Abstract

M. Wagner, C. Ponton, R. Tech, M. Fuchs, & J. Kastner (Hamburg, Germany & Charlotte, USA) Non-Parametric Statistical Analysis of EEG/MEG Map Topographies and Source Distributions on the Epoch Level

In Event-Related Potential and Event-Related Field experiments, stimuli – often of several different types – are presented repeatedly, and the subject's brain response is recorded using Electroencephalography (EEG) or, in the ERF case, Magnetoencephalography (MEG). After removing artifacts and epoching the data, many repetitions per stimulus type are available, which are later usually averaged and compared. At this stage, though, it is no longer possible to establish whether and for which latencies the averaged waveforms are significantly different between stimulus types, nor whether the epochs for a given stimulus type yield significant averages in the first place. A statistical analysis of all individual epochs can provide exactly this information. Topographic Analysis of Variance (TANOVA) and Statistical non-Parametric Mapping performed on the results of Current Density Reconstructions (CDR SnPM) are non-parametric permutation or randomization tests which have previously been published but mainly been used to process per-subject averaged EEG data in the context of group studies. This paper describes how to apply TANOVA and CDR SnPM to individual epochs on a sample-by-sample basis, even in the context of single-subject data. A multiple comparison correction approach for the analysis of subsequent samples based on spectral properties of the data is presented. Methods are demonstrated using filtered and unfiltered simulated dipole data and data from a Continuous Performance Task (CPT) EEG experiment eliciting Mismatch Negativity. While TANOVA is able to identify latencies of significantly different map topographies, CDR SnPM extracts – per latency – the locations of significant source activation differences between stimulus types, albeit at the price of reduced overall sensitivity. Using simulated data, the proposed multiple comparison correction approach is illustrated. Significant peaks and source locations obtained for the CPT data are consistent with existing knowledge.

Keywords: Electroencephalography, Magnetoencephalography, Event-Related Fields, Event-Related Potentials, Continuous Performance Task, Statistical Analysis, Randomization Statistics, Non-Parametrical Statistics, Topographical Analysis of Variance, Current Density Analysis, Statistical non-Parametric Mapping

M. Wagner, C. Ponton, R. Tech, M. Fuchs, & J. Kastner — Non-Parametric Statistical Analysis of EEG/MEG Map Topographies and Source Distributions on the Epoch Level

M. Wagner[1], C. Ponton[2], R. Tech[1], M. Fuchs[1], & J. Kastner[1]
[1]Compumedics Germany GmbH,
Heußweg 25, 20255 Hamburg, Germany,
[2]Compumedics USA Inc., 6605 West WT Harris Blvd, Charlotte, NC 28269, USA
mwagner@neuroscan.com

Introduction

In an Event-Related Potential (ERP) or Event-Related Field (ERF) experiment, an Electroencephalography (EEG) or Magnetoencephalography (MEG) device records the brain response related to a sensory, cognitive, or motor event. Depending on the experimental design, events (stimuli or responses) may be of the same or of different types. Data segments with distortions such as ocular, cardiac, or muscle artifacts are later detected and artifacts are either reduced or excluded from further processing. After splitting the data into epochs time-locked to events, many repetitions per event type are available and usually averaged and compared. After averaging, though, it is no longer possible to establish whether and for which latencies the averaged waveforms differ significantly between event types, nor whether the trials (epochs) of a given type yield significant averages in the first place. A statistical analysis of all individual epochs can provide this information.

Traditional statistical measures in channel space such as the *t*-test make disputable assumptions regarding repeatability and independence (Murray, Brunet, & Michel, 2008; Koenig & Melie-García, 2009). Therefore, a new non-parametric family of methods has recently attracted attention as it became computationally feasible for the analysis of ERP group studies (Murray et al., 2004). Although – misleadingly – referred to as Topographic Analysis of Variance (TANOVA, Koenig & Melie-García, 2010), no analysis of variance is being conducted, but rather a non-parametric permutation or randomization test. TANOVA is usually applied to per-subject averaged data in the context of group studies and yields similarities within and differences between groups of subjects.

If distributed source analysis methods such as Current Density Reconstruction (CDR) are used to localize the neural generators, the additional question of where in the brain significantly different source topographies occur may be asked. Statistical non-Parametric Mapping (SnPM) by non-parametric permutation or randomization tests using a maximum statistic to control the Family-Wise Error Rate (FWER) provides an assumption-free environment to answer this question (Nichols & Holmes, 2002). In the context of EEG, CDR SnPM is typically applied to source analysis results obtained for averaged data in the context of group studies and used to assess differences between groups of subjects (J. S. Kim, Han, Park, & Chung, 2008; Y. Y. Kim et al., 2009).

In this contribution, a framework is described that allows the application of the existing methods TANOVA and SnPM not only to individual averages in the context of an ERP/ERF group

study but to the individual epochs themselves (Wagner, 2014), something that is even possible for single-subject data. Unlike described in previous publications, the statistical analysis is conducted sample-by-sample as opposed to using a maximum statistic over all samples (Pantazis, Nichols, Baillet, & Leahy, 2003; R. E. Greenblatt & Pflieger, 2004), thus following the approach presented by (Koenig & Melie-García, 2009) for group data, but using a multiple comparison correction that is based on the spectral properties of the data. For CDR SnPM, in addition to the test for significant differences between conditions, a within-condition consistency test is proposed which can be used to justify testing for differences on a sample-by-sample basis. Standardized Low Resolution Electromagnetic Tomography (sLORETA) in a realistically shaped head model is employed for source localization, because it yields low localization error for focal activity, uniform spatial sensitivity, and is robust with respect to regularization (Pascual-Marqui, 2002; Wagner, Fuchs, & Kastner, 2004).

A simulation study is used to validate the implementation, and a visual Continuous Performance Task (CPT) EEG experiment eliciting Mismatch Negativity (MMN) is used to demonstrate the methods. The following Methods section is ordered by data acquisition and processing steps. However, the simulation study is described last, as it builds upon methods described previously.

Methods

Visual Continuous Performance Task Experiment

The numbers "1" and "2" were used as visual stimuli. In a visual CPT paradigm, "1" was used as the target stimulus and "2" as the distractor stimulus. The subject, a healthy adult, had EEG electrodes attached while watching the presentation of stimuli on a computer screen. 31 EEG electrodes were placed according to an extended 10-20 system with additional *FPz, FCz, CPz, Oz, FC3/4, CP3/4, FT7/8,* and *TP7/8* contacts (Fig. 1). The stimulus duration was 200 ms, with a randomized inter-stimulus interval (ISI) of between 800 ms and 1300 ms. 41 target and 165 distractor stimuli were presented in randomized order using the STIM system (Compumedics, Charlotte, NC, USA). The subject was instructed to press a button following the presentation of each target stimulus. EEG and VEOG data were recorded using a 32-channel Neuroscan system (Compumedics, Charlotte, NC, USA) with a sampling frequency of 250 Hz and a high-pass filter of 0.15 Hz.

Signal processing was performed in the Curry 7 software (Compumedics, Charlotte, NC, USA). Data were re-referenced to a Common Average Reference (CAR), because the subsequently applied statistical and source analysis methods require CAR data, and filtered using a 40 Hz low-pass filter. Eye blink effects were reduced using a regression analysis in combination with artifact averaging (Semlitsch et al., 1986). Data were epoched from 200 ms before to 500 ms after stimulus onset (Fig. 1). Epochs with signals exceeding ± 30 µV were excluded, since a visual inspection of the common average referenced epochs showed that signals of this magnitude were likely due to artifact. Averages for both stimulus types were

computed (Fig. 2). In the following, this dataset will be referred to as the low-pass filtered CPT data. In order to explore the effects of filtering on the statistics outcome, data were alternatively processed and epoched without any additional filtering besides the 0.15 Hz data acquisition filter, and with a high-pass of 1 Hz, 2 Hz, and 3 Hz in addition to the 40 Hz low-pass. To investigate statistics outcome for reduced epoch counts, derived versions of the low-pass filtered CPT data with only 75 %, 50 %, and 25 % of the total number of epochs were created. These three new, decimated datasets were obtained by excluding every fourth or every second epoch, or retaining only every fourth epoch, respectively.

Topographic Analysis of Variance

In the context of a TANOVA, two different non-parametric randomization tests were performed for all epochs: a consistency test per event type, and a test for differences between event types. Both tests were already described in (Koenig & Melie-García, 2010) and are summarized here for reference only.

The consistency test evaluates field topography (map) similarity across epochs. It is performed independently for each event type and each sample. Here, the Null Hypothesis is that epochs of the same event type are unrelated, i.e. that random maps have been measured. If the Null Hypothesis holds, randomly perturbing channels within each epoch's maps should not deteriorate the average map across all epochs.

For each sample s and Ec epochs of event type c, the test is performed as follows: First, the observed mean global field power (MGFP) $P_{s,c,0}$ of the average over all epochs e of the individual maps $d_{s,c,e}$ is computed as

$$P_{s,c,0} = \mathrm{mgfp}\left(\frac{1}{E_c}\sum_{e=1}^{E_c} d_{s,c,e}\right) \tag{1}$$

with

$$\mathrm{mgfp}(d) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(d_i - \frac{1}{M}\sum_{j=1}^{M} d_j\right)^2}$$

where M is the number of channels. Then, for a total of R repetitions, the channels within each map are randomly shuffled or perturbed. Typically, perturbation is used if the total number of perturbations is computationally feasible, while randomization is used in all other scenarios, including most real-world applications. For each repetition r, this yields new randomized maps $d_{s,c,e,r}$, and a new global field power $P_{s,c,r}$ can be computed according to

$$P_{s,c,r} = \mathrm{mgfp}\left(\frac{1}{E_c}\sum_{e=1}^{E_c} d_{s,c,e,r}\right). \tag{2}$$

The probability $p_{s,c}$ of the Null Hypothesis is the fraction of values $P_{s,c,r}$ that are larger than or equal to $P_{s,c,0}$. Small values of p, traditionally $p < 0.05$, indicate rejection of the Null Hypothesis, or consistency between epochs of the same event type. The number of possible permutations is $M!$ (Nichols & Holmes, 2002) and must be larger than the number of randomizations, which is typically the case for 8 and more channels.

The test for differences between event types is again performed independently for each sample. Here, the Null Hypothesis is that there is no difference between event types, i.e. that the same maps occur regardless of event type. If the Null Hypothesis holds, randomly perturbing maps across event types should not alter the average maps per event type.

When just two event types are compared, the MGFP of the difference of the averaged maps per

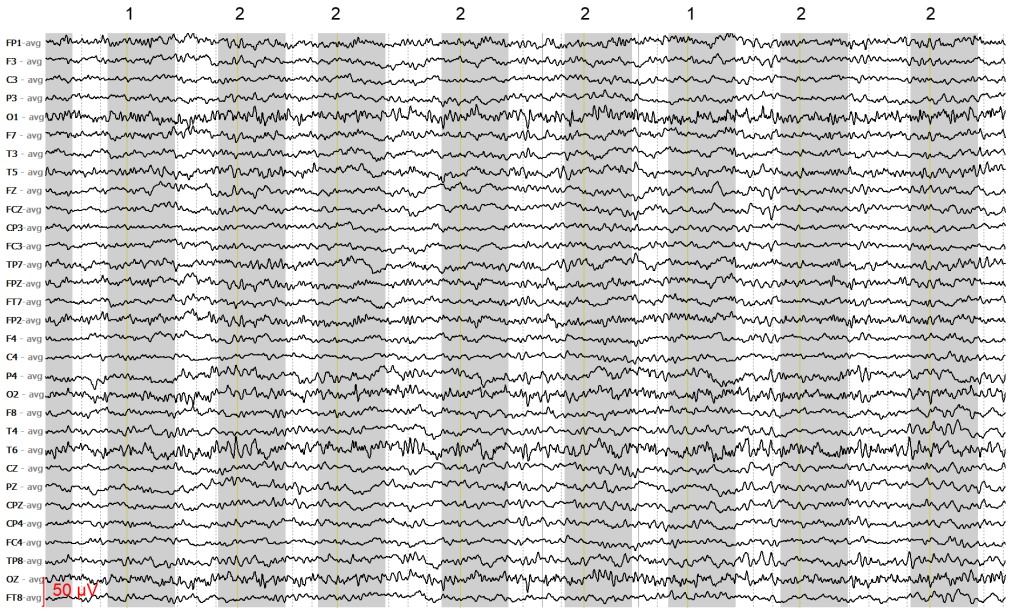


Figure 1: A 10 s page of ongoing EEG data re-referenced to CAR, with stimulus types at the top, where "1" stands for target and "2" represents distractor stimuli. Latency ranges marked in gray were used for epoching, from 200 ms pre- to 500 ms post-stimulus onset.

event type can serve as the measure. For each sample, the test is performed as follows: In a first step, the observed global field power $P_{s,0}$ of the difference of the averages over all epochs of event types $c = 1$ and $c = 2$ is computed as

$$P_{s,0} = \mathrm{mgfp}\left(\frac{1}{E_1}\sum_{e=1}^{E_1} d_{s,1,e} - \frac{1}{E_2}\sum_{e=1}^{E_2} d_{s,2,e}\right). \quad (3)$$

For R repetitions, maps are then randomly shuffled across event types. For each repetition r, randomized maps $d_{s,c,e,r}$ are obtained and the global field power $P_{s,r}$ can be computed according to

$$P_{s,r} = \mathrm{mgfp}\left(\frac{1}{E_1}\sum_{e=1}^{E_1} d_{s,1,e,r} - \frac{1}{E_2}\sum_{e=1}^{E_2} d_{s,2,e,r}\right). \quad (4)$$

Again, the probability p_s of the Null Hypothesis is the fraction of values $P_{s,r}$ that are larger than or equal to $P_{s,0}$. Small values of p indicate significant map differences between event types. Because an omnibus measure of map similarity has been used, no correction for multiple testing is necessary. The number of possible permutations, which again must be larger than the number of randomizations, is $(E_1+E_2)!/E_1!\cdot E_2!$ (Nichols & Holmes, 2002), which is typically the case for 8 and more epochs per type.

As both the consistency test and the difference test are performed sample by sample, false positives are to be expected. A test for the significance of consecutive rejections of the Null Hypothesis can establish, whether such periods of significance are significant themselves and is described in Koenig and Melie-García (2009).

Optionally, data may be collapsed across samples of interest to increase the Signal-to-Noise Ratio (SNR). Averaged maps may be normalized before computing the difference, in order to ignore absolute effect sizes. An extension to more

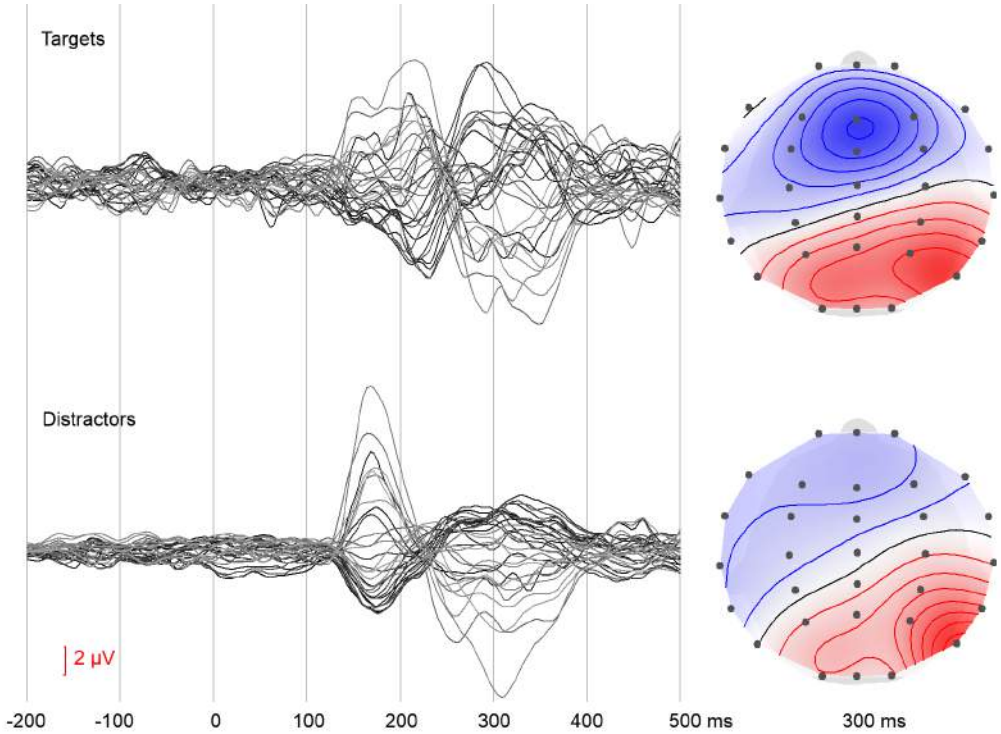


Figure 2: Butterfly plots of the average waveforms (left) and voltage topography maps for the 300 ms latency (right) for both stimulus types.

than two and to different categories of event types using a measure called global dissimilarity is described in Murray et al. (2008). TANOVA computation times scale linearly with the number of samples, number of randomizations, and number of channels.

TANOVA analysis was performed using the Curry software. For this paper, values of $p < 0.05$ were regarded as significant. As suggested by Manly (2006), the corresponding required number of repetitions was chosen to be $R = 50/p = 1000$. Map normalization was used for the difference tests, such that the MGFP per map was equal to 1. The complete time range from -200 ms to 500 ms was analyzed for the differently filtered and decimated data sets.

Source Analysis

A standardized Low Resolution Electromagnetic Tomography (sLORETA) analysis (Pascual-Marqui, 2002; Wagner et al., 2004) was performed for each sample of every epoch. As the head model, a realistically shaped three-compartment Boundary Element Method (BEM) model comprising 8043 nodes and 16074 triangles was used (Fuchs, Wagner, & Kastner, 2001). Conductivities were assumed to be 0.33, 0.0132, and 0.33 S/m for the skin, skull, and brain compartment, respectively. Source locations were distributed on a 7 mm regular grid throughout the brain but excluding the cerebellum, yielding a total of $N = 4786$ locations. The regularization parameter was determined

according to the discrepancy principle (Fuchs, Wagner, Köhler, & Wischmann, 1999) using an sLORETA analysis of the grand average of all epochs and subsequently kept fixed throughout the individual epoch analyses.

For the remainder of this paper, the resulting source strength distribution for sample s, event type c, and epoch e will be referred to as a source image $I_{s,c,e}$ and each source location will be referred to as a voxel n, with $i_{s,c,e,n}$ the source intensity at that particular voxel. In spite of this notation, the concept of a voxel readily generalizes to cases where sources are computed on the folded cortical surface only (Wagner, Fuchs, Wischmann, Ottenberg, & Dössel, 1995).

Statistical non-Parametric Mapping

The input data for SnPM can be CDR source images, but also beamformer results or voltage topographies (R. E. Greenblatt & Pflieger, 2004). In this case, sLORETA source images are used, which have the property to be normalized to the standard deviation per voxel (Pascual-Marqui, 2002). As a consequence, using sLORETA images as input data for SnPM yields uniform spatial sensitivity of the statistical test (Pantazis et al., 2003). As for TANOVA, a consistency test per event type and a test for differences between event types can be performed.

The consistency test evaluates source image similarity across epochs. It is performed independently for each event type and each sample. Here, the Null Hypothesis is that epochs of the same event type are unrelated, i.e. that random source images have been computed. If the Null Hypothesis holds, randomly perturbing voxels within each epoch's source image should not deteriorate the average source image across all epochs.

For each sample s and event type c, the test is performed as follows: First, for each voxel n, a t-value is obtained using a one-sample t-test. This t-value $t_{s,c,n,0}$ scores the hypothesis that the mean voxel intensity across all E_c source images is zero (a zero-centered distribution of the positive voxel intensities can be created by normalization and log-transformation, both of which are described further below):

$$t_{s,c,n,0} = \frac{\overline{i_{s,c,*,n}}}{\sigma(i_{s,c,*,n})/\sqrt{E_c}}. \tag{5}$$

Then, for a total of R repetitions, the voxels within each source image are randomly shuffled. For each repetition r, this yields new randomized source images $I_{s,c,e,r}$, and new t-values $t_{s,c,n,r}$ can be computed according to

$$t_{s,c,n,r} = \frac{\overline{i_{s,c,*,r,n}}}{\sigma(i_{s,c,*,rn,})/\sqrt{E_c}}. \tag{6}$$

The standard deviation (SD) σ used for computing the t-values in equations 5 and 6 is special in that it is additionally spatially smoothed as described in Nichols and Holmes (2002). This smoothing may be performed by simply taking the average across all voxels, or alternatively using a smoothing kernel. After randomization, a significance threshold $t_{s,c}$ is computed as the $(1-p)\cdot 100^{\text{th}}$ percentile across repetitions, based on the largest t-values across all voxels per repetition. This maximum t-statistic controls the FWER and is a means of multiple comparison correction across voxels (Westfall & Young, 1993). For all voxels with $t_{s,c,n,0} < t_{s,c}$ the Null Hypothesis is confirmed, while for all other voxels consistency across epochs has been established. To visualize the locations-of-consistency, a t-statistic image can be generated based on $t_{s,c,n,0}$ where values of t below the significance threshold are set to zero.

The test for differences between event types is again performed independently for each sample. Here, the Null Hypothesis is that there is no difference between event types, i.e. that the same source images occur regardless of event type. If the Null Hypothesis holds, randomly perturbing source images across event types should not alter the average source images per event type.

For each sample s, the test is performed as follows: First, an F-test is performed using a one-way Analysis of Variance (ANOVA) for each voxel n where the event types c are regarded as factors (Maxwell & Delaney, 2004). The F-value $F_{s,n,0}$ thus obtained measures the hypothesis that the voxel means of all E_c source images per event type are equal. For R repetitions, source images are then randomly shuffled across event types. For each repetition r, randomized source images $I_{s,c,e,r}$ are obtained and F-values $F_{s,n,r}$ can be computed per voxel. Next, a significance threshold F_s is computed as the $(1-p)\cdot 100^{\text{th}}$ percentile across repetitions, based on the largest F-values across all voxels per repetition (maximum F-statistic). For all voxels with $F_{s,n,0} < F_s$ the Null Hypothesis is confirmed, while for all other voxels it has been established that they are significantly different. To visualize the locations of significance, an F-statistic image can be generated based on $F_{s,n,0}$ where values of F below the significance threshold are set to zero. Because a global measure of source image difference has been used, no further correction for multiple testing across voxels is necessary. While the F-test per se is known to be non-robust against deviations from normality, in the context of SnPM it is only the ordering of, not the absolute F values, that determine significance.

Again, a test for the significance of consecutive rejections of the Null Hypothesis can be performed (Koenig & Melie-García, 2009). Optionally, data may be collapsed across samples of interest to increase the SNR. Source images may be normalized and/or log-transformed before entering the calculations. Normalization allows comparing relative as opposed to absolute source magnitudes. A log-transformation can make the distribution of the (always positive) voxel intensities more symmetric and – if voxel intensities have previously been normalized so that their sum-of-squares equals the number of voxels – zero-centered. Neither normalization nor log-transformation are strictly required for CDR SnPM, though, as non-parametric statistics per se are robust with respect to unknown or skewed distributions (Nichols & Holmes, 2002). An extension to different categories of event types is possible using ANOVA for multiple factors (Maxwell & Delaney, 2004). CDR SnPM computation times scale linearly with number of samples, number of randomizations, and number of source locations.

CDR SnPM analysis was performed using the Curry software. Again, values of $p < 0.05$ were regarded as significant, with $R = 1000$ the number of repetitions. Source distribution normalization and log-transformation were used and σ-averaging was applied. Again, all of the differently filtered and decimated data sets were processed.

Multiple Comparison Correction

Neither TANOVA nor CDR SnPM per se require a multiple comparison correction across sensors or voxels, because complete voltage topography maps are used for TANOVA and a maximum statistic is employed in CDR SnPM. However, both methods are performed for each sample and neighboring samples yield to multiple compar-

isons, if the spectral content of the data is impaired by low-pass filtering or otherwise limited.

To assess the number of independent comparisons n that occur due to low-pass filtering the data but still analyzing each sample, one should keep in mind that, according to the Nyquist–Shannon sampling theorem, after filtering using a cutoff frequency of f_c, data may be resampled at $2f_c$ without losing information. If the original sampling frequency is f_s, there are $f_s/2f_c$ ways to perform this resampling depending on which out of $f_s/2f_c$ samples is picked as the first sample. This ratio equals the number of comparisons n to consider when analyzing low-pass filtered data sample by sample:

$$n = f_s/2f_c. \tag{7}$$

The corresponding multiple comparison-corrected significance level α using the Šidák correction is

$$\alpha = 1 - (1 - \bar{\alpha})^{1/n} = 1 - (1 - \bar{\alpha})^{2f_c/f_s} \tag{8}$$

with $\bar{\alpha}$ the experiment-wide significance level.

As a consequence, for the analysis of the low-pass filtered data sets with $f_c = 40\,\text{Hz}$ and $\bar{\alpha} = 0.05$, a corrected significance threshold of $\alpha = 0.0163$ was used, including an adapted number of repetitions $R = 50/\alpha = 3071$.

Simulated Data

In order to test the statistical methods, an additional simulated dataset was created, using the same electrode layout, sampling rate, prestimulus and post-stimulus times as described above for the CPT experiment. This dataset comprised 100 epochs each of two different types. One epoch type ("dipole + noise") was created by simulating a current dipole source in the postcentral gyrus, 15 mm beneath the inner skull layer of the same realistic BEM head model as described above (Fig 3). Its dipole moment was modeled to be zero before 0 ms and linearly rise to 100 µA mm at 500 ms. The second epoch type ("noise") contained zero data. White Gaussian noise with a standard deviation of 10 µV was added to all epochs. A second version of this simulated data set was obtained by applying a low-pass filter with $f_c = 10\,\text{Hz}$. For the statistical analyses described above, values of $p < 0.05$ were regarded as significant, the number of repetitions was $R = 1000$, and a corrected significance threshold of $\alpha = 0.0041$ according to Eq. 8 was used for $f_c = 10\,\text{Hz}$, with the corresponding adapted number of repetitions $R = 12210$.

In order to independently assess the TANOVA results for this simulated dataset, the time-varying SNRs for the first principal component (Hastie, Tibshirani, & Friedman, 2009) of the averaged "dipole + noise" epochs was plotted and latencies with an $\text{SNR} < 1$ were marked as insignificant. The first principal component was chosen because it represents the simulated dipole topography in this case where the simulated dipole is clearly large enough to dominate the noise. Before applying Principal Component Analysis (PCA), data were SNR-transformed (whitened) by multiplication with the inverse standard deviation of the noise, estimated from the signal-free latencies before 0 ms (R. Greenblatt, 1995; Fuchs et al., 1998).

Results

For the averaged "dipole + noise" epochs of the simulated dataset, the estimated standard deviation of the noise was $\sigma_{\text{noise}} = 1\,\mu\text{V}$. For the low-pass filtered version, $\sigma_{\text{noise}} = 0.231\,\mu\text{V}$.

For the low-pass filtered CPT data set, after excluding epochs with signals exceeding $\pm 30\,\mu\text{V}$,

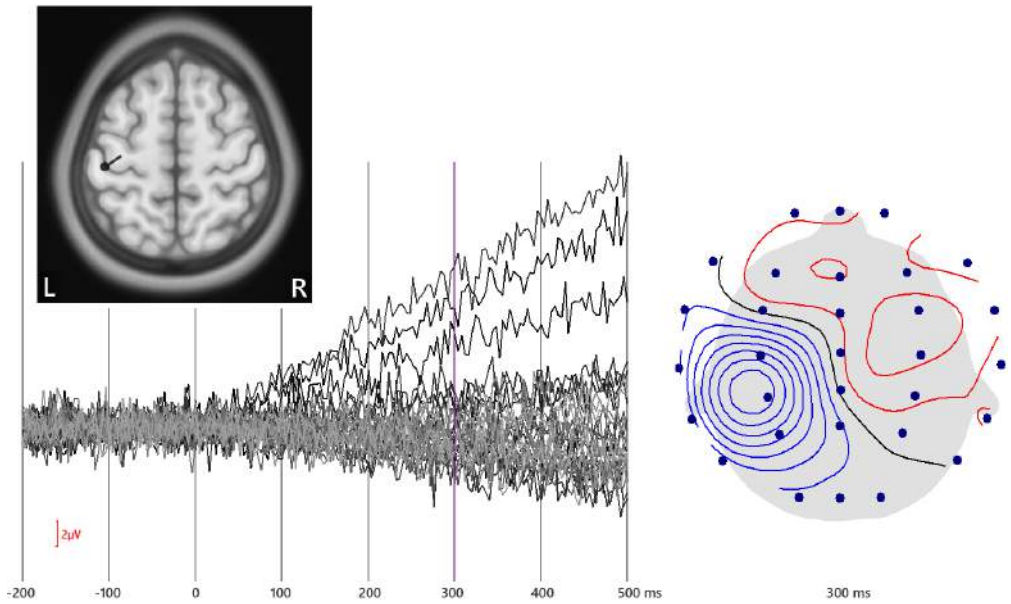


Figure 3: Simulated dipole location (top left), butterfly plot of the average "dipole + noise" waveforms (lower left) and voltage topography map for the 300 ms latency (right).

166 epochs remained: 34 of type "target" and 132 of type "distractor". Epoch counts for all other data sets derived from the CPT data are listed in the upper rows of Table 1. Data were subjected to TANOVA and sLORETA-based SnPM analyses.

Topographic Analysis of Variance

For the simulated dataset, the TANOVA consistency test (Fig. 4) established consistency of the "dipole+noise" epochs for all latencies after 84 ms, with additional, shorter periods of consistency from 60 to 64 ms and from 72 to 76 ms, as well as from -12 to -8 ms. Single significant samples failed the test for significance of contiguous rejections of the Null hypothesis, which in this case required a minimum of two samples. The "noise" epochs yielded no significant periods of consistency but a total of 9 isolated samples with $p < \alpha$. For comparison, the number of false positives to be expected is α times the number of samples which equals $0.05 \cdot 176 = 8.8$. The TANOVA test for differences between epoch types was satisfied from 92 ms onwards with the exception of the 129 ms sample. Again, two or more consecutively significant samples were required to establish significance of contiguous samples, and as a consequence 7 significant samples were rejected, including the 84 ms sample. The first principal component, when represented in SNR units, had an SNR ≥ 1 from 84 ms on.

In the case of the low-pass filtered CPT dataset, the TANOVA consistency test for the target stimuli yielded periods of consistency from -88 ms to -72 ms, from 60 ms to 68 ms and from 132 ms to 500 ms, with a total of 101 significant samples. For the distractors, consistency periods occurred from -140 ms to -112 ms, from -80 ms to -68 ms, from -56 ms to 120 ms, and from 132 ms to 500 ms. The total number of significant samples was 150. The test for differences between targets

Table 1: Dataset characteristics, significance levels, number of repetitions, and number of significant samples for TANOVA and CDR SnPM.

	Differently Filtered Datasets					Decimated No. Epochs		
Data Characteristics								
Low Cutoff [Hz]	0.15	0.15	1	2	3	0.15	0.15	0.15
High Cutoff [Hz]	125	40	40	40	40	40	40	40
Target No. Epochs	34	34	36	37	37	25	16	7
Distractor No. Epochs	131	132	154	158	158	100	67	35
Total No. Epochs	165	166	190	195	195	125	83	42
Statistics Parameters								
Significance Level α	0.05	0.0163	0.0163	0.0163	0.0163	0.0163	0.0163	0.0163
No. of Repetitions R	1000	3071	3071	3071	3071	3071	3071	3071
TANOVA	Samples $p < \alpha$							
Target	108	101	106	105	101	104	99	114
Distractor	156	150	137	122	116	132	115	87
Difference	67	57	64	66	65	64	55	35
CDR SnPM	Samples $p < \alpha$							
Target	147	139	162	168	161	117	73	45
Distractor	176	176	176	176	176	176	176	127
Difference	37	26	30	17	10	27	24	10

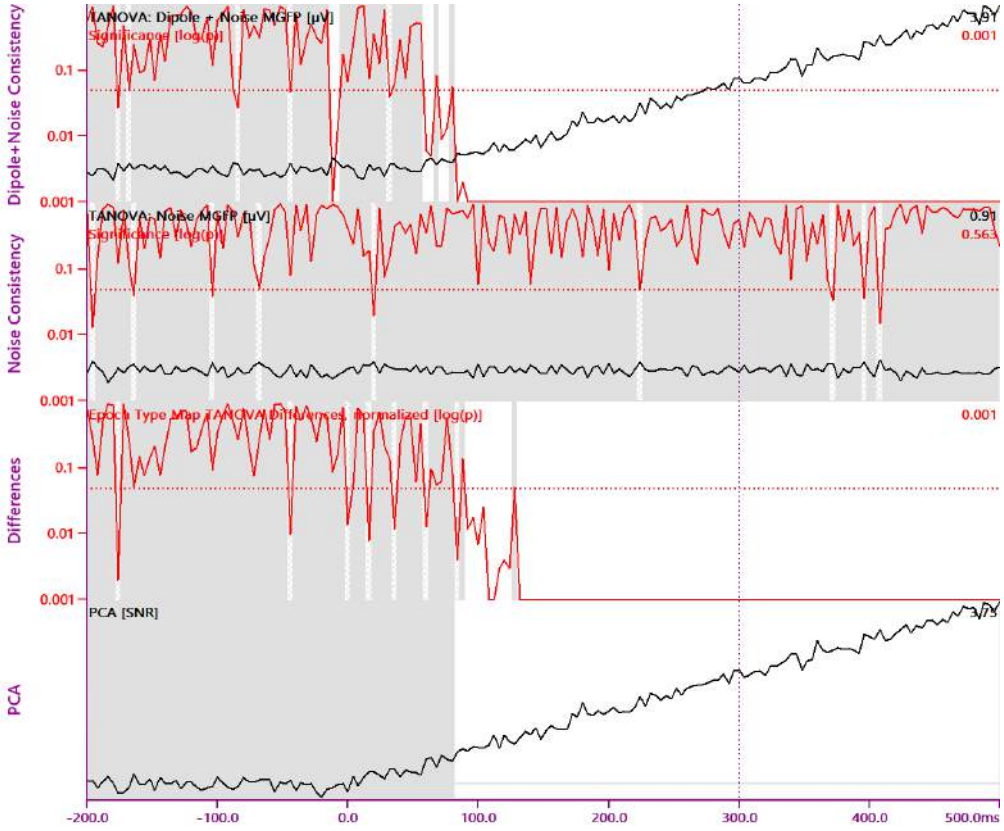


Figure 4: TANOVA results for the unfiltered simulated dataset. Red waveforms are p-Values depicted in a logarithmic scale, and white areas indicate significance, with $p < \alpha$. Short hatched gray areas indicate significance but failure to establish significance of consecutive rejections. The significance level $\alpha = 0.05$ is visualized as a horizontal red dotted line. Rows 1 and 2 show consistency test results for "dipole+noise" and "noise" epochs, respectively. Black waveforms are MGFPs of the average per event type, scaled equally. Row 3 shows differences between epoch types. Row 4 shows the SNR of the first principal component. Here, white areas mark SNRs ≥ 1. Numbers in the upper right corners of each waveform are the numerical values of that waveform for the 300 ms time point, indicated by a dotted vertical line.

and distractors yielded significance latencies from 164 ms to 184 ms, 204 ms to 380 ms, and 436 ms to 452 ms, and a total of 57 significant samples. A single significant sample at -76 ms was rejected because two or more contiguously significant samples were required to establish significance of consecutive rejections of the Null hypothesis. The results are shown in Fig. 5, and the number of significant samples is also presented in the TANOVA rows of Table 1. The computation time for the TANOVA tests with 3071 randomizations each for 31 EEG channels and 166 epochs, performed for all 176 samples per epoch at 250 Hz was 18 seconds on a 2.3 GHz Core i7-4850HQ CPU.

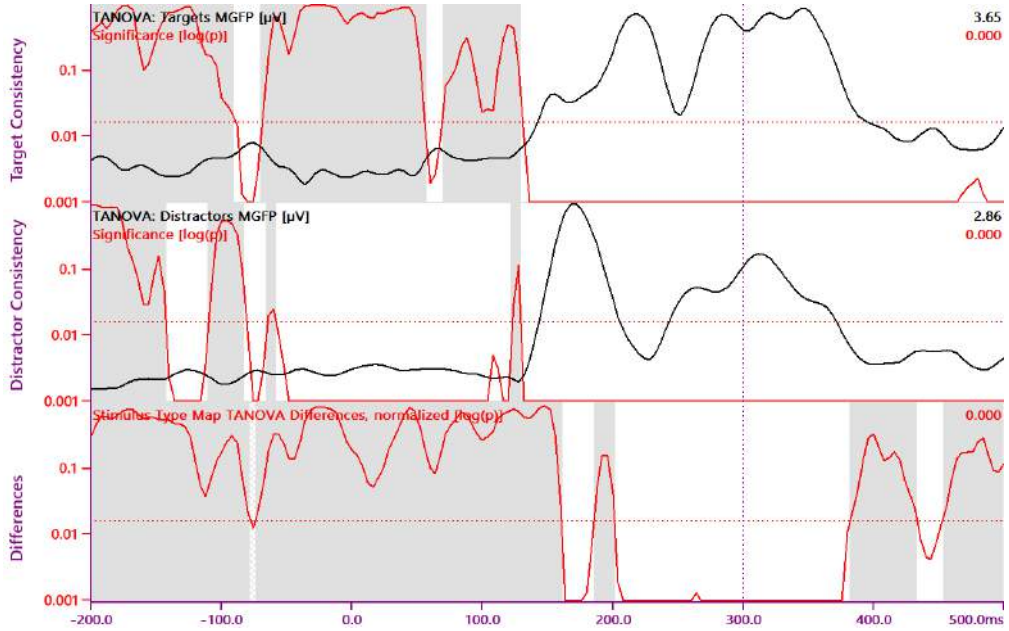


Figure 5: TANOVA results for the 40 Hz low-pass filtered CPT dataset. Red waveforms are p-Values depicted in a logarithmic scale, and white areas indicate significance, with $p < \alpha$. Short hatched gray areas indicate significance but failure to establish significance of consecutive rejections. The significance level $\alpha = 0.0163$ is visualized as a horizontal red dotted line. Rows 1 and 2 show consistency test results for target and distractor stimuli, respectively. Black waveforms are MGFPs of the average per stimulus type, scaled equally. Row 3 shows differences between targets and distractors. Numbers in the upper right corners of each waveform are the numerical values of that waveform for the 300 ms time point, indicated by a dotted vertical line.

Statistical non-Parametric Mapping

For the simulated dataset, consistency of the "dipole+noise" sLORETA results could be established for latencies from 180 ms on, with the exception of the 188 ms, 192 ms, and 204 ms samples. Single significant samples were rejected by the test for significance of contiguously significant samples, which required a minimum of two samples. The "noise" epoch consistency test failed, with a total of four single-sample false positives rejected. The test for differences of sLORETA results between epoch types yielded latencies of 196 ms and later, except for the 204 ms, 224 ms, and 228 ms samples. Results are shown in Fig. 6. The locations of significantly different activity between epoch types are illustrated in Fig. 7 for the 300 ms latency, with differences in the postcentral gyrus area and in the frontal lobes.

In the case of the low-pass filtered CPT dataset, the SnPM consistency tests established consistency (Null Hypothesis rejected for at least one voxel) for the target stimuli at a total of 139 out of 176 samples (-200 to -192 ms, -180 to -164 ms, -152 to -108 ms, -88 to -40 ms, -32 to 56 ms, 64 to 84 ms, 108 to 244 ms, 264 to 384 ms, 396 to 400 ms, 452 to 472 ms, 480 to 488 ms), and for the distractor stimuli at all samples. The test for

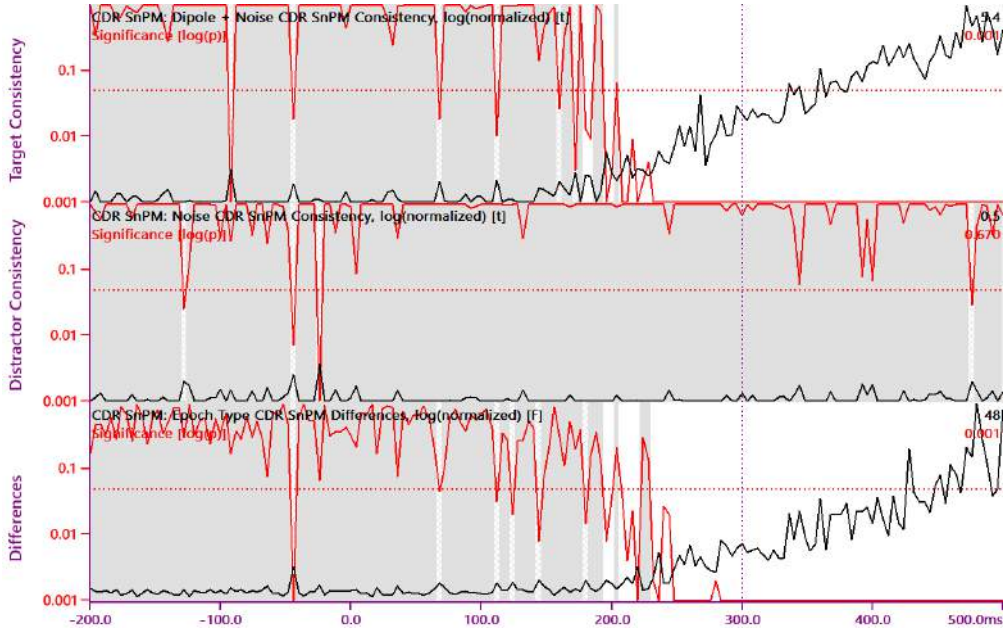


Figure 6: CDR SnPM results for the unfiltered simulated dataset. Red waveforms are p-Values depicted in a logarithmic scale, and white areas indicate significance, with $p < \alpha$. Short hatched gray areas indicate significance but failure to establish significance of consecutive rejections. The significance level $\alpha = 0.05$ is visualized as a horizontal red dotted line. Rows 1 and 2 show consistency test results for target and distractor stimuli, respectively. Black waveforms are the maximum t-values, scaled equally. Row 3 illustrates differences between targets and distractors. Here, the black waveform represents maximum F-values. The numbers in the upper right corners of each waveform are the numerical values of that waveform for the 300 ms time point, indicated by a dotted vertical line.

differences between targets and distractors yielded significant latencies (Null Hypothesis rejected for at least one voxel) from 172 ms to 180 ms, 228 ms to 244 ms, and 276 ms to 336 ms. Single significant samples at 148 ms and 344 ms were rejected because two or more contiguously significant samples were required to establish significance of consecutive rejections of the Null hypothesis. Results are presented in Fig. 8, and the number of significant samples is also presented in the SnPM rows of Table 1. Fig. 9 shows the spatial distribution of F-values above the significance level for the 300 ms time point, highlighting frontocentral and posterior right hemisphere differences. While the cerebellum was excluded from source analysis, some source symbols visually overlap with cerebellar areas due to the 7 mm resolution of the sLORETA source space. The computation of sLORETA CDR results for all epochs and samples took 40 seconds. Computation times for the CDR SnPM tests for all epochs and samples and 3071 randomizations were 33 minutes, of which 7 minutes and 10 seconds were used for the CDR SnPM difference analysis.

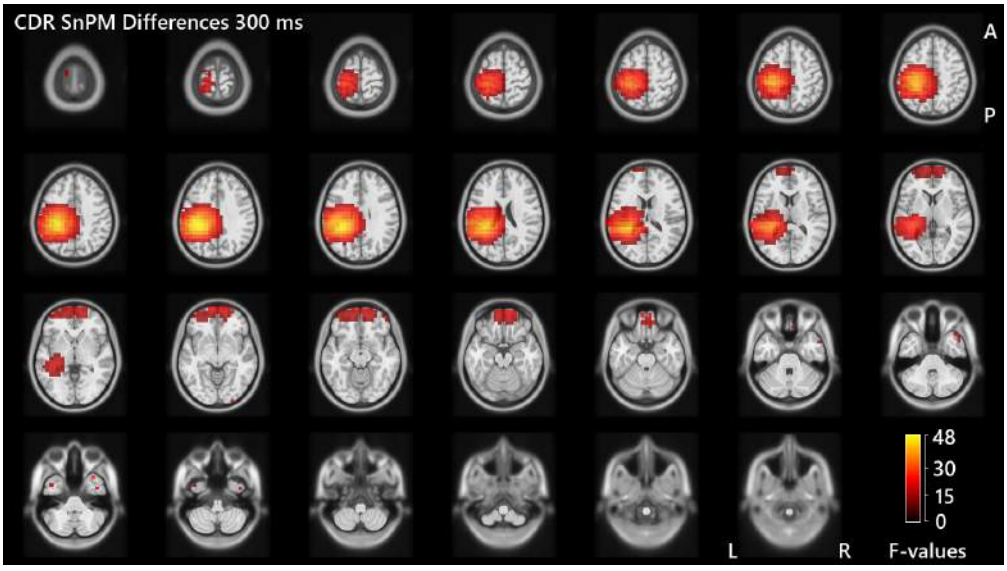


Figure 7: CDR SnPM F-values for the unfiltered simulated dataset. Thumbnails show template MRI slices ordered from top to bottom, with labels indicating axis orientations where A represents anterior, P posterior, L left, and R right. Red-yellow colored overlays are the significance-thresholded F-value image for the 300 ms latency.

Different Filter Frequencies and Epoch Counts

For the simulated dataset, Fig. 10 shows a comparison of the TANOVA and CDR SnPM difference tests already summarized above with results for a version of the dataset low-pass filtered at 10 Hz. While for the unfiltered dataset, significant TANOVA differences start at 92 ms, the filtered dataset yields significant differences already from 28 ms onwards. Both latencies are in line with the PCA results, where the SNR of the first principal component representing the simulated dipole map rises above one at 84 ms and 28 ms, respectively. The CDR SnPM differences show less sensitivity with significance established at 196 ms and 88 ms, respectively.

The results of submitting differently filtered versions of the CPT dataset to TANOVA and CDR SnPM difference tests are shown in Fig. 12. The second row of these figures represents the same 40 Hz low-pass filtered dataset that was characterized above and presented in Figs. 5 and 8. Table 1 summarizes the number of significant samples for both tests. For high-pass filter frequencies of 2 Hz and 3 Hz, the number of significant samples is reduced for the CDR SnPM tests (Fig. 11a). The impact of filtering on the TANOVA results is smaller compared to CDR SnPM. For the less-filtered data, more prestimulus samples are significant. Significant differences for the 300 ms latency cannot be detected for the 3 to 40 Hz-filtered dataset.

Reducing the number of epochs yields the results shown in Fig. 13 (here the top row represents the same dataset as in Figs. 5 and 8), as well as in Table 1. For only 42 epochs, the overall number of significant samples is markedly reduced (Fig. 11b), while for TANOVA, with smaller epoch

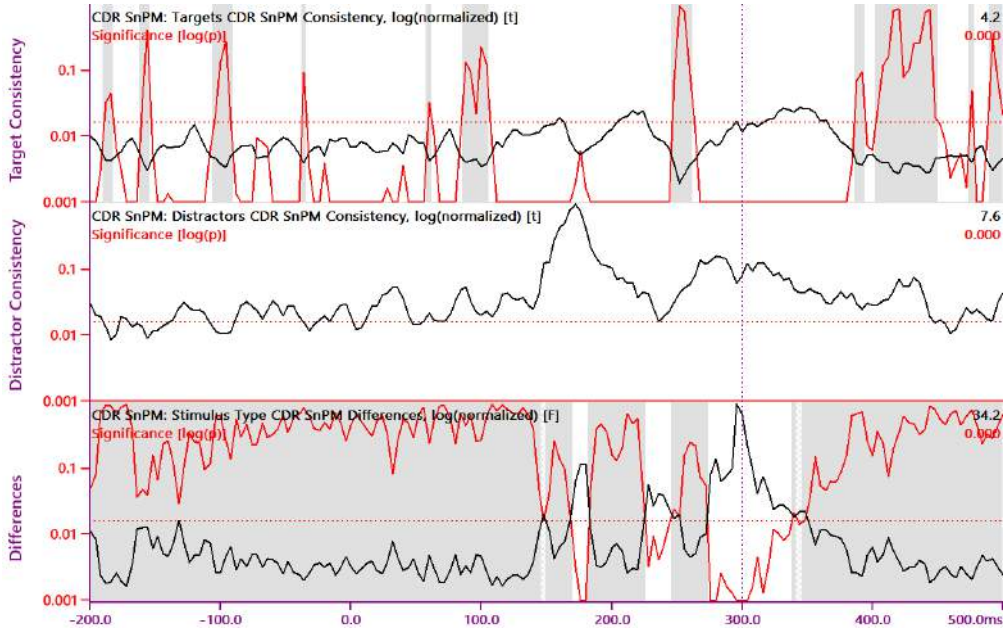


Figure 8: CDR SnPM results for the 40 Hz low-pass filtered CPT dataset. Red waveforms are p-Values depicted in a logarithmic scale, and white areas indicate significance, with $p < \alpha$. Short hatched gray areas indicate significance but failure to establish significance of consecutive rejections. The significance level $\alpha = 0.0163$ is visualized as a horizontal red dotted line. Rows 1 and 2 show consistency test results for target and distractor stimuli, respectively. Black waveforms are the maximum t-values, scaled equally. Row 3 illustrates differences between targets and distractors. Here, the black waveform represents maximum F-values. The numbers in the upper right corners of each waveform are the numerical values of that waveform for the 300 ms time point, indicated by a dotted vertical line.

counts, more pre-stimulus samples were found significant. A significant difference for the 300 ms latency cannot be established.

Discussion

For the simulated dataset, TANOVA identified latencies of consistency within epoch type for the "dipole+noise" epochs, while for the "noise"-only epochs just the expected number of false positives was reported, none of which passed the test for contiguous rejections of the Null hypothesis. Latencies with significant differences between epoch types agreed with latencies where the SNR of the first principal component of the averaged "dipole+noise" epochs exceeded 1.

CDR SnPM, when applied to the simulated dataset, highlighted two distinct brain regions of significant differences, one of them in the post-central gyrus area showing hyper-activity, the other in the frontal lobe area with hypo-activity, due to the sharper fall-off beyond the maxima of sLORETA activity for the "dipole+noise" epochs as compared to the "noise"-only epochs where the smoothing-effect of regularization dominates the sLORETA images due to lack of features in the data.

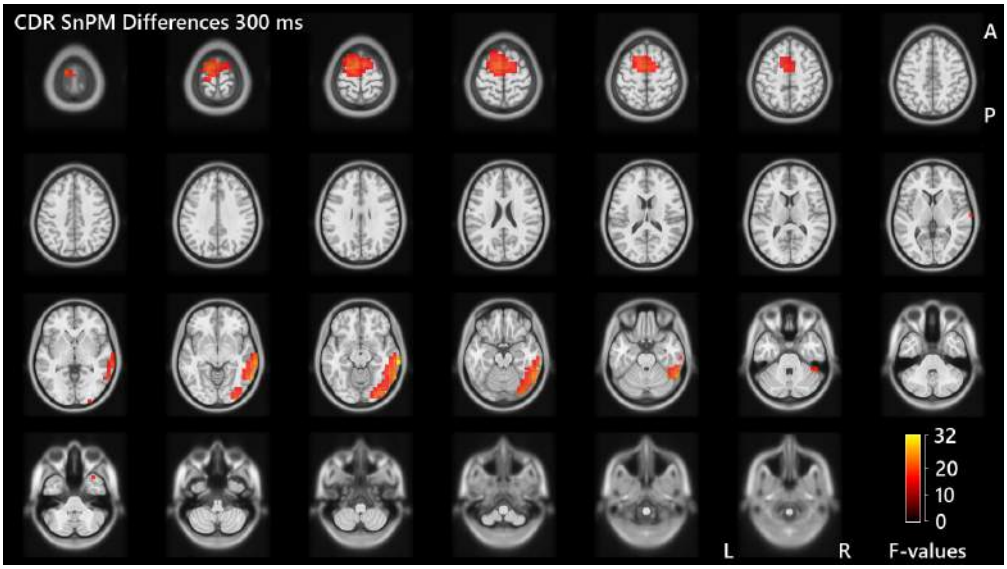
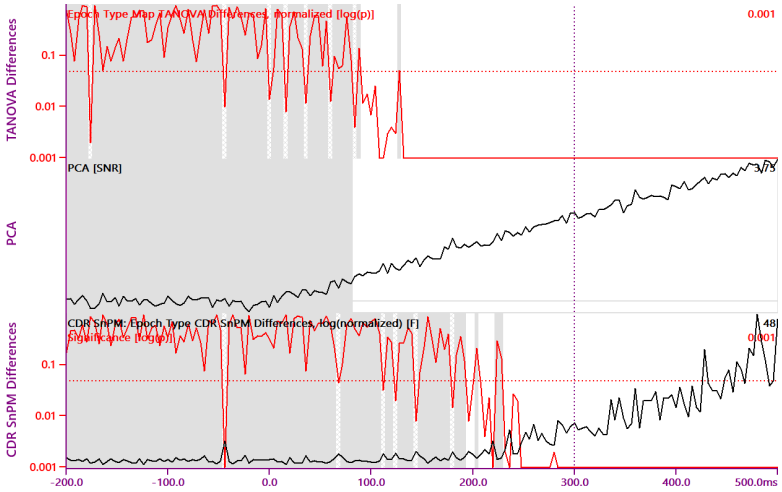
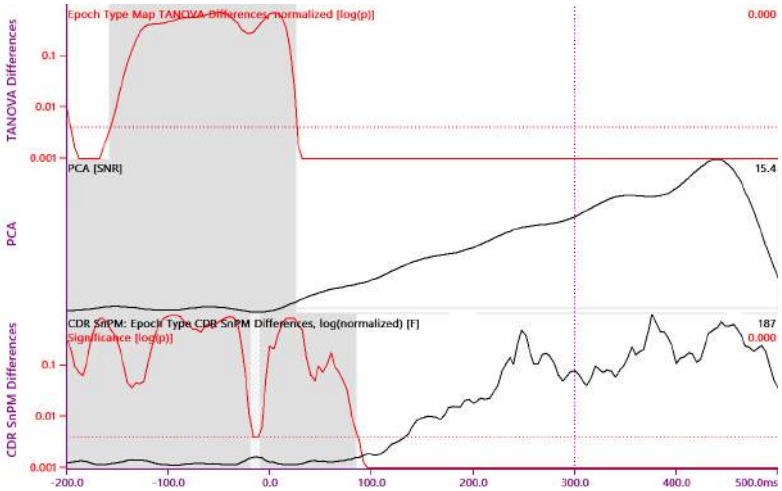


Figure 9: CDR SnPM F-values for the 40 Hz low-pass filtered CPT dataset. Thumbnails show template MRI slices ordered from top to bottom, with labels indicating axis orientations where A represents anterior, P posterior, L left, and R right. Red-yellow colored overlays are the significance-thresholded F-value image for the 300 ms latency.

The sensitivity of CDR SnPM was lower compared to TANOVA, which can be seen from the later onset of the respective periods of significant latencies in the difference tests. The maximum statistic used in CDR SnPM summarizes the individual voxel statistics into a single measure, thus addressing the spatial multiple comparison problem while at the same time retaining spatial resolution. Nichols and Holmes (2002) describe a tradeoff between statistical power and the ability to localize significant voxels. Thus, the lower sensitivity of SnPM as compared to TANOVA, which employs an omnibus measure to summarize map topography, may be explained. Furthermore, scalp topography maps contain information about source orientation, while the voxel intensities analyzed in CDR SnPM do not. It is beyond the scope of this paper to clarify whether the differences in statistical outcome between TANOVA and CDR SnPM observed here are due to the transition to source space, the representation of source activity as absolute values, or the maximum statistic employed to elicit locations of significance. It should be noted, however, that the outcome of sLORETA source localization are not oriented sources but voxel intensities. Rather, sLORETA was chosen because of its low localization error, its uniform spatial sensitivity due to the inherent normalization to the standard deviation per voxel, and because its outcome has been shown to be robust with respect to changes in the regularization parameter. The spatial resolution to be expected from CDR SnPM depends on the underlying source analysis method, with "low resolution" even a part of the sLORETA acronym. When just a representation of significant peaks as source images is required – as opposed to establishing significance for certain source locations – a

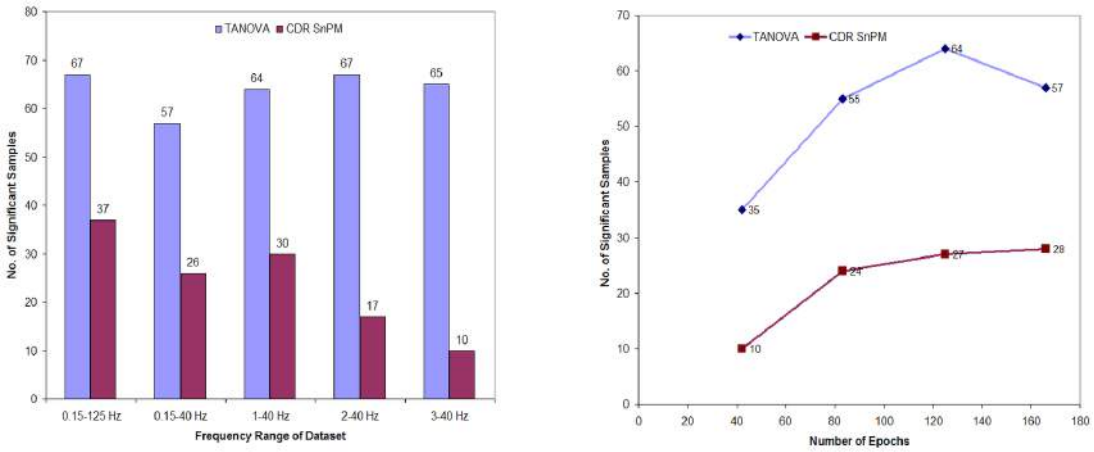


(a) Unfiltered dataset.



(b) 10 Hz low-pass filtered dataset.

Figure 10: Comparison of difference test results for the unfiltered and 10 Hz low-pass filtered simulated datasets. Red waveforms are p-Values depicted in a logarithmic scale, and white areas indicate significance, with $p < \alpha$. Short hatched gray areas indicate significance but failure to establish significance of consecutive rejections. The significance levels α are visualized as horizontal red dotted lines. Row 1 shows TANOVA results. Row 2 shows the SNR of the first principal component, where white areas mark SNRs ≥ 1. Row 3 are CDR SnPM results, where the black waveform represents maximum F-values.



(a) Differently filtered datasets.

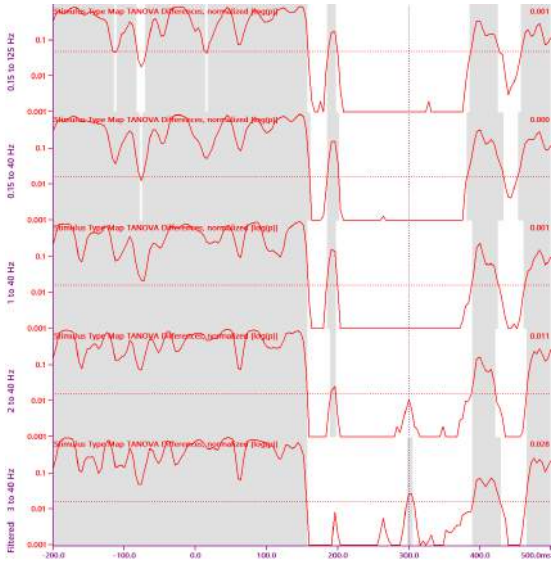
(b) Different epoch counts.

Figure 11: Number of significant samples for the TANOVA and CDR SnPM difference tests for a) differently filtered datasets and b) different epoch counts.

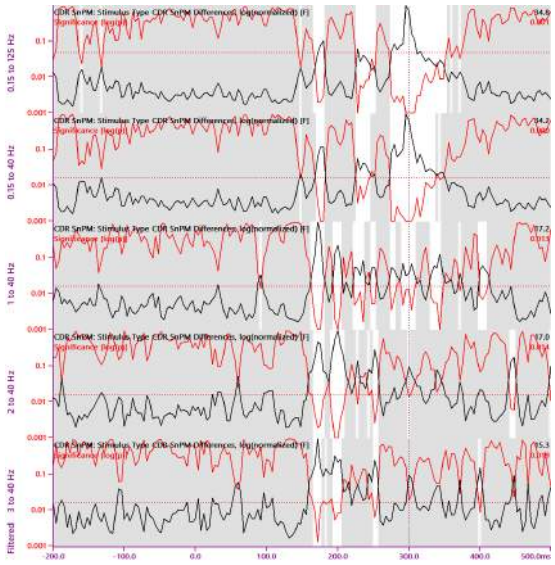
dipole or current density analysis of the samples-of-significance identified by TANOVA would be an alternative strategy to CDR SnPM.

The visual CPT data showed sufficient consistency within stimulus type to warrant a comparison of the different stimulus types. The TANOVA method, which processes complete voltage topography maps, detected significant differences between stimulus types for more latencies than the CDR SnPM method. TANOVA uses an omnibus measure of map topographies for establishing significance, with the consistency test Null hypothesis that random maps have been measured. There are certainly situations, where, even without stimulus-related brain activity, map topographies are not completely random. Examples would be subjects with strong alpha in a group of posterior electrodes or residual muscle spiking in temporal electrodes, which may both show up as significant consistencies. The TANOVA consistency test should therefore not be seen as an indication of proper artifact reduction or removal. In the case of the CPT data, when looking at the TANOVA consistency test results and also at the difference tests for small epoch counts, significant pre-stimulus latencies can be observed. Without ISI randomization, it is common for slow effects to show up as consistent signals in the pre-stimulus latencies. Even with ISI randomization, as used in this study, such effects may not be totally suppressed, especially for the Bereitschaftspotential (Gladwin, Lindsen, & Jong, 2006). Furthermore, an amplifier-side 0.15 Hz high-pass filter was used in this study, while a comparison of the differently filtered datasets showed how, with increasing high-pass filter frequencies, the number of pre-stimulus significances became smaller, indicating the additional possibility of dispersion caused by low-frequency high-pass filtering, of which the commonly used baseline correction is just a special case.

CDR SnPM was able to extract brain locations with significantly different sLORETA activations between target and distractor stimuli. The time

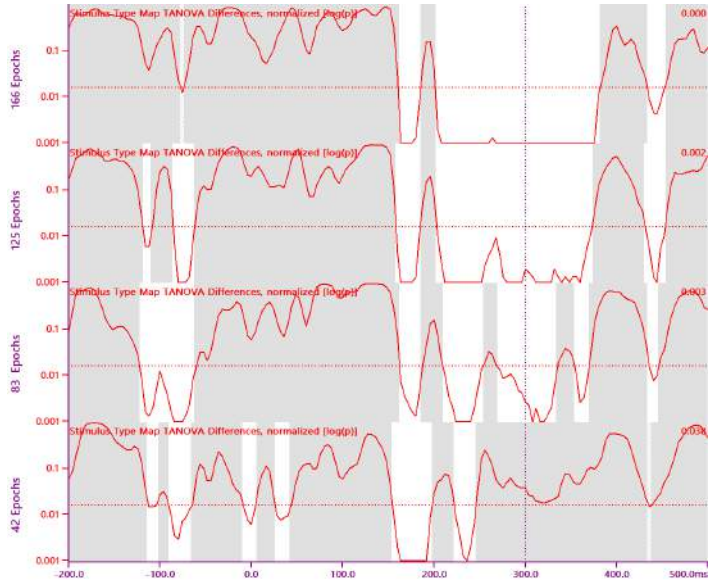


(a) TANOVA.

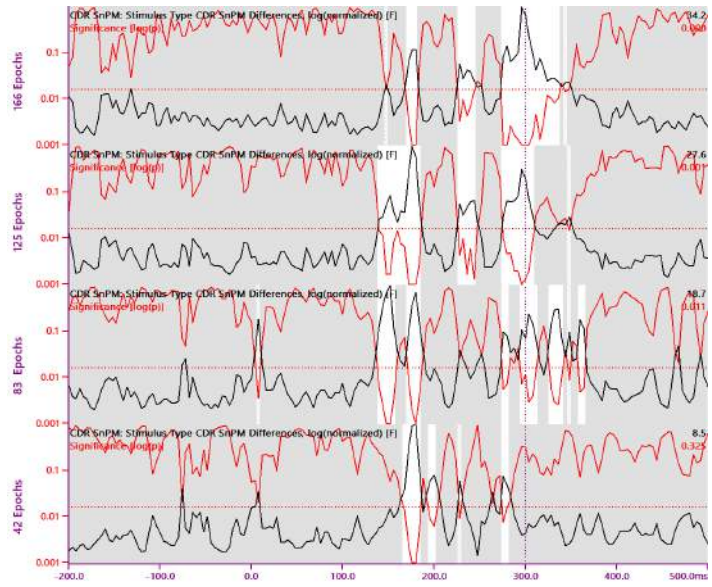


(b) CDR SnPM: black waveforms represent maximum F-values.

Figure 12: Comparison of difference test results for the differently filtered CPT datasets. Red waveforms are p-Values depicted in a logarithmic scale, and white areas indicate significance, with $p < \alpha$. Short hatched gray areas indicate significance but failure to establish significance of consecutive rejections. The significance levels α are visualized as horizontal red dotted lines. a) TANOVA b) CDR SnPM: black waveforms represent maximum F-values.



(a) TANOVA .



(b) CDR SnPM: black waveforms represent maximum F-values

Figure 13: Comparison of difference test results for different epoch counts of the 40 Hz low-pass filtered CPT dataset, with consecutive rows showing results for all, 75 %, 50 %, and 25 % of epochs used for the analysis. Red waveforms are p-Values depicted in a logarithmic scale, and white areas indicate significance, with $p < \alpha$. Short hatched gray areas indicate significance but failure to establish significance of consecutive rejections. The significance level $\alpha = 0.0163$ is visualized as horizontal red dotted lines. a) TANOVA b) CDR SnPM: black waveforms represent maximum F-values.

ranges and brain regions uncovered for the 300 ms latency are consistent with what is known about the P300, with frontocentral and posterior right hemisphere differences (Bledowski et al., 2004). The results of CDR SnPM consistency tests are of limited value beyond establishing that the underlying source images are suited for difference testing, as they show consistency for nearly all samples. This can be an effect of regularization, which for noisy data produces smooth source images of small amplitude, which are then amplified by normalization.

Filtering EEG data is a well-established technique for changing waveform morphology in such a way as to accent characteristic peaks for visual inspection, or to enhance SNR by low-pass filtering the data. With the low-pass filtered simulated dataset and temporal multiple comparison correction, the effects of this SNR enhancement are clearly visible. Significance could be established for earlier latencies with smaller dipole activity as for the unfiltered simulated data. This is corroborated by the very similar latencies where the SNR of the first principal component exceeded 1. In general, however, filtering disseminates signal energy onto nearby samples and can potentially impair features in the data that may be crucial for establishing significance. While TANOVA results were less affected by filtering, CDR SnPM seems to work best on slightly or unfiltered data. For the CPT dataset analyzed in this study, epoch counts of 125 and higher produced more significant samples than epoch counts of 83 and lower. One possible extrapolation of this observation is, that the number of epochs generally accepted as sufficient for creating average ERPs is also sufficient for performing TANOVA and SnPM analyses.

The proposed method for multiple comparison correction in the time domain uses the ratio between twice the maximum frequency in the data and the sampling rate as its comparison count parameter. A corrected significance threshold is used if this ratio is smaller than one, leading to higher numbers of required repetitions and longer computation times. The question of how to choose or estimate the maximum frequency remains to be discussed. If data have previously been filtered, as is the case in most ERP/ERF studies, an upper limit for the maximum frequency is certainly given by the low-pass filter frequency and transition width. As the observed brain processes might work on even slower scales, it becomes clear that this is by no means a conservative method for multiple comparison correction. However, it accounts at least for the common practice of recording data at 1 or 2 kHz but later low-pass filtering to some 40 or 70 Hz, which effectively reduces environmental noise. Consequently, in this paper only the low-pass filter frequency was used for determining the significance threshold.

Conclusion

It was shown how TANOVA and CDR SnPM can be applied to the individual epochs obtained in an ERP experiment. The TANOVA analysis established data plausibility and identified latencies-of-interest for further analysis. The SnPM analysis, in addition, identified brain regions of consistent activity within stimulus type and of significantly different activity between stimulus types.

Obviously, the approach presented here is not limited to EEG data analysis but can also be performed on MEG data. It can easily be extended to group or longitudinal studies. In some cases, it is then necessary to shuffle within-subject only. For group or longitudinal studies, either individual averages per stimulus type can be processed,

or all acquired epochs of all datasets. Furthermore, SnPM can be employed to identify significant channels when performed on voltage topographies instead of source distributions.

References

Bledowski, C., Prvulovic, D., Hoechstetter, K., Scherg, M., Wibral, M., Goebel, R., et al. (2004). Localizing p300 generators in visual target and distractor processing: a combined event-related potential and functional magnetic resonance imaging study. *J Neurosci*, *24*(42), 9353–9360.

Fuchs, M., Wagner, M., & Kastner, J. (2001). Boundary element method volume conductor models for eeg source reconstruction. *Clin Neurophysiol*, *112*(8), 1400–1407.

Fuchs, M., Wagner, M., Köhler, T., & Wischmann, H. A. (1999). Linear and nonlinear current density reconstructions. *J Clin Neurophysiol*, *16*(3), 267–295.

Fuchs, M., Wagner, M., Wischmann, H. A., Köhler, T., Theissen, A., Drenckhahn, R., et al. (1998). Improving source reconstructions by combining bioelectric and biomagnetic data. *Electroencephalogr Clin Neurophysiol*, *107*(2), 93–111.

Gladwin, T. E., Lindsen, J. P., & Jong, R. de. (2006). Pre-stimulus eeg effects related to response speed, task switching and upcoming response hand. *Biol Psychol*, *72*(1), 15–34.

Greenblatt, R. (1995). Biomagnetism: fundamental research and clinical applications. In C. Baumgartner, L. Deecke, G. Stroink, & S. Williamson (Eds.), (pp. 402–405). Amsterdam: Elsevier Science / IOS Press.

Greenblatt, R. E., & Pflieger, M. E. (2004). Randomization-based hypothesis testing from event-related data. *Brain Topogr*, *16*(4), 225–232.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). NY: Springer.

Kim, J. S., Han, J. M., Park, K. S., & Chung, C. K. (2008). Distribution-based minimum-norm estimation with multiple trials. *Comput Biol Med*, *38*(11-12), 1203–1210.

Kim, Y. Y., Roh, A. Y., Namgoong, Y., Jo, H. J., Lee, J.-M., & Kwon, J. S. (2009). Cortical network dynamics during source memory retrieval: current density imaging with individual mri. *Hum Brain Mapp*, *30*(1), 78–91.

Koenig, T., & Melie-García, L. (2009). Electrical neuroimaging. In C. Michel, T. Koenig, D. Brandeis, L. Gianotti, & J. Wackermann (Eds.), (pp. 169–189). Cambridge: Cambridge University Press.

Koenig, T., & Melie-García, L. (2010). A method to determine the presence of averaged event-related fields using randomization tests. *Brain Topogr*, *23*(3), 233–242.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Murray, M. M., Brunet, D., & Michel, C. M. (2008). Topographic erp analyses: a step-by-step tutorial review. *Brain Topogr*, *20*(4), 249–264.

Murray, M. M., Michel, C. M., Peralta, R. G. de, Ortigue, S., Brunet, D., Andino, S. G., et al. (2004). Rapid discrimination of

visual and multisensory memories revealed by electrical neuroimaging. *Neuroimage*, *21*(1), 125–135.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp*, *15*(1), 1–25.

Pantazis, D., Nichols, T. E., Baillet, S., & Leahy, R. M. (2003). Spatiotemporal localization of significant activation in meg using permutation tests. *Inf Process Med Imaging*, *18*, 512–523.

Pascual-Marqui, R. D. (2002). Standardized low-resolution brain electromagnetic tomography (sloreta): technical details. *Methods Find Exp Clin Pharmacol*, *24 Suppl D*, 5–12.

Wagner, M. (2014). Magnetoencephalography. In S. Supek & C. J. Aine (Eds.), (chap. Non-Parametric Statistical Analysis of Map Topographies on the Epoch Level). Heidelberg: Springer. (In Press)

Wagner, M., Fuchs, M., & Kastner, J. (2004). Evaluation of sloreta in the presence of noise and multiple sources. *Brain Topogr*, *16*(4), 277–280.

Wagner, M., Fuchs, M., Wischmann, H., Ottenberg, K., & Dössel, O. (1995). Biomagnetism: fundamental research and clinical applications. In C. Baumgartner, L. Deecke, G. Stroink, & S. Williamson (Eds.), *Biomagnetism: Fundamental research and clinical applications: Proceedings of the 9th international conference on biomagnetism* (pp. 352–356). Amsterdam: Elsevier Science / IOS Press.

Westfall, P., & Young, S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment* (Vol. 279). John Wiley & Sons.