Spatiotemporal Localization of Significant Activation in MEG using Permutation Tests[1]

Dimitrios Pantazis[1], Thomas E. Nichols[2], Sylvain Baillet[3] and Richard M. Leahy[1]

[1] Signal & Image Processing Institute, University of Southern California,
Los Angeles, CA 90089-2564, USA
{pantazis, leahy}@sipi.usc.edu
[2] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA
nichols@umich.edu
[3] Neurosciences Cognitives & Imagerie Cerebrale CNRS UPR640-LENA,
Hospital de la Salpetriere, Paris, France
sylvain.baillet@chups.jussieu.fr

Abstract. We describe the use of non-parametric permutation tests to detect activation in cortically-constrained maps of current density computed from MEG data. The methods are applicable to any inverse imaging method that maps event-related MEG to a coregistered cortical surface. To determine an appropriate threshold to apply to statistics computed from these maps, it is important to control for the multiple testing problem associated with testing 10's of thousands of hypotheses (one per surface element). By randomly permuting pre- and post-stimulus data from the collection of individual epochs in an event related study, we develop thresholds that control the familywise (type 1) error rate. These thresholds are based on the distribution of the maximum intensity, which implicitly accounts for spatial and temporal correlation in the cortical maps. We demonstrate the method in application to simulated data and experimental data from a somatosensory evoked response study.

1 Introduction

Cortically constrained spatio-temporal maps of neural activity can be computed from event related MEG data using linear inverse methods to estimate source current densities within pyramidal cells in the cortex. One of the most commonly used approaches extracts a representation of the cerebral cortex from a coregistered MR image, tessellates the result, and solves a linear inverse problem for elemental sources located at each of the vertices of the tessellated surface. The problem is hugely underdetermined, so that regularization methods are typically used [1,2]. The resulting current density maps (CDMs) are in general low resolution; interpretation is further confounded by the presence of additive noise exhibiting strong spatial correlation. As

[1] This work was supported by grant R01 EB002010 from the National Institute of Biomedical Imaging and Bioengineering.

with fMRI images, objective assessment of CDMs requires a principled approach to identifying regions of significant activation.

Dale et al. [2] normalize the CDMs using an estimate of the background noise variance at each cortical element. These normalized images follow a t-distribution under the null hypothesis of Gaussian background noise. Thresholding the images will produce maps of significant activation. However, testing at each surface element gives rise to the multiple comparisons problem: if significance is set at the p=.05 level, for example, then statistically 5% of the surface elements will give false positives. We wish to determine a threshold to achieve the desired family-wise error rate (FWER). The simplest solution is the Bonferroni correction which scales the p-value by the number of tests performed (i.e. number of surface elements). This is of little practical value in neuroimaging experiments since, due to strong spatial dependence, it is very conservative. The most widely used methods in analysis of neuroimaging data use random field theory and make inferences based on the maximum distribution. The maximum plays an essential role in controlling FWER. Consider a statistic image T_i, thresholded at u; if the null hypothesis is true everywhere, then the FWER is

$$\begin{aligned} P(FWE) &= P(\cup_i \{T_i > u\}) \\ &= P(\max_i T_i > u) \\ &= 1 - F_{\max T}(u). \end{aligned} \quad (\mathbf{1})$$

That is, a familywise error occurs when one or more T_i are above the threshold *u*, but this can only occur when the maximum of the T_i is above *u*. Hence, to control the FWER at level α, one needs to find the $(1-\alpha)100^{\text{th}}$ percentile of the maximum distribution,

$$u = F_{\max T}^{-1}(1-\alpha). \quad (\mathbf{2})$$

The random field methods proceed by fitting a general linear model to the data. The parameters of this model are estimated and then contrasted (using t-tests, F-tests, paired t-tests, ANOVA or others) to produce a statistic image. In this framework, a closed form approximation for the tail of $F_{\max T}$ is available, based on the expected value of the Euler characteristic of the thresholded image.

The parametric framework is valid for PET and smoothed fMRI data. However, the assumptions for the p-value local maxima and the size of the suprathreshold clusters do not hold directly for MEG data because of spatially variant noise correlation on the cortical surface. One solution to this problem is to use a transformation that warps or flattens the image into a space where the data are isotropic [3]. This approach can be applied directly to MEG, for example to determine an appropriate threshold for the noise-weighted maps described in Dale et al [2]. An application of this framework to MEG is described in [4] but the method is specifically tailored to beamforming methods rather than the linear inverse methods of interest here. These parametric random field methods require the usual parametric assumption of normality at each spatial location, in addition to random field assumptions of a point spread function with two derivatives at the origin, sufficient smoothness to justify the appli-

cation of the continuous random field theory, and a sufficiently high threshold for the asymptotic results to be accurate.

Non-parametric methods rely on minimal assumptions, deal with the multiple comparisons problem and can be applied when the assumptions of the parametric approach are untenable. They have also outperformed the parametric approaches in the case of low degrees-of-freedom t images [5]. Non-parametric permutation tests have been applied in a range of functional imaging applications [5,6,7,8]. Permutation tests are attractive for the application to MEG data since they are exact, distribution free and adaptive to underlying correlation patterns in the data. Further, they are conceptually straightforward and, with recent improvements in desktop computing power, are computationally tractable. Blair et al [6] describe an application of this approach to analysis of EEG data as recorded at an array of electrodes; in contrast the work presented here is applied to inverse solutions in which the maps are estimates of cortical activation.

2 Method

Our goal is to detect spatial and temporal regions of significant activity in MEG-based cortical maps while controlling for the risk of any false positives. We find global or local thresholds on statistics computed from the cortical maps that control the FWER . The method is introduced in a general framework to demonstrate its flexibility and adaptability to different experiments; we then describe the specific tests used in our experimental studies.

2.1 Permutation Approach

We assume that MEG data are collected as a set of N stimulus-locked event-related epochs (one per stimulus repetition) each consisting of a pre- and post-stim interval of equal length. Each epoch consists of an array of data representing the measured magnetic field at each sensor as a function of time. A cortical map is computed by averaging over all N epochs and applying a linear inverse method to produce an estimate of the temporal activity at each surface element in cortex. Our goal is to detect the locations and times at which activity during the post-stim experiment period differs significantly from the background pre-stim period. The method, as described below, can be readily extended to address more complex questions involving multiple factors.

To apply the permutation test, we must find permutations of the data that satisfy an exchangeability condition, i.e. permutations that leave the distribution of the statistic of interest unaltered under the null hypothesis. Permutations in space and time are not useful for these applications because of spatio-temporal dependence of the noise. Instead we rely on the exchangeability of the pre- and post-stimulus data for each epoch. Given N original epochs, we can create $M \leq 2^N$ permutation samples, each consisting of N new epochs. Since the inverse operator is linear, we can equivalently apply the inverse before or after averaging the permuted epochs. Consequently, we

describe the permutation tests in terms of permutation of the images formed from individual epochs, although in practice it is more computationally efficient to average the permuted data before applying the inverse operator.

Our modeling proceeds by succesively summarizing the information contained in the current density maps as illustrated in Figure 1. Current density maps for each epoch are denoted $Y_{ijk}(t)$ with t the time index, *i* the spatial index, *j* the permutation index, and *k* the epoch index, with $j = 0$ representing the original non-permuted data. We first summarize the data over epochs, finding the average effect $E_{ij}(t)$ of all epochs at each time point and spatial location. Then we summarize the data over time, creating an image T_{ij} of the effect of interest. Finally we summarize over space to gauge the overall effect of the experiment, S_j:

$$E_{ij}(t) = \text{summary statistic}_k\{Y_{ijk}(t)\} \quad \textbf{(2)}$$

$$T_{ij} = \text{summary statistic}_t\{E_{ij}(t)\} \quad \textbf{(3)}$$

$$S_j = \text{summary statistic}_i\{T_{ij}\} \quad \textbf{(4)}$$

Appropriate summary statistics include mean, mean absolute value, mean squared value, and absolute maximum value. Due to the nonparametric nature of the test, any test statistic can be used. However, as noted above, the maximum statistic captures the necessary information to control the FWER. Put another way, using the maximum statistic, we can return and make inferences in this dimension using the empirical maximum distribution.

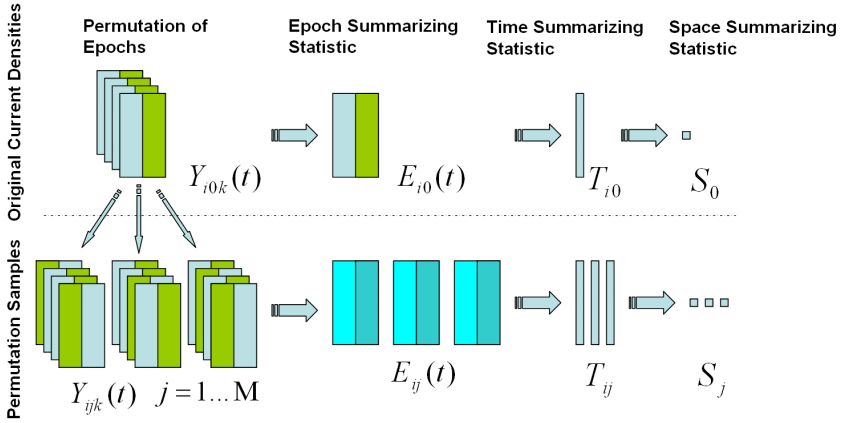

Fig. 1. Illustration of the summarizing procedure used to construct empirical distributions from the permuted data: M permutation samples $Y_{ijk}(t)$ are produced from the original data $Y_{i0k}(t)$. The data are then summarized successively in epochs, time and space according to equ (2)-(4) respectively, to produce S_j. The empirical distribution of S_j can be used to draw statistical inferences for the original data.

If we were interested in making inferences among epochs, such as looking for habituation effects, we would have to use a maximum statistic in equ (2). However, in our case we will assume no structured experimental variation among epochs and use the average, which is also consistent with the standard procedure for analyzing event related MEG data. For the time summarizing statistic in equ (3) we use the maximum over all post-stimulus data. This allows us to maintain resolution in the time domain and later check the temporal activation profile of the sources. Finally, using a maximum summarizing statistic in equ (4) to compute S_j allows us to retain spatial as well as temporal resolution.

After summarizing all data with respect to epochs, time and space, we can use the distribution of the S_j statistic to define a global threshold that controls the FWER, i.e. if we pick a global threshold with p-value equal to 0.05 with respect to the distribution of S_j, we have a 5% probability of one or more false positives throughout the entire spatio-temporal data set. We can then use this value to threshold the image at each point in time at each surface element to determine those regions for which we can reject the null hypothesis and hence detect significant activation.

2.2 Achieving Uniform Sensitivity

Permutation tests are always valid given the assumption of exchangeability under the null hypothesis. However, if the null distribution varies across space or time, there will be uneven sensitivity in that dimension. For example, with a maximum statistic over space, surface elements for which background noise variance is high will contribute more to the maximum distribution than others with low noise; the impact is a relatively generous threshold for the high-noise variance locations and a stringent threshold for the other locations. We can overcome this problem by including some form of normalization in the summary statistic. Thus, before computing the maximum statistic T_{ij}, we first normalize the data at each surface element by the sample standard deviation at that element computed from the pre-stim data (this is equivalent to the noise normalization performed in [2], except we measure our noise in the surface element domain, instead of the detector domain). We assume homogeneous variance over time, so do not perform any normalization in this dimension.

Under the assumption that the data are Gaussian, the noise normalization converts the data, under the null hypothesis, to a t-distribution. In this case, the permutation test will yield uniform spatial sensitivity. However, if the data are non-Gaussian, then simply normalizing by the standard deviation may not be sufficient for this purpose (Fig. 2). An alternative is to normalize based on the p-values themselves, i.e. at each spatial location we compute the empirical distribution across permutations and then replace the statistic T_{ij} for each permutation sample with its p-value. The p-value at surface element *i* for permutation *j*, called T_{ij}^p, is defined by:

$$T_{ij}^p = p_i(T_{ij}), \quad p_i(t) = \frac{1}{M}\sum_j H(T_{ij} - t), \quad H(x) = \begin{cases} 1 & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases} \qquad (5)$$

where $p_i(t)$ is the p-value function for surface element *i*, the proportion of permutations as large or larger than *t*. For each *i*, $\{T_{ij}^p\}$ has a uniform distribution under the null hypothesis, and hence is normalized. We next compute the summary statistic as the distribution of the minimum of T_{ij}^p for each surface element over the entire cortical surface (minimum p-value plays the same role as the maximum statistic in FWER). From this we compute the threshold on the T_{ij}^p values to achieve the desired FWER, and from this compute the corresponding threshold to apply at each individual spatial location.

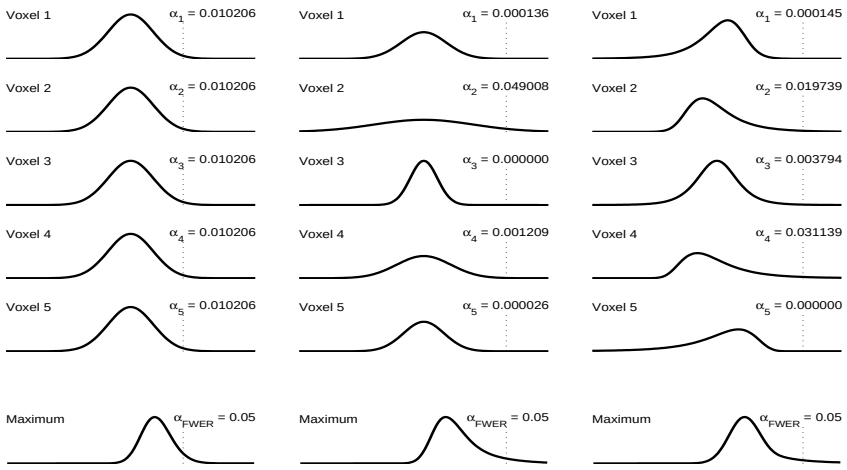

Fig. 2. Illustration of the impact of heterogeneous voxel null distributions on a 5% FWER threshold. Shown are null distributions of 5 surface elements in three cases: all sharing the same normal distribution, each having different variances and each having different skewed distributions. The first case (left) shows that with homogenous nulls the false positive rate at each surface element is homogeneous. The second case (middle) demonstrates the variable false positive rate when test statistics are not normalized (e.g. raw CDM values, $E_{ij}(t)$). The last case demonstrates the impact of non-Gaussianity, even when variance is normalized, and motivates the use of p-values to normalize T_{ij}. Note that in all cases FWER is controlled at 5%.

One practical problem with this approach is the discreteness of the p-values T_{ij}^p, which in turn causes S_j to be discrete. If many T_{ij}^p have the smallest possible value (1/M), then small α-levels for S_j may be unattainable. For example, one Monte Carlo experiment with M=1,000 found that 30% of the permutations had a minimum T_{ij}^p of value 0.001 and hence the smallest possible FWER threshold corresponded to $\alpha = 0.3$. Therefore, this p-value normalization approach, while makes no assumptions on differing shapes of the local distributions, requires many permutations.

2.3 Two Detection Methods

We have described above the procedure we use for generating the summary statistics from which we compute thresholds to detect significant activation, as well as the available normalization procedures. The two methods we will examine further in our simulations are summarized in Table 1. Both methods use the mean statistic to summarize epochs, as well as maximal statistics to summarize in time and space. However, method 1 does not normalize the time-summarizing T_{ij}, while method 2 transforms T_{ij} into p-values, essentially normalizing T_{ij} with the local permutation distribution. They subsequently use the maximum (method 1) or minimum (method 2) to summarize space. We can then use the empirical distribution of the space-summarizing statistic S_j to define a global threshold S^{th} that achieves a 5% FWER.

Table 1. Summary statistics and normalization schemes for the detection methods

	Epoch-Summarizing $E_{ij}(t)$	$E_{ij}(t)$ Normalized $E_{ij}^n(t)$	Time-Summarizing T_{ij}	T_{ij} Normalized T_{ij}^n	Space-Summarizing S_j
Method 1	$mean_k\{Y_{ijk}(t)\}$	$E_{ij}(t)/s_{ij}$	$\max_t\{\mid E_{ij}^n(t)\mid\}$	T_{ij}	$\max_i\{T_{ij}^n\}$
Method 2	$mean_k\{Y_{ijk}(t)\}$	$E_{ij}(t)/s_{ij}$	$\max_t\{\mid E_{ij}^n(t)\mid\}$	$p_i(T_{ij})$	$\min_i\{T_{ij}^n\}$

The process for testing the original data against S^{th} is as follows. For method 1 the threshold can be applied directly to the normalized CDM's, $E_{i0}^n(t)$; any source with $E_{i0}^n(t) \geq S^{th}$ at any time can be declared significant. For method 2 the test S^{th} has units of p-values and cannot be directly applied to $E_{i0}^n(t)$. Moreover, the p-values were computed separately for each source *i*, so the same p-value at different sources will correspond to different values of $E_{i0}^n(t)$. Method 2's variable thresholds are found with the inverse p-value transformation, where source *i* at time *t* is significant if $E_{i0}^n(t) \geq p_i^{-1}(S^{th})$.

3 Simulation Studies

In this section we present simulation results to evaluate the two methods summarized in Table 1. A cortical surface was extracted from an MRI scan using BrainSuite, a brain surface extraction tool [9] and coregistered to the MEG sensor arrangement of a CTF Systems Inc. Omega 151 system. The original surface contained approximately 520,000 faces and was down-sampled to produce a 15,000 face (7481 vertices) surface suitable for reconstruction purposes. Further, the original surface was smoothed to assist easy visualization of CDMs. An orientation constraint was applied to the reconstruction method using surface normals estimated from the original dense cortical surface. The forward model was calculated using overlapping spheres. The inverse matrix H was regularized using the Tikhonov method with $\lambda = 4 \cdot 10^{-7}$.

3.1 Source Simulation Experiment

We simulated two sources on the left and right hemisphere of the brain, as shown in Figure 3. The timecourses represent an early and a delayed response to a stimulus (Fig. 4).

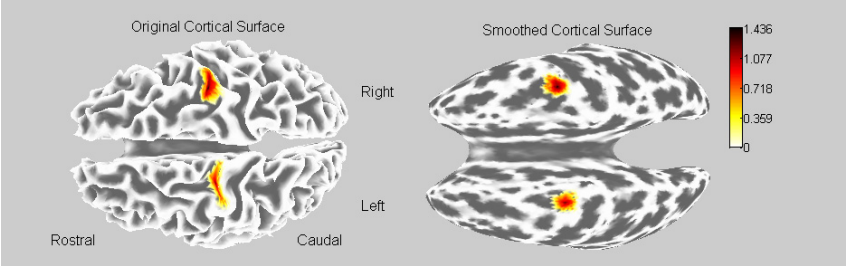

Fig. 3. Source 1 (left) and source 2 (right) are shown on the original and smoothed version of a cortical surface.

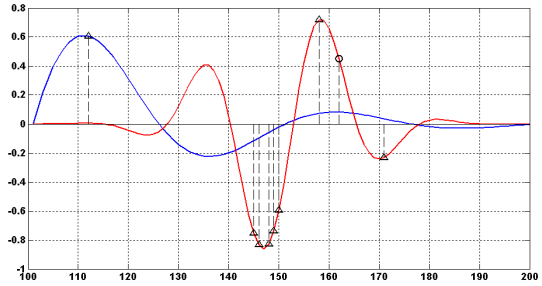

Fig. 4. Timecourse of simulated sources and points of source identification at α=0.123. Triangles for both methods, circles for only method 1

A total number of 100 epochs were generated, each consisting of 100 pre-stimulus and 100 post-stimulus time points. Gaussian i.i.d. noise with power 2000 times the average signal power was added to the channel measurements. The epochs were resampled producing $M = 1000$ permutation samples for method 1 and $M = 10000$ for method 2 (the larger number of resamples required for method 2 was due to the discreteness of the p-values as discussed in Section 2.2). We then applied the inverse operator H to all data, to produce CDMs $Y_{ijk}(t)$. All further processing for the extraction of the empirical distribution of S_j is summarized in Table 1.

The global α=0.05 threshold S^{th} for method 1 was $S_1^{th} = 5.236$. Due to the discreteness problem, the smallest possible threshold with method 2 was α=0.123; for this level, $S_1^{th} = 5.0503 \cdot 10^{-4}$ and $S_2^{th} = 0.0001$; note that the first is a threshold on maximum statistics while the second is a p-value threshold. We applied these thresholds to the original data as described in Sect. 2.3. For α=0.05, method 1 identified 6 time points as containing significant activations. For α=0.123 method 1 identified 9

and method 2 identified 8 time instances as containing significant activations (Fig. 4). Note that both methods successfully identify regions with either source 1 or source 2 active. Importantly, neither give any false positives in regions where there is no source.

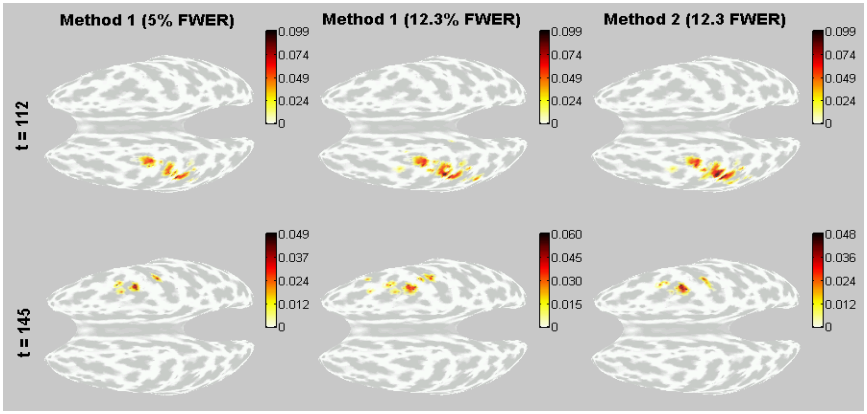

Fig. 5. Examples of significant activation maps for method 1 and 2 for two time instances. Reconstruction appears spread on the smooth cortical surface, but active sources are in neighboring sulci in the original cortical surface. The lowest achieved FWER for method 2 is α=0.123

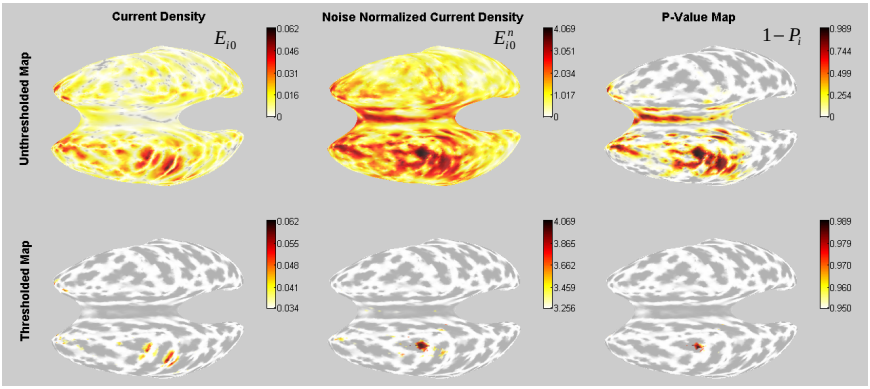

Fig. 6. Thresholded and Unthresholded maps of the current density (E_{i0}), the noise normalized current density (E^n_{i0}) and (1-p)-value map at $t = 113$. The E_{i0} and E^n_{i0} maps are thresholded subjectively while the (1-p) value map is thresholded at a p=.05 for each source.

Figure 5 shows that method 1 and method 2 produce very similar results. In simulation, this is expected since the noise is Gaussian. We should comment here that permutation tests do not address the limited resolution of MEG reconstruction. All

inverse methods are ill-posed and CDMs tend to mislocalize source activation. If the inverse method demonstrates experimental variation in some regions, permutation tests will identify these regions regardless of the presence of an actual source there.

We can display the unthresholded p-value maps of method 2, transforming the CDMs of the original data into p-values. Even though this does not address the multiple comparisons problem, it is interesting to compare the achieved localization of CDMs, noise-normalized CDMs and p-value maps. Such a result is given in Figure 6.

3.2 Noise Simulation Experiment

In order to test both methods for specificity, we applied permutation tests using noise-only data. We estimated the thresholds for method 1 using standard Gaussian noise; we did not evaluate method 2 due to the discreteness problem. Then, we created 100 measurements, each consisting of 100 epochs. The epochs had 100 pre- and 100 post-stimulus time points. We tested these data for significant activation, keeping in mind that the approximate Monte Carlo standard error for a true 0.05 rejection rate is 2.2; hence we expect 5+/-2 false positives. Method 1 exhibited false positives only 6 out of 100 times, consistent with being an exact test.

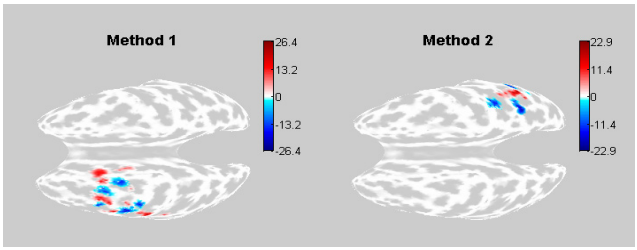

Fig. 7. Examples of false positives for method 1 and 2. Due to high correlation, false positive sources are in neighboring areas on the cortical surface

4 Real Data Experiment

The effectiveness of the proposed algorithm was evaluated using data from a real somatosensory experiment. The data acquisition was done using a CTF Systems Inc. Omega 151 system. The somatosensory stimulation was an electrical square-wave pulse delivered randomly to the thumb, index, middle and little finger of each hand of a healthy right-handed subject. For the purposes of the current experiment, only data from the right thumb where tested for reconstruction.

This experiment demonstrated that method 2 is more sensitive than method 1. Also, the discrepancies in the significant activation maps is an indication that the data are not Gaussian, as they were in the simulation experiments. As shown in Figure 8, in $t = 22ms$ only method 2 detected significant activity. Further, it seems to correct the CDM, which shows the main activity in the ipsilateral hemisphere. Significant

activation in the left somatosensory cortex is expected, as the experiment involved stimulation of the right thumb, so method 2 produces reasonable results. For $t = 28ms$ the same remarks for sensitivity are true. Figure 9 shows the thresholds applied by each method. Again, due to discreteness, the lowest achieved FWER by method 2 is α=0.086.

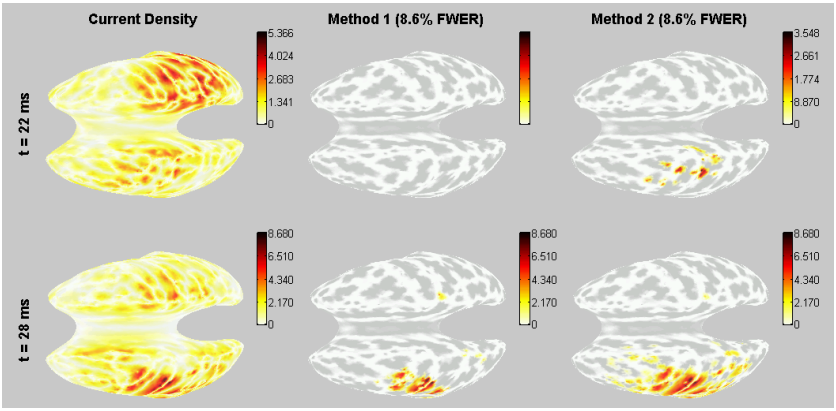

Fig. 8. Reconstruction and Significant maps from methods 1 and 2 for two time instances. All maps are scaled by 10^{12}

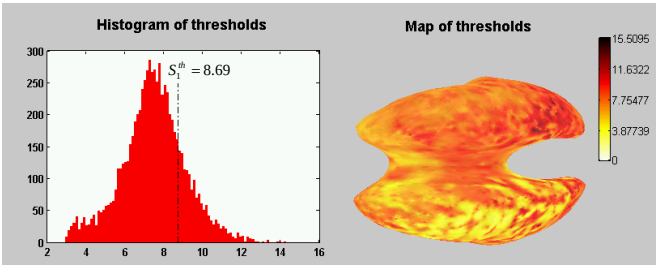

Fig. 9. Global threshold applied by method 1 (S_1^{th}) at level α=0.086, as compared to the histogram of the thresholds applied to each source by method 2. Also, a map of the thresholds on the cortical surface is given on the right. Most of the individual thresholds are below S_1^{th}

5 Conclusion

We have presented a method to apply permutation tests for processing of MEG data and extracting maps of significant brain activation. It can be combined with any inverse imaging method and is flexible in terms of available statistics and normalization procedures. The method is exact (i.e. it achieves the specified FWER) providing confidence that activation is present in the cortical regions that do test as significant.

One limitation of the method is that the pre- and post-stimulus size of the data should be the same for the permutation scheme to work; we will be considering bootstrap alternatives to avoid this requirement. Also, this work does not address the limited resolution of the inverse methods in MEG. If the CDMs demonstrate experimental variation in some regions, permutation tests, or indeed other tests based on the rejection of H_0, will identify these regions regardless of the presence of an actual source at that location. Thus there is the potential to detect significant brain activation but for the sites to be misplaced relative to true activation area. It is important to take this effect into account when interpreting maps of cortical activation derived from MEG data.

Acknowledgements

We thank Sabine Meunier for providing the experimental data.

References

1. Phillips, J.W.; Leahy, R.M.; Mosher, J.C.: MEG-Based Imaging of Focal Neuronal Current Sources, IEEE Transactions of Medical Imaging, Vol. 163, June (1997) 338-348
2. Dale, A. M., Liu, A. K., Fischi, R. B., Buckner, R. L., Belliveau, J. W., Lewine, J. D., Halgren, E., Dynamic Statistical Parametric Mapping: Combining fMRI and MEG for High-Resolution Imaging of Cortical Activity. Neuron, Vol. 26 (2000) 55-67
3. Worsley, K. J., Andermann, M., Koulis, T., MacDonald, D., Evans, A. C.: Detecting Changes in Nonisotropic Images, Human Brain Mapping 8 (1999) 98-101
4. Barnes G. R. Hillebrand, A.: Statistical Flattening of MEG Beamformer Images, Human Brain Mapping 18 (2003) 1-12
5. Nichols, T. E., Holmes, A. P.: Nonparametric Permutation Tests For Functional Neuroimaging: A Primer with Examples, Human Brain Mapping 15 (2001) 1-25
6. Blair, R. C., and Karnisky, W.: Distribution-Free Statistical Analyses of Surface and Volumetric Maps. Functional Neuroimaging: Technical Foundations, Academic Press, San Diego, California, (ed.) Thatcher, R. W., Hallett, M., Roy, J. E., Huerta, M. (1994)
7. Arndt, S., Cizadlo, T., Andreasen, N.C., Heckel, D., Gold, S., and O'Leary, D.S.: Tests for comparing images based on randomization and permutation methods, Journal of Cerebral Blood Flow and Metabolism, 16, (1996) 1271-1279
8. Holmes, A.P., Blair, R.C., Watson, J.D.G., and Ford, I.: Nonparametric analysis of statistic images from functional mapping experiements, Journal of Cerebral Blood Flow and Metabolism, 16, (1996) 7-22
9. Shattuck, D. W., Leahy, R. M.: BrainSuite: An Automated Cortical Surface Identification Tool. Medical Image Analysis, 6 (2) (2002) 129-142