculate *t*-value, noted as t_{point}, for every point of interest in the temporal, frequency or spatial domain. All points with t_{point} not exceed the point-level threshold corresponding to certain α level, noted as α_{point}, are ignored. Cluster the remaining points in connected sets on the basis of temporal, frequency or spatial adjacency. Calculate cluster-level statistics, noted as $t_{cluster}$, by taking the sum of t_{point} within a cluster. Let t_{max} be the most extreme value of $t_{cluster}$.
2. Perform the permutation test: Repeat the permutation process and calculate their t_{max} to generate the permutation distribution. With the distribution, the cluster-level threshold is determined by certain α level, noted as $\alpha_{cluster}$. Any cluster under the true labels with its $t_{cluster}$ above the threshold will be checked out as statistical significant.

The whole procedure is controlled by two level thresholds. Cluster-level threshold, the threshold for cluster detection determined by $\alpha_{cluster}$, guarantees the FWER under the null hypothesis. Point-level threshold, the threshold for cluster inclusion determined by α_{point}, controls the points included in the cluster. The point-level threshold does not affect the FWER for the MCP, but does affect the sensitivity of the test. Since the sensitivity of this test is determined by the largest cluster in each permutation, the sensitivity for the smaller clusters (the second, third, …, largest clusters) will be reduced.

III. Existing Problems

In this section, we use a simulation model to demonstrate the problem of sensitivity in the conventional cluster-based permutation test.

A. Simulation Model

Considering the ERP signal

$$X(t) = S(t) + \mathcal{E} \quad (1)$$

where $X, S, \mathcal{E} \in \mathcal{R}^{N\times T}$, $X(t)$ is 1 channel EEG signal with N=8 trials and T=1000 time points, S is the signal, and $\mathcal{E}\sim\mathcal{N}(0,1)$ is the white noise. To illustrate the problem of the sensitivity for smaller clusters detection in cluster-based permutation test, we simulate three types of multiple comparison experiments in time domain.

Exp1: background noise, i.e. S(t)=0;

Exp2: ERP with a smaller effect. S(51:100)=3, S(t)=0 for else;

Exp3: Except the smaller effect, a larger effect is included. S(51:100)=3, S(501:900)=10, S(t)=0 for else.

B. Results for α_{point} = 0.05 and 0.0112 with $\alpha_{cluster}$=0.05

Here we investigate the sensitivity of the smaller cluster in Exp1, 2 and 3 with different value of α_{point} (0.05 and 0.0112) but keep $\alpha_{cluster}$=0.05.

With α_{point} = 0.05, the point-level *t*-value for the points in the smaller cluster in theory is

$$t_{point} = \frac{\bar{x}-0}{std(x)/\sqrt{N}} \approx \frac{3-0}{1/\sqrt{8}} = 8.49$$
$$> 2.36 = t(\alpha_{point} = 0.05, df = N-1).$$

Hence, these points would be clustered together with α_{point} = 0.05. And the cluster-level *t*-value for the smaller cluster is

$$t_{cluster} = t_{point} \times cluster\_size \approx 8.49 \times 50 = 424.5.$$

In practice, due to the inclusion of neighboring spurious points, the cluster-level *t*-value is 483.82 in the experiment. The cluster threshold with $\alpha_{cluster}$=0.05 is 10.11, 27.35 and 889.5 for Exp 1-3, and the permutation distribution is shown in Fig. 1a. It is clear that the existence of the larger cluster makes the cluster level threshold dramatically increased. The smaller cluster, which would be detected in Exp2, does not survive in Exp3. It could also be noted that in Exp2 with the smaller cluster, the cluster level threshold is also a little bit higher than Exp1 with background noise.

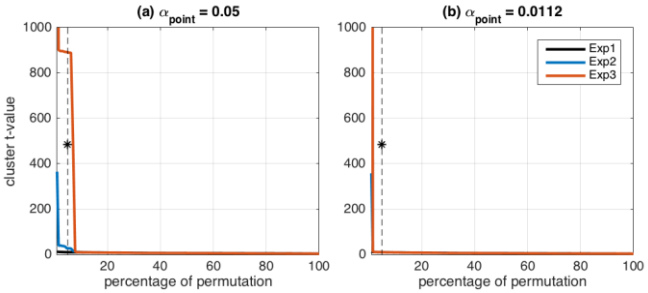

Figure 1. Permutation distribution for Exp 1-3 with (a) α_{point} = 0.05 and (b) α_{point} = 0.0112. The cluster-level threshold for the smaller cluster is marked with an asterisk.

If we reduce the α_{point} from 0.05 to 0.112, the points in the smaller cluster with their *t*-value

$$t_{point} \approx 8.49 > 3.42 = t(\alpha_{point} = 0.0112, df = N-1).$$

will still be clustered together. However, with αpoint = 0.112 the cluster threshold in Exp 2 and 3 is 10.44 and 10.11, which are similar close to 10.00 in Exp 1 with background noise (Fig. 1b).

IV. Guideline for point-level Threshold Selection

In the last section, with inappropriate point-level α_{point}=0.05, the existence of larger clusters induces lower sensitivity for the detection of smaller clusters. Hence, how to decide the value of point-level threshold is key important to the statistic power of cluster-based permutation test. In this section, an empirical guideline for threshold selection in the cluster-based permutation test is proposed. The performance is tested on both simulation dataset and real dataset.

A. Point-level threshold selection

Cluster-level threshold is determined by the main effect corresponds to largest cluster at the cluster-level $\alpha_{cluster}$ of the permutation distribution. In the study above, with certain α_{point} level, the value of the cluster-level threshold will be cumulated by the size of the main effect in the permutation, which is related to the size of the largest cluster with the true label. Hence the existence of the

larger cluster may influence the sensitivity for the smaller cluster detection. Furthermore, if we decrease α_{point} to certain value (increase the point-level threshold) until the *t*-value for each points in the largest cluster with the true label in the permutation at $\alpha_{cluster}$ of the permutation distribution below the point-level threshold, the cluster-level threshold will not be influenced by the largest cluster with the true label. In this case, the cluster-level threshold will only be determined by the background noise and the sensitivity for the smaller cluster will not be influenced by the larger cluster any more.

For single comparison, the *t*-value for a single point with the true labels would be infinite, but there is always a boundary for the threshold in the permutation distribution at certain α level with sample size $N > \log_2 \frac{1}{\alpha} + 1$. For example, with sample size N=8, the threshold in permutation test with α=0.05 would be always less than or equal to $\sqrt{35/3}$, which is determined by the sequence [1, 1, 1, 1, 1, 0, 0 ,0].

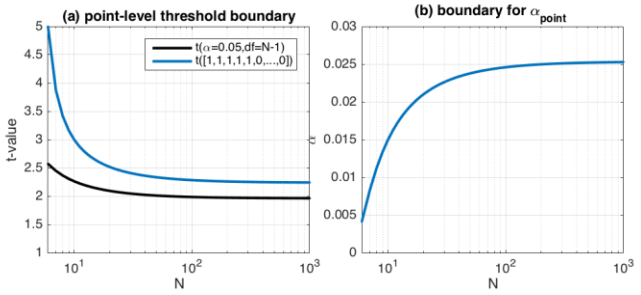

Figure 2. The boundary for (a) point-level threshold and (b) point-level α_{point} in blue curve at cluster-level $\alpha_{cluster}$=0.05 with N from 6 to 1000.

Hence in the cluster-based permutation test for multiple comparison, if point-level threshold is larger than the boundary (point-level α_{point} less than the boundary) at certain $\alpha_{cluster}$ with sample size $N > \log_2 \frac{1}{\alpha_{cluster}} + 1$, then the cluster-level threshold will not be cumulated by the size of the largest cluster with the true label. In practice, we are unable to get an effective way to calculate the boundary with any $\alpha_{cluster}$ and N. But we could still provide the boundary for point-level threshold with some special value of cluster-level $\alpha_{cluster}$. Empirically, at $\alpha_{cluster}$=0.05 with sample size $N \geq 6$, the largest *t*-value corresponding $\alpha_{cluster}$ of the permutation distribution is determined by the sequence [1, 1, 1, 1, 1, 0, ……,0] with five 1s, and 0s for the rest $N-5$ values. Hence the boundary for point-level threshold is $\sqrt{\frac{5(N-1)}{N-5}}$ (Fig. 2), the corresponding boundary for point-level α_{point} is shown in Fig. 3b. Hence, for N=8, the boundary for the point-level α_{point} is 0.0112. Similarly, at $\alpha_{cluster}$=0.01 with sample size $N \geq 8$, the boundary for point-level threshold is $\sqrt{\frac{7(N-1)}{N-7}}$.

B. Simulation dataset study

It should be noted that the guideline of point-level threshold selection is proposed by the study of single comparison. Whether the boundary is hold in cluster-level multiple comparison is a question. Here, the simulation model Eq. (1) is used to explore this problem. Only one large effect is kept with S(501:900)=D and S(t)=0 for else, where the value of D is used to control the signal noise ratio, $\alpha_{cluster}$ is always kept to 0.05 in the study. Fig. 3 shows the value of cluster-level threshold with the different value of D and α_{point}. With D=0, there is no signal. It is exactly the same as Exp1 with only background noise. With the α_{point} increases from 0.001 to 0.05, the cluster-level threshold is increased from 88.51 to 97.16, but not so greatly. With D=30, the signal-noise ratio for the larger cluster is great. The cluster-level threshold is directly determined by the size of the larger cluster with α_{point}>0.03. While if α_{point}<0.015, the cluster-level threshold is not influenced by the larger cluster any more. With a moderate signal-noise ratio, like D=1 for α_{point}=0.0112, under the complex effect of signal and noise, the cluster-level threshold is still larger than that with D=0, but the cumulate effect is not happened. In fact, the guideline for the point-level threshold selection provides the low boundary to prevent the cumulate effect happening. With certain level of signal-noise ratio, the cluster-level threshold will still be influenced by some clusters.

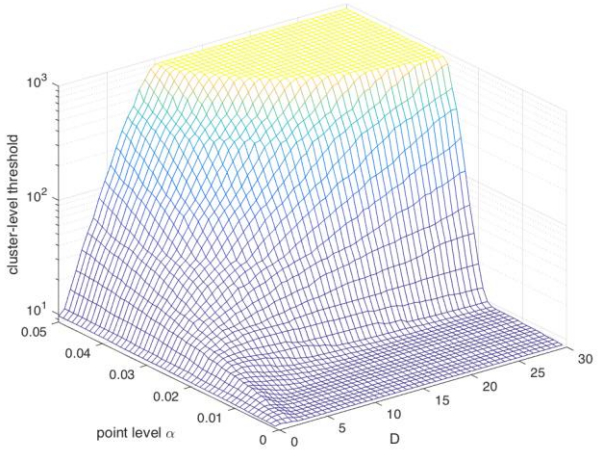

Figure 3. The cluster-level threshold with different value of α_{point} and D.

C. Real dataset study

The cluster-based permutation test in temporal-spatial domain was studied with a visual oddball ERP dataset from Groppe et al. 2009, Experiment 3 [11]. Groppe et al. also used this dataset as an example in the review for the method in MCP [8]. The dataset was collected from -0.1 to 0.92s with 8 subjects, 26 channels and sampling rate 250Hz. Fig. 4 shows the averaged ERP for target and standard conditions. Fig. 5 and 6 show the procedure of cluster-based permutation test with α_{point}=0.05 and 0.0112. For cluster detection, an electrode's spatial neighborhood was defined as all electrodes within approximately 5.4 cm, resulting in 3.8 neighbors for each electrode on average [8].

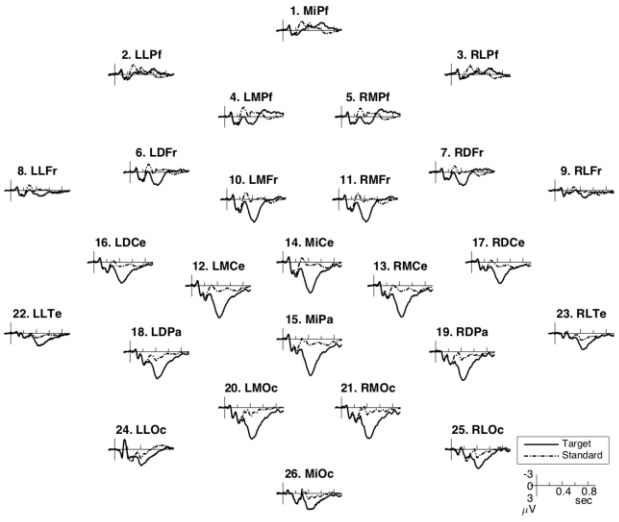

Figure 4. The ERPs for the visual oddball experiment.

With α_{point}=0.05, all the points above the point-level threshold form several clusters in Fig. 5a. In the cluster-based permutation test, the cluster-level threshold is determined by the largest cluster in the 7th permutation (Fig. 5d), which is mainly determined by the largest cluster with the true label. In result, the cluster-level threshold is larger than the cluster-level *t*-value for the second largest cluster (Fig. 5c). Hence, the second largest cluster is not survived in result (Fig. 5b). By carefully comparing the permutation distribution of the Fig. 1a and Fig. 5c, there is the similar inflection points at 7.81% in both real and simulation dataset, which corresponds to the 10th permutation in the distribution.

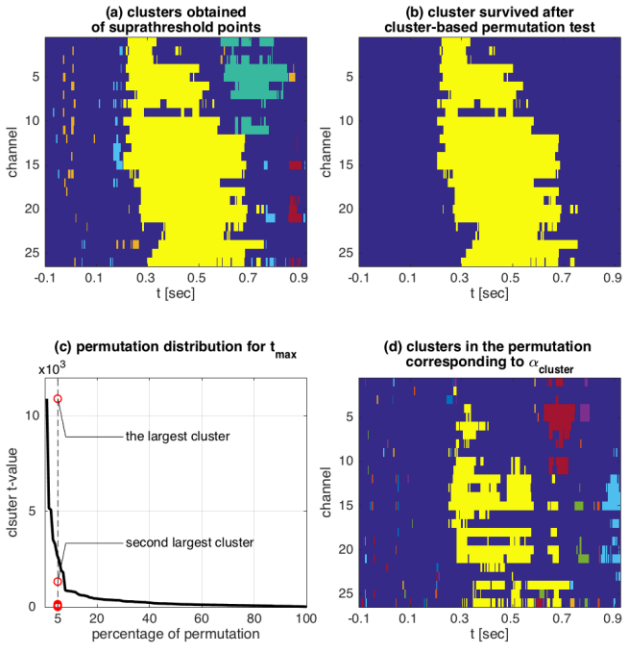

Figure 5. The procedure for cluster-based permutation test with α_{point}=0.05. (a) clusters obtained from temporal-spatial adjacent points with their t_{point} above point-level threshold, (b) the largest cluster survived in result, (c) permutation distribution for t_{max}, (d) the clusters in the 7th permutation of the distribution, the largest one (the yellow cluster) determines the cluster-level threshold.

With α_{point}=0.0112, as the boundary we get in the Section IV.A, the cluster-level threshold is determined by the cluster from channel 12 to 21 at around 0.05s (Fig. 6d). It is not influenced by the largest cluster with the true label any more. Hence, both the largest and the second largest cluster are survived (Fig. 6b).

V. Conclusion and Discussion

The performance of cluster-based permutation test is controlled by the two parameters. Cluster-level $\alpha_{cluster}$ controls the FWER, and point-level α_{point} will influence the sensitivity. In this paper, it is found that with the inappropriately parameter setting of point-level α_{point} (i.e. point-level threshold for cluster inclusion), which is commonly used in the research literature, the existence of larger cluster will dramatically decrease the sensitivity for smaller cluster detection. The reason is that the value of the cluster-level threshold may be cumulated by the size of the main effect in the test. Decreasing the point-level α_{point} (increase the point-level threshold) would avoid this problem effectively. Empirically, the guideline for point-level threshold selection is given for cluster-level $\alpha_{cluster}$=0.05 and 0.01 with different sample size. Both the simulation and real data show the proposed guideline can improve the sensitivity of cluster-based permutation test for detecting small clusters while retaining the same FWER.

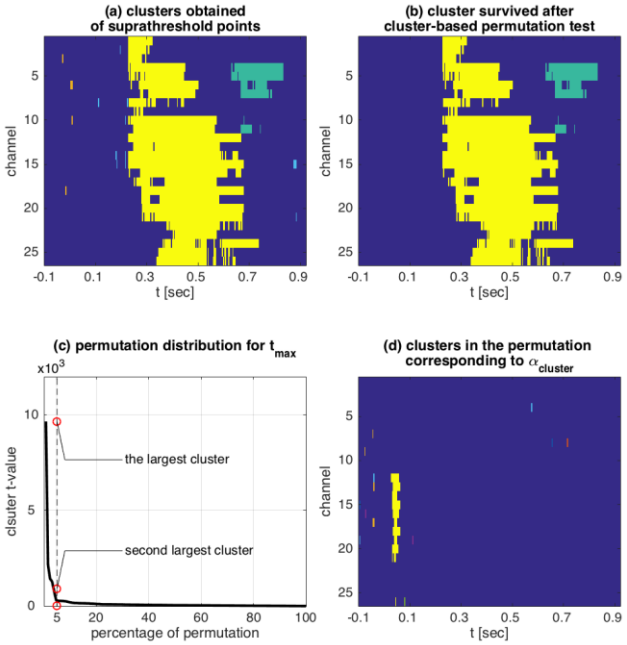

Figure 6. The procedure for cluster-based permutation test with α_{point}=0.0112. (a) clusters obtained from temporal-spatial adjacent points with their t_{point} above point-level threshold, (b) the largest two clusters survived in result, (c) permutation distribution for t_{max}, (d) the clusters in the 7th permutation of the distribution, the largest one (the yellow cluster) determines the cluster-level threshold.

References

[1] T.C. Handy, "Event-related potentials: A methods handbook," MIT press, 2005.

[2] O. J. Dunn, "Multiple comparisons among means," J. Am. Stat. Assoc., vol. 56, no. 293, pp. 52–64, 1961.

[3] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," J. R. Stat. Soc. Ser. B, pp. 289–300, 1995.

[4] Y. Benjamini, A. M. Krieger, and D. Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," Biometrika, vol. 93, no. 3, pp. 491–507, 2006.

[5] R. C. Blair and W. Karniski, "An alternative method for significance testing of waveform difference potentials," Psychophysiology, vol. 30, no. 5, pp. 518–524, 1993.

[6] E. Maris and R. Oostenveld, "Nonparametric statistical testing of EEG-and MEG-data," J. Neurosci. Methods, vol. 164, no. 1, pp. 177–190, 2007.

[7] D. M. Groppe, T. P. Urbach, and M. Kutas, "Mass univariate analysis of event-related brain potentials/fields II: Simulation studies," Psychophysiology, vol. 48, no. 12, pp. 1726–1737, 2011.

[8] D. M. Groppe, T. P. Urbach, and M. Kutas, "Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review," Psychophysiology, vol. 48, no. 12, pp. 1711–1725, 2011.

[9] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: a primer with examples," Hum. Brain Mapp., vol. 15, no. 1, pp. 1–25, 2002.

[10] A. P. Holmes, R. C. Blair, J. D. G. Watson, and I. Ford, "Nonparametric analysis of statistic images from functional mapping experiments," J. Cereb. Blood Flow Metab., vol. 16, no. 1, pp. 7–22, 1996.

[11] D. M. Groppe, S. Makeig, and M. Kutas, "Identifying reliable independent components via split-half comparisons," Neuroimage, vol. 45, no. 4, pp. 1199–1211, 2009.