ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 39 (2008) 1051 – 1063

Quantifying inter-subject agreement in brain-imaging analyses

Dik Kin Wong,[a,*,1] Logan Grosenick,[a,1] E. Timothy Uy,[a] Marcos Perreau Guimaraes,[a] Claudio G. Carvalhaes,[a,b] Peter Desain,[c] and Patrick Suppes[a]

[a] *Center for the Study of Language and Information, Ventura Hall, 200 Panama St., Stanford University, CA, USA*
[b] *Instituto de Matematica e Estatistica, Universidade do Estado do Rio de Janeiro, Brazil*
[c] *Nijmegen Institute for Cognition and Information, University of Nijmegen, Nijmegen, The Netherlands*

Received 5 April 2007; revised 29 June 2007; accepted 27 July 2007
Available online 23 August 2007

In brain-imaging research, we are often interested in making quantitative claims about effects across subjects. Given that most imaging data consist of tens to thousands of spatially correlated time series, inter-subject comparisons are typically accomplished with simple combinations of inter-subject data, for example methods relying on group means. Further, these data are frequently taken from reduced channel subsets defined either a priori using anatomical considerations, or functionally using *p*-value thresholding to choose cluster boundaries. While such methods are effective for data reduction, means are sensitive to outliers, and current methods for subset selection can be somewhat arbitrary. Here, we introduce a novel "partial-ranking" approach to test for inter-subject agreement at the channel level. This non-parametric method effectively tests whether channel concordance is present across subjects, how many channels are necessary for maximum concordance, and which channels are responsible for this agreement. We validate the method on two previously published and two simulated EEG data sets.

Introduction

Random variables are "concordant" if they tend to all take high or low values together (Joe, 1990), i.e. show a high degree of agreement. In the case of brain-imaging data, where variables are large numbers of spatially correlated signals from different "channels" (e.g., electrodes, sensors, voxels), we expect that some cluster(s) of channels will consistently outperform others across subjects. Such subsets are useful, as we would like to make quantitative claims about the quality and consistency of a signal in time or space across subjects. Further, they are important for methodological reasons, allowing us, for example, to define "best sensors" or functional "regions of interest" (ROIs). These reduce data dimensionality and improve statistical power. Indeed in prediction analyses, decreasing the number of nuisance predictors can markedly increase prediction accuracy.

While there are various ways to localize signals in brain imaging, there are few statistical approaches that quantify inter-subject agreement at the channel level or that validate the optimal size and shape of concordant channel sets across subjects—i.e. subsets of channels that generally yield optimal or near optimal performance across subjects for some desired criteria (such as signal amplitudes, frequency band maxima, univariate "statistical maps", or classification rates). This is not surprising, as quantitatively defining the size and location of "best" subsets can be difficult, and analyzing many correlated channels simultaneously across subjects requires non-trivial multivariate methods. As a result, current methods for finding channel "clusters" or "best sensors" typically:

I. Use reductions of the data sensitive to outliers. For example, within-group data averaging across subjects (Polich, 1997; Hagoort, 2003, Hagoort et al., 2004; Chen et al., 2003; Schendan et al., 2003), channel-wise *t*-statistics between treatment groups (Sereno et al., 2002; Van Boven et al., 2005), or the strict intersection across subjects of all channels meeting some amplitude criteria (Simon et al., 2002; Johnson-Frey et al., 2003; Grill-Spector and Kanwisher, 2005; for a review of some common data pooling methods, see Lazar et al., 2002).
II. Use independent anatomical evidence to choose a set of voxels (Stark and Squire, 2001; Johnson-Frey et al., 2003; Schendan et al., 2003; Huettel et al., 2004; Phelps et al., 2004; Van Boven et al., 2005; Etkin et al., 2006) or arranging sensors more densely (Buchner et al., 1995; Lal et al., 2004; Li et al., 2005) or exclusively (Nashmi et al., 1994; Li et al., 2004; Tang et al., 2006) over a scalp region thought to best represent the signal of interest across subjects (Nashmi et al., 1994; Li et al., 2004; Tang et al., 2006).

* Corresponding author.
E-mail address: dkwong@stanford.edu (D.K. Wong).
[1] These authors contributed equally to this work.
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2007.07.064

III. Choose "functional" ROIs using *p*-value thresholds to limit cluster size and number while controlling for family-wise error (Grill-Spector et al., 1998; O'Craven and Kanwisher, 2000; Stark and Squire, 2001; Sereno et al., 2002; Huettel et al., 2004; Phelps et al., 2004; Eickhoff et al., 2005; Van Boven et al., 2005; Etkin et al., 2006).

Such basic data reduction methods are very useful, as they simplify hypothesis testing and increase test power. However, in some cases they are based more on intuition or convention than in statistical analysis (Tang et al., 2006), while in others they can suffer from a degree of arbitrariness in the *p*-value thresholding used to limit family-wise error and define cluster size (which may be adjusted over a wide range, resulting in different distributions and densities of "activation") (Lazar et al., 2002). Many approaches employ *p*-value correction methods (such as a Bonferroni correction) meant for independent variables, which in the case of spatially and temporally correlated data are underpowered (but cf. Taylor et al., 2007, for a more refined approach). Further, such methods are often applied to data averaged across subjects, or on a "statistical parametric map" (Friston et al., 1995), or use simple intersection methods to combine information across subjects. Such methods can make strong assumptions about the data and often rely on sample means, which have a breakdown point of zero are thus arbitrarily sensitive to outliers (Brammer et al., 1997; Lazar et al., 2002; Meriaux et al., 2006). Finally, in cases where data from individual subjects are presented, it is not uncommon to show "best" channels from a representative subject, suggesting the importance of a particular region of physiological interest (D'Esposito et al., 2000; O'Craven and Kanwisher, 2000; Liu et al., 2002; Schwartz et al., 2002; Lal et al., 2004). Such exemplars – while informative and intuitively satisfying – are not equivalent to quantitative accounts of inter-subject agreement and channel performance across subjects.

For the above reasons, a robust "assumption-free" method able to statistically test for inter-subject agreement and yield a near optimal size and shape of concordant channels is desirable. Defining such an approach is not simple, however, as in brain-imaging data problems "best" individual sensors or voxels of interest are unlikely to appear in a strictly invariant order across subjects. Rather, certain channels typically appear more frequently in the top several channels across subjects, but within this top subset may be in no particular order. Methods able to quantify such general subset concordance are therefore needed.

Our test has several important advantages over existing statistical approaches. First, it is a simple extension of a well-characterized and reliable non-parametric hypothesis testing method. Ranking has played a central role in non-parametric statistics since their inception, and its application here is straightforward and natural. Second, the results are easily interpretable, as the concept of "best" channel subsets across conditions is intuitive—and already frequently used without proper statistical justification. Third, it has straightforward applications in the context of brain-imaging data. Indeed, concerns with channel concordance follow naturally from problems involving feature selection and channel elimination in brain–computer interfaces (BCI) (Guger et al., 2000; Blankertz et al., 2002; Schroder et al., 2003; Lal et al., 2004) and are easily extended to conceptually related problems such as choosing "best" channels across subjects. Fourth, our method introduces "partial ranking", a simple but novel approach that effectively detects concordance in subsets of the data that would remain hidden in a full data set. Fifth, this partial-ranking method may be described from the more classical viewpoints of parameterized hypothesis testing (Shaffer, 1995) and empirical risk minimization (Lehmann and Casella, 1998), in which we are minimizing the expectation of a loss function on each of multiple tests with highly correlated outcomes (see Supplementary information). Finally, we believe this new method is easily extensible to a variety of problems, as the user is left the freedom to choose the performance criteria for the channels, and since looking across subjects is easily generalized to looking across other conditions (such as different classifiers, spatial co-registration methods, and so on).

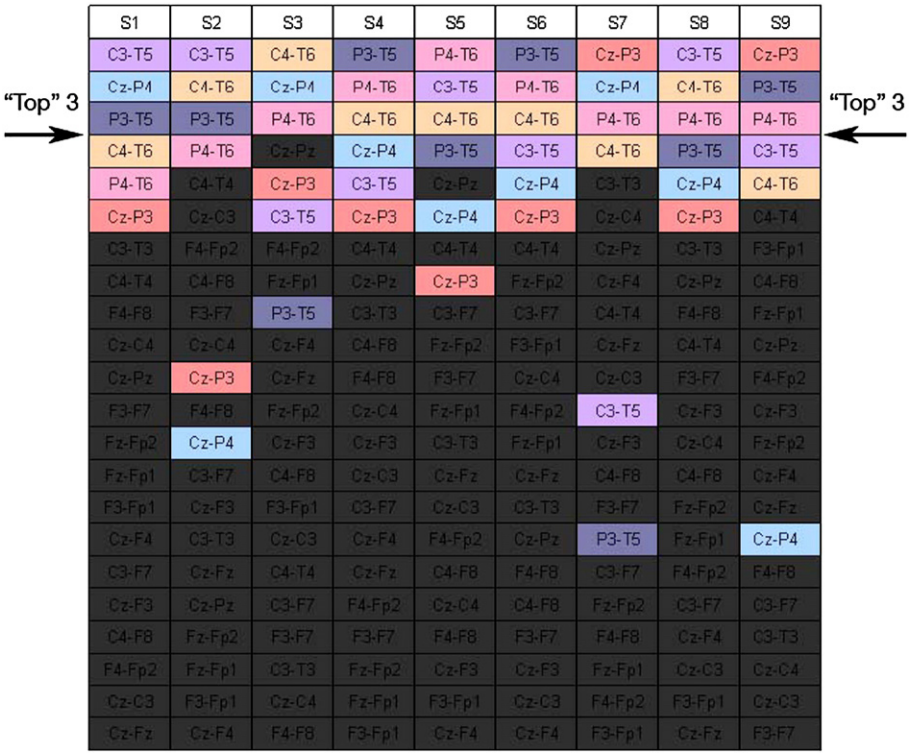
As proof of concept we validate our method on two previously published papers involving pattern classification of EEG brain waves (Suppes et al., 1999a; Lal et al., 2004—using in the first case the channel-wise analysis of the data reported in Wong et al., 2004), and on two simulated data sets. These are easy data on which to validate our method as the number of channels is not too large—and analogous problems (e.g. ROIs with tens to hundreds of channels) should be similarly easy to solve. However, we also discuss optimizations of the computational methods that would allow one to apply the method across 20 subjects with 10,000 channels in less than 24 h on a standard computer. For the two particular experiments we are evaluating, both of which involve single-trial classification of EEG data, the criteria of interest are classification rates. That is, we are interested in how well a model, when trained only on un-averaged EEG trials from a subject recorded during the presentation of a particular stimulus, can correctly identify "new" trials of that stimulus from the brain waves (trials the model was not trained on) (for a general reference, see Hastie et al., 2001; for fMRI classification, see Cox and Savoy, 2003; for EEG classification, see Wong et al., 2004; for MEG classification, see Perreau-Guimaraes et al., 2007). However, it is easy to see that our method can apply to other criteria of interest, such as signal amplitude at a particular latency, frequency band maxima, or any signal features of interest—as long as these features can be ranked and are available at the channel level. The success of this method is also demonstrated by two sets of simulated data, one statistical and the other physical.

We emphasize that the concentration of this paper is on a new statistical approach, and not on the already published experimental results. However, we will discuss the results of applying our new method to these data, which are to the best of our knowledge the first clear use of multivariate hypothesis testing to show channel concordance across brain-imaging subjects.

Methods

Qualitative approach

When applying pattern classification methods to brain-imaging data, it is natural to expect that some channels (e.g., control channels) should not help the classification. Further, we expect that channels closely associated spatially with the signal of interest will yield better results than those that are not. We may qualitatively evaluate these expectations by ranking our channels for each condition according to our criteria (in this case, how successful we are at classifying different trials using data from a particular sensor), and then comparing the rankings to see if certain channels are consistently higher or lower than others in rank across subjects. In Fig. 1, we give an example of such a rank comparison, with the channels of each subject from the 48-sentence experiment listed in



S1	S2	S3	S4	S5	S6	S7	S8	S9
C3-T5	C3-T5	C4-T6	P3-T5	P4-T6	P3-T5	Cz-P3	C3-T5	Cz-P3
Cz-P4	C4-T6	Cz-P4	P4-T6	C3-T5	P4-T6	Cz-P4	C4-T6	P3-T5
P3-T5	P3-T5	P4-T6	C4-T6	C4-T6	C4-T6	P4-T6	P4-T6	P4-T6
C4-T6	P4-T6	Cz-Pz	Cz-P4	P3-T5	C3-T5	C4-T6	P3-T5	C3-T5
P4-T6	C4-T4	Cz-P3	C3-T5	Cz-Pz	Cz-P4	C3-T3	Cz-P4	C4-T6
Cz-P3	Cz-C3	C3-T5	Cz-P3	Cz-P4	Cz-P3	Cz-C4	Cz-P3	C4-T4
C3-T3	F4-Fp2	F4-Fp2	C4-T4	C4-T4	C4-T4	Cz-Pz	C3-T3	F3-Fp1
C4-T4	C4-F8	Fz-Fp1	Cz-Pz	Cz-P3	Fz-Fp2	Cz-F4	Cz-Pz	C4-F8
F4-F8	F3-F7	P3-T5	C3-T3	C3-F7	C3-F7	C4-T4	F4-F8	Fz-Fp1
Cz-C4	Cz-C4	Cz-F4	C4-F8	Fz-Fp2	F3-Fp1	Cz-Fz	C4-T4	Cz-Pz
Cz-Pz	Cz-P3	Cz-Fz	F4-F8	F3-F7	Cz-C4	Cz-C3	F3-F7	F4-Fp2
F3-F7	F4-F8	Fz-Fp2	Cz-C4	Fz-Fp1	F4-Fp2	C3-T5	Cz-F3	Cz-F3
Fz-Fp2	Cz-P4	Cz-F3	Cz-F3	C3-T3	Fz-Fp1	Cz-F3	Cz-C4	Fz-Fp2
Fz-Fp1	C3-F7	C4-F8	Cz-C3	Cz-Fz	Cz-Fz	C4-F8	C4-F8	Cz-F4
F3-Fp1	Cz-F3	F3-Fp1	C3-F7	Cz-C3	C3-T3	F3-F7	Fz-Fp2	Cz-Fz
Cz-F4	C3-T3	Cz-C3	Cz-F4	F4-Fp2	Cz-Pz	P3-T5	Fz-Fp1	Cz-P4
C3-F7	Cz-Fz	C4-T4	Cz-Fz	C4-F8	F4-F8	C3-F7	F4-Fp2	F4-F8
Cz-F3	Cz-Pz	C3-F7	F4-Fp2	Cz-C4	C4-F8	Fz-Fp2	C3-F7	C3-F7
C4-F8	Fz-Fp2	F3-F7	F3-F7	F4-F8	F3-F7	F4-F8	Cz-F4	C3-T3
F4-Fp2	Fz-Fp1	C3-T3	Fz-Fp2	Cz-F3	Cz-F3	Fz-Fp1	Cz-C3	Cz-C4
Cz-C3	F3-Fp1	Cz-C4	Fz-Fp1	F3-Fp1	Cz-C3	F4-Fp2	F3-Fp1	Cz-C3
Cz-Fz	Cz-F4	F4-F8	F3-Fp1	Cz-F4	Cz-F4	F3-Fp1	Cz-Fz	F3-F7

Fig. 1. Qualitative "coloring" example. The columns of the table above are the channel rankings of the 9 subjects of the 48-sentence experiment. We have colored all channels that appear at least once above a cut-off of three channels – i.e. in the "top" three – with each channel assigned a different color. Qualitatively, some of the channels are clustered in the top several slots, demonstrating channel concordance. The distribution of these channels within these top slots lacks a clear ordering, and it is unclear if choosing three channels as a cut-off is best.

descending order of our criterion: classification rate. A similar qualitative table is shown for the ranked BCI data in the published paper (Lal et al., 2004) and such tables have been used elsewhere to visually evaluate channel agreement. To show how we might use such a table to identify "best" channels, we have colored all channels that are ranked in the top I channels at least once (arbitrarily setting $I=3$ for this example). Interestingly, only 6 channels meet this criterion, suggesting that the channel ranks are far from randomly distributed. Further, of the remaining table rows (ranks less than 3), only 27 of the 171 possible channel positions are colored, and these 27 colored channels are distributed non-uniformly, with 20 of them ranking 4th, 5th or 6th.

From these results, it is qualitatively clear that across subjects a certain subset of channels seem to outperform the rest. What is less clear is precisely which channels these are, and how we might quantitatively define a "best" subset (e.g., how we might choose I). In Fig. 1, several channels from the same scalp region appear reliably in the top several positions, while channels in the rest of the table seem to be more randomly distributed. Such subset concordance is based on the dependence of rank permutations across multiple subjects, and as such is difficult to describe quantitatively. Several classical ranking statistics provide an easy way to compare the full rankings to one another. However, on a full ranking the concordance of a few good channels could be lost in the noise of the remaining channels. Further, even if such a classical rank test did detect concordance, it does not provide a ready method for choosing a subset. Following this qualitative example, we examine quantitative solutions to these questions in the context of EEG channel concordance across subjects.

Quantitative approach

Statistical methods for ranking have a long history, and numerous statistical measures for comparing two ranks have been developed (Kendall, 1938; Kendall and Smith, 1940; Moran, 1947; Kruskal, 1958; Stephens et al., 1977). In this paper, we will concentrate on Spearman's rho (ρ), which is widely used, has been extensively studied (Hotelling and Pabst, 1936; Kendall et al., 1939; Student, 1921), and is easily extensible to other correlation methods (Fieller et al., 1957).

Let $\mathbf{X}=(\mathbf{X}_1,\ldots,\mathbf{X}_N)$ and $\mathbf{Y}=(\mathbf{Y}_1,\ldots,\mathbf{Y}_N)$ be random vectors of length N. Given a ranking function $r(\bullet)$ which takes a vector and returns a ranking of its elements, we can find the ranks $r(\mathrm{x})=(r(\mathbf{X}_1),\ldots,(\mathbf{X}_{N_C}))$ and $r(\mathbf{Y})=(r(\mathbf{Y}_1),\ldots,(\mathbf{Y}_{N_C}))$, where N_C is the number of channels, bold-faced letters represent vectors, and capital italic letters with subscripts are elements of these vectors. Spearman's ρ_{N_C} between $r(\mathbf{X})$ and $r(\mathbf{Y})$ is then:

$$\rho_{N_C} = 1 - \frac{6\sum d_i^2}{N_C(N_C^2-1)} \tag{1}$$

where $\Sigma d_i^2=\Sigma_{i=1}^{N_C}[r(\mathbf{X}_i)-r(\mathbf{Y}_i)]^2$. Using this measure, we can compute ρ_{N_C}'s for all possible pairings of subject channel ranks, with each ρ_{N_C} defined by Eq. (1) applied to the channel rankings of any two subjects in the experiment. It is important to note that because this test takes into account the entire ranking, the concordance of a small subset of signal channels could go undetected in the presence of many noisy channels. Furthermore, even if concordance is detected, the test does not provide a clear method for choosing a "best" subset.

Partial ranking

If we simply want to know whether the overall channel rankings are similar, Eq. (1) will suffice. However, if we are interested in finding the most concordant subset of channels across subjects, we must extend our test to include rankings of subsets of the data. Such a formulation seems to be missing in the classical ranking literature. This absence is not surprising, as intentionally ignoring a large percentage of available data has not traditionally been a strategy employed by statisticians. However, such "feature selection" is common in the data mining and machine learning applications that have arisen in the last decade or so. These new methods must cope with high-dimensional data sets with low signal-to-noise ratios and therefore employ various means of dimension reduction, including systematic data exclusion. In the spirit of such feature selection, we have extended the classical Spearman's ρ to test for concordance on limited subsets of the channels.

In using partial rankings, there are two clear approaches to selecting channel subsets. Having ranked the data for each subject (as in Fig. 1), we could either choose an overall criterion (e.g., averaged performance for all subjects) and use it to select subsets from each ranking – which may give us non-consecutive subsets within individual subject rankings – or we could examine the similarity of consecutive top subsets across individual subjects (as in the "top I" approach explored in the qualitative example). Having explored both approaches, we have found the latter to be more effective, and so leave our discussion of the former to the Supplementary information for completeness.

Bivariate intersection method

Our "intersection method" essentially seeks to maximize the size of the channel intersection across subjects given a specific "cut-off" I across subjects, i.e. the number of channels that appear for all subjects if we consider only the top I channels. This number is then compared to a simulated null distribution of the size of intersections between randomly permuted channel rankings also truncated below I. It therefore resembles the qualitative "coloring" example given above, where we noted that the "top" 3 channels across subjects (all channels present for $I=3$) were far from a random sampling of the possible channels. Here, however, we evaluate across all possible values of I to find the results furthest from a simulated null distribution, that is, furthest from a random permutation of the rankings.

To do this we repeatedly divide the subjects into two groups: Group 1 and Group 2. Using subjects in Group 1, we derive a mean ranking based on the individual channel classification rates (or ranks) of subjects in the group, and then renormalize this vector to a ranking. This "prototype" ranking is then compared with the ranking for each of the subjects in Group 2 as follows. First the I best channels are chosen from the prototype ranking, and the J best channels are chosen for each subject in Group 2 (note J need not be equal to I). We then count the number of channels present in both the prototype and the selected Group 2 subject rankings (the intersection of the truncated rankings) and repeat this for every subject in Group 2 and for all possible sizes of Group 1. For N_s subjects there are thus 2^{Ns}-2 possible two-group partitions (e.g., the 48-sentence experiment has 9 subjects and thus 510 possible partitions). For a specific I and J, and under the null hypothesis of the two ranks being independent, the expected number of overlapping channels is then:

$$\mu = \frac{I \cdot J}{N_C \min(I,\ J)} \tag{2}$$

where N_C is the number of channels (for derivation, see Appendix A).

This equation can be used in cases where I and J are not equal. For simplicity, we consider only the case $I=J$. For some intersection of size k, with a block size $I=J$, it can be shown that the bivariate distribution function under the null hypothesis is

$$f_I(k) = \frac{[I!(N_C - I)!]^2}{N_C!k![(I-k)!]^2(N_C - 2I + k)!} \tag{3}$$

(for derivation, see Appendix A).

The p-value for an intersection of fixed size c can then be obtained in the standard way by summing the probabilities of intersections of size c or greater under the null distribution.

Multivariate intersection method

While we were able to derive a closed-form solution for the bivariate density function, extending this distribution to the multivariate joint distribution is a difficult problem (Oja, 1999). Simulating this distribution, however, is straightforward, and we provide a simple algorithm for doing so in the Methods section. From this distribution, we can evaluate significance using the distance of the observed data from the simulated null hypothesis, with the number of standard deviations away from the simulated distribution of $f_I(k)$ providing the distance measure. An optimal I may be chosen in this way, and the corresponding "best" subset found by taking all channels appearing at and above the truncation rank I. Details of the actual algorithm can be found in Appendix A.

Experimental results

In this section, we apply classical Spearman ranking and our novel "partial ranking" tests to real experimental data from two EEG data sets. The essentials of the experiments used in the validation are as follows. In the first experiment (Suppes et al., 1999a), nine subjects were shown 48 sentences about European geography over 480 trials, and EEG data were recorded from 22 bipolar sensors with standard placement (Jasper, 1958). A subsequent paper classified the un-averaged trials from the 48-sentence experiment to establish single-trial classification rates for each channel (Wong et al., 2004). We use these channel-wise classification rates in our validation of the method presented here. In the second experiment (Lal et al., 2004), the EEG was recorded from 39 monopolar sensors with standard placement (Oostenveld and Praamstra, 2001) while subjects imagined either left- or right-hand movement for 400 trials. Time series data from five subjects were fit with an autoregressive (AR) model, and a novel non-linear feature selection method was used to rank the channels. Both experiments therefore yielded classification rates for each channel (i.e. how well the data from just that channel predicted differences in the task), which are used as the performance criteria in the applications below. A complete exposition of the experiments is available in the published papers.

Applying classical Spearman ranking to the EEG data

Applying Eq. (1) for all the pair-wise combinations of subject rankings (i.e. between each of the $N_s \times (N_s - 1)/2$ unordered subject pairs) for the 48-sentence experiment yielded the ρ's and p-values shown in Supplementary Table 1 (p-values were calculated with the AS 89 algorithm for the bivariate Spearman's ρ) (Best and Roberts, 1975), implemented in *R* (R Foundation for Statistical Computing, 2006). While the expected value of each ρ under the null hypothesis is 0, those shown in the table have a mean of 0.52. In this experiment, then, the classical test does appear to detect concordance at a significance level of $\alpha = 0.05$, as the values in the table show that the p-value for this mean ρ falls $0.02 > p > 0.01$ (as the p-value is a monotonically decreasing function in ρ, and $0.54 > \rho > 0.50$ in the tabled values). While this is perhaps an interesting finding by itself, it is all we can say. We do not have a clear way of determining which channels in particular are responsible for the concordance.

Similarly, in Supplementary Table 2, we show the ρ's and p-values for all pair-wise combinations of subjects in the BCI experiment. While the expected value of each ρ under the null hypothesis is 0, those shown in the table have a mean of 0.3745. Again, we have detected some concordance significant at the $\alpha = 0.05$ level, with $0.046 > p > 0.01$ according to the table. While this is promising, we lack a natural method for selecting a best subset for evaluating individual channels.

Intersection method applied to the EEG experiments

In Fig. 2a, we show the case where $I = J$ for the 48-sentence experiment. The bold line shows the actual ranking of the data (n_{exp}), while the other lines were computed by repeatedly (1000 times) creating nine random permutations of independent rankings (n_{sim}), one for each subject. The expected value of n_{sim} is computed using Eq. (2). For each I, the number of standard deviations is shown in Fig. 2b.

With such large standard deviations, we can easily evaluate the degree of concordance for different I. For example, looking at just the top channel ($I = 1$), we find that the intersection rate is not significantly different from chance. The probability of an element being in the intersection is less than 0.10 while chance level is 5% (1 out of 22). This suggests that finding a single uniformly "best" channel across subjects is highly unlikely. On the other hand, using $I = 6$, which has a chance level of 0.27 (6 out of 22), the rate is more than 80%. This is more than 15 standard deviations away from the expectation under the null hypothesis. Further, these plots peak at specific subset sizes, suggesting an optimal subset size for maximizing channel concordance, and making them well-suited for subset selection.

In Fig. 2c, we again show the case where $I = J$. The bold line is n_{exp} and the other lines are n_{sim} where n_{exp} comes from the data and the n_{sim} are computed by repeatedly (1000 times) creating five randomized independent rankings of 39 items. The expected value of n_{sim} is again computed using Eq. (2). For each I, the number of standard deviations is shown in Fig. 2d. There is a clear peak at $I = 3$ channels corresponding to an intersection rate of more than 50%—20 standard deviations from the expectation of the simulated null hypothesis. The BCI experiment is of additional interest here as it includes a documented outlying subject, who unlike the other subjects showed task-relevant muscle activity (Lal et al., 2004). Removing this outlier results in a larger concordant set emerging (corresponding to $I = 8$; see Supplementary information).

Finally, our approach yields similar results for all possible sizes of Group 1 for both data sets (Fig. 3). This suggests that in some circumstances (e.g., if we have a very large number of channels)

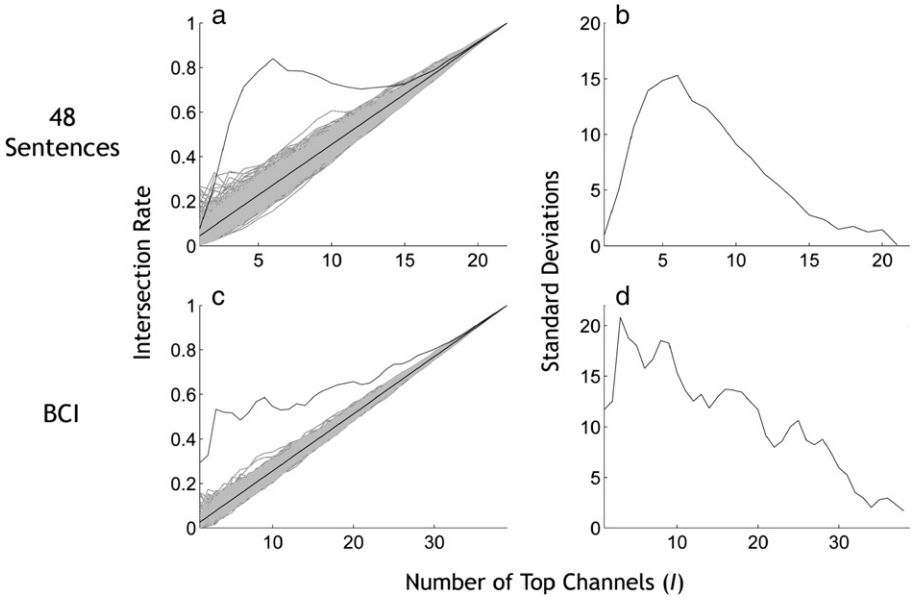


Fig. 2. Intersection method results for the 48-sentence and BCI experiments. Results of applying the intersection method are shown here for both the 48-sentence and BCI experiments. Plots in the upper row refer to the 48-sentence experiment while those in the lower row correspond to the BCI experiment. In the two figures on the left, the black line is n_{exp}, corresponding to the observed data, and the other 1000 light grey lines are the simulated null distribution n_{sim}. In both we show the intersection rate as a function of I, the number of "top" channels being compared. In the figures on the right, we show the number of standard deviations the intersection rate of the observed data was from the expectation under the null hypothesis, also as a function of channel number. The plots yield clear maxima.

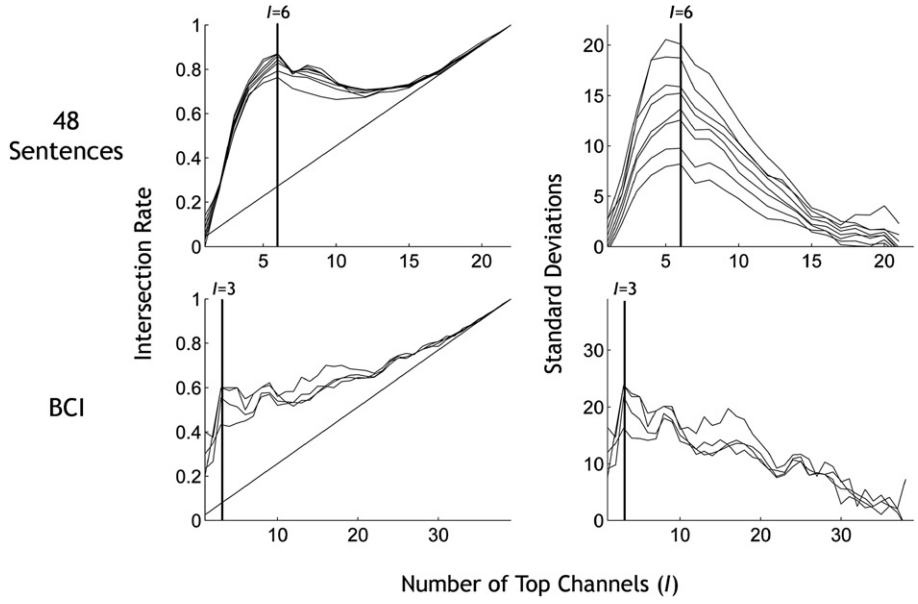


Fig. 3. Intersection rates and standard deviations for different sizes of Group 1. The plots on the left show the intersection rates of the Group 1 and Group 2 rankings for different sizes m of Group 1 for the two experiments, plotted as a function of the number of channels I. The plots on the right correspond to the numbers of standard deviations the observed data is from the expectation of the simulated null distribution at different sizes m, again plotted as a function of I. Note that for all m, the plots peak around $I=6$ for the 48-sentence experiment and around $I=3$ for the BCI experiment. This suggests we might be able to use just one value of m in our analysis, greatly reducing computational complexity.

we may be able to use just the largest group size for Group 1 (i.e. the size of Group 1 is N_s-1), reducing the number of computations from $O(2^{Ns})$s to $O(N_s)$, a convenient reduction in computational complexity. We validate this computational simplification on both our simulated and real data sets.

Subset comparisons

Given the results above, the "intersection method" for partial ranking works well for precise channel subset selection, providing clear maxima in terms of the standard deviations of different subset sizes from the null simulation. In terms of hypothesis-driven evaluation of concordance over subsets selected *a priori*, where we are simply evaluating pre-specified subsets of channels for the existence of significant concordance (as we do below), such precision is perhaps less important, and the "criterion method" presented in the Supplementary information might also be used. Still, we will use the "intersection method" presented above for the rest of the evaluations in this paper, as its null distribution is easy to simulate and so allows a clearer statistical comparison than the averaged or median ρ scores generated by the "criterion method".

48-sentence experiment: simple comparison of anterior vs. posterior channels

Previously published work on sentence classification (Suppes et al., 1998, 1999a,b; Wong et al., 2004) suggests that channels behind and including the coronal midline dominate in terms of single-trial classification rates in sentence/word classification, and other semantic tasks have been found to show peak EEG amplitude over these channels (e.g., Hagoort et al., 2004). We might therefore be interested in examining this claim across subjects in a hypothetical region of interest chosen *a priori* based on these functional results. To this end, we could compare channels in the ROI subset to their complement (the remaining frontal and central channels) to see if these channels are randomly distributed across subjects while the channels of interest show strong concordance. As a simple evaluation, we partitioned the channels into "anterior" and "posterior" subsets of equal size and ran separate partial ranking tests on each subset (11 bipolar channels per subset—with "posterior" channels defined as any bipolar pair including electrodes from behind or strictly on the coronal midline T3–C3–Cz–C4–T4, and "anterior" pairs being those that included at least one member anterior to the midline). The top panels of Fig. 4 display the results of these tests, and Fig. 5a shows the *a priori* subset. It is clear that the concordance property holds for the posterior channels, which show intersection rates more than 7 standard deviations from the simulated mean. The anterior channels, however, show no difference from the null hypothesis of random rankings, remaining less than 1 standard deviation from the mean. Such results are encouraging, as they confirm that we can pick improved subsets of channels that continue to demonstrate rank concordance while excluding channels that do not contribute to our classification across subjects. They also refute claims that concordance might be due simply to, for example, more noise on front channels making back channels uniformly better. Such validation may seem quite basic, but is lacking from current assessments of channel quality, and is easily extended to more detailed applications.

BCI experiment: simple comparison of motor vs. non-motor channels

For the BCI experiment, it is natural to ask if channels positioned over motor areas of the cortex will show greater concordance than those positioned elsewhere. We would expect this to be the case, given that these channels should reliably yield

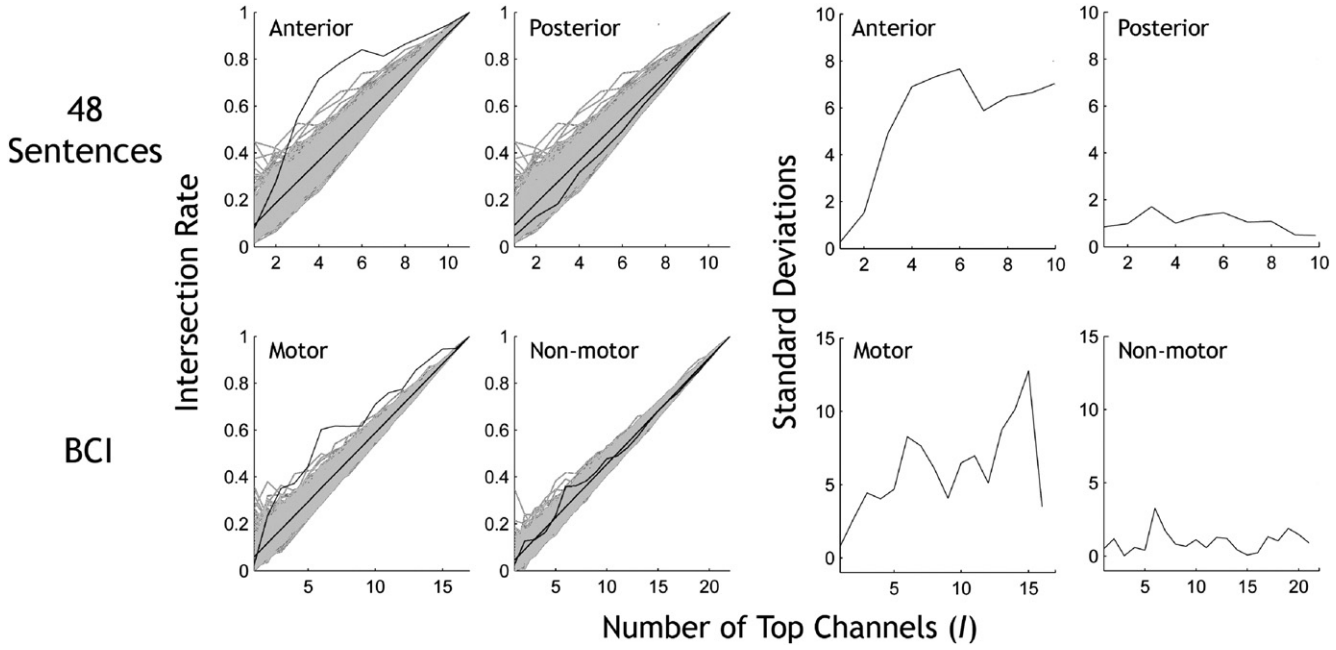


Fig. 4. Plots of intersection rates and their standard deviations for the subset comparisons as functions of the number of best channels. The top row shows the intersection rates for observed data (black line) and simulated data (light grey lines) for the "anterior" and "posterior" subsets of the 48-sentence experiment as well as the number of standard deviations away from the simulated null distribution the observed data fell. The bottom row shows similar results for the "Motor" and "Non-motor" subsets of the BCI data.

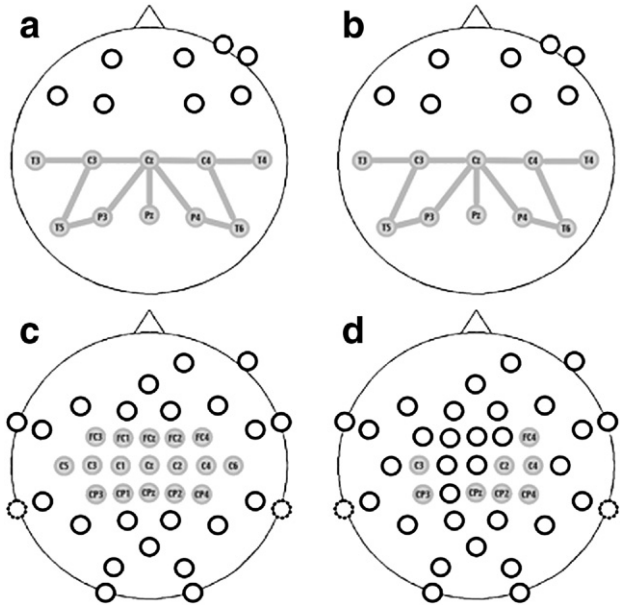


Fig. 5. *A priori* and data-driven subset selection. In panel a we show the channel subset chosen *a priori* for the 48-sentence data according to previous functional results with related stimuli. In panel b we show the subset given by applying the intersection method to all channels of the same data set. The subsets are identical, confirming the functionally defined ROI to be near optimal in terms of concordance of classification rates across subjects. In panel c we show the best channel subset chosen *a priori* by Lal et al. (2004) in the BCI experiment based on anatomical considerations. Again, the data driven approach shown in panel d yields a reduced subset that is perfect subset of the anatomical ROI.

higher classification rates than, for example, artifact detection channels or channels over areas unrelated to the task. We divided the monopolar channels into two partitions: "motor" and "non-motor", using the 17 channels over or near motor areas as defined in Lal et al. (2004) as the region-of-interest (motor subset, Fig. 5b). The remaining 22 channels constituted the non-motor subset. The results of running separate intersection tests for concordance on these subsets are shown in the bottom panels of Fig. 4. Again, it is clear that there is significant concordance in the subset of interest, with the motor subset showing intersection rates approaching 6 standard deviations from the simulated mean, while the non-motor subset shows intersection rates of less than one standard deviation from the mean. It is thus apparent that channels in the heuristically defined task relevant region are significantly concordant across subjects, while channels elsewhere are not.

Subset selection/validation and regions of interest

Using the results of the intersection method over all channels, we can use a data-driven method to select "best" subsets and eliminate channels that do not show promising results across subjects. This may be used directly in subset selection or in conjunction with other methods to independently validate or reduce existing ROIs without making any assumptions about channel spatial distribution (such as contiguity). In the 48-sentence experiment for example, applying the optimal cut-off of six channels across subjects ($I=6$), and applying it to all nine subjects yields a subset of 11 bipolar pairs: C3–T3, C3–T5, P3–T5, Cz–C3, Cz–P3, Cz–Pz, C4–T4, C4–T6, P4–T6, Cz–C4, and Cz–P4 (shown in Fig 5b). This specific subset is obtained by including all channels in the table of channel rankings that appear in or above the sixth rank, i.e. the union of all subject's top six channels. The channels chosen are all channels behind and including the

coronal midline, and thus the same subset we chose *a priori* as the task-relevant subset in the subset comparison above based on heuristic functional considerations (Fig. 5a). Returning to the BCI results on the matching test, we see that an optimal cut-off of three channels ($I=3$) applied to the ranking table for five BCI subjects gives a subset of 8 monopolar channels above and including the fifth rank across subjects: C3, CP3, CPz, CP2, C2, CP4, C4, and FC4 (shown in Fig. 5d). These are a perfect subset of the task-relevant subset chosen *a priori* in Lal et al. (2004) localized bilaterally over the motor cortices (Fig. 5c).

A final remark on the subset selection is necessary. Although we are applying a truncation rank across the ranking table at the same level (truncating below the same rank for all subjects), it is entirely possible, indeed likely, that the border separating the optimal subset in the ranking table from the remaining channels might be "jagged". When we cut straight across the table at a particular rank, constraining the number of top channels included for each subject to be the same, we introduce a slight liberal bias (towards being overly inclusive), as a channel that ranks high for only one of the subjects may end up lying above the rigid truncation rank. Empirically this most often occurs in the last row just above the cut-off rank. In our simulations and in the experimental data sets, we have found that removing channels appearing only once above the truncation rank just above the cut-off alleviates this slight bias—eliminating channels that were included once because they appeared in the same row as concordant channels at the boundary. Even with the bias, however, our method does quit well in the simulations below. We have added further discussion of this point to the Supplementary information.

Further validation on simulated data

While the intersection method certainly seems to pick subsets consistent with our anatomical and functional examples, as these are real-world data sets we cannot compare the results of our method to a known best subset of channels. For this reason, we also validate out method on two simulated data sets in which the best channels are known by construction. For comparison, we apply the *t*-test, Wilcoxon rank sum test, and our partial ranking method on both of the simulated data sets. The *t*-test and Wilcoxon rank sum test both test the null hypothesis that the classification rates obtained are not distinguishable from chance level (which for these binary data sets is 50%).

Simulated data set 1: basic model

To create the first set of simulated data we used MATLAB code available online (http://www.cs.bris.ac.uk/~rafal/phasereset/), which has been used previously in (Yeung et al., 2004; Yeung et al., 2007). This code generates single EEG trials by adding statistically independent signal and noise components. The noise components are generated by summing together 50 sinusoids of randomly varying frequency and phase (frequencies varied randomly between 0.1 and 125 Hz, phases varied randomly between 0 and 2 pi). The maximum amplitude of any single frequency component of the background EEG at 0.1 Hz is set to be 20 mV, and given this constraint, the amplitude of the sinusoid generated at each frequency is scaled to match the power spectrum of the EEG estimated from empirical data. More details concerning the generation of the EEG noise are available on the Web site and in the published papers.

We generated two different classes of 1000 ms signal trials (each down-sampled 4 times to yield 250 time points per trial). The first class had signal peaks at 100 ms and 150 ms, the second at 150 ms and 200 ms. Each signal peak was generated using the appropriate MATLAB function, which created a sinusoid—half of the cycle of which was taken to form the peak. For the first class, the peak at 100 ms had a frequency of 8 ms and the second peak at 150 ms a frequency of 40 ms. In the second class trials, the peak at 150 ms had a frequency of 40 ms and the second peak at 200 ms a frequency of 8 ms. All peaks had a normal random per-trial jitter with a standard deviation of 32 ms. To add inter-subject differences, we scaled the first peak in each trial type by a uniformly random scalar between 1 and 5.

For this artificial data set, we generated 300 trials of each class at 31 channels for 10 different subjects. We added signal trials to noise trials to get channels 1 through 8 and left the remaining channels as noise only. To add inter-subject variation per channel we added signal to between 1 and 3 of channels 9 through 31, chosen randomly for each subject. Thus each subject had signal on channels 1–8 and 1 to 3 channels numbered greater than 8, and noise on the remainder. It is important to note that data at each channel was generated separately, so there are no spatial corre-

Table 1
*Includes fractions for each channel numbered >8 representing how many subjects had randomly added signal on those channels

t-test			Wilcoxon			Partial ranking		True*
α=05	α=.01	α=.0016	α=.05	α=.01	α=.0016	Strd	Corr	
1	1	1	1	1		1	1	1
2	2	2	2	2		2	2	2
3	3	3	3	3		3	3	3
4	4	4	4	4		4	4	4
5	5	5	5	5		5	5	5
6	6	6	6	6		6	6	6
7	7	7	7	7		7	7	7
8	8	8	8	8		8	8	8
9	9	9	9	9				
11	11		11	11		11		1/10
12			12			12		2/10
								1/10
14			14					1/10
15	15		15	15				2/10
16	16		16	16				
17			17					
18	18		18	18				1/10
19						19		3/10
20			20					
21	21	21	21	21				1/10
22	22		22					
								1/10
								1/10
27			27					1/10
								2/10
29	29		29	29				1/10
30	30	30	30	30				
31	31	31	31	31				

lations between channels. For each subject, there are either channels with or without signal, and across all 10 subjects the first eight channels carry signal. While very artificial, the data set allows us to assess how well our method and other methods are able to identify channels that we know to be best across subjects (here the first eight channels).

After generating 300 single-trials per class per subject, we used linear discriminant classification (LDC) (implemented in MATLAB's "classify" function) to obtain classification rates for each individual channel within each subject. Having this table of channel performance data (here classification rates) by channel and subject, we would now like to determine which channels are "best" across subjects. Table 1 shows the results of applying (1) *t*-tests across subjects for each channel, (2) Wilcoxon rank sum tests across subjects for each channel, and (3) our partial ranking method. For (1) and (2), we show which channels would be chosen at three standard *p*-value thresholds: $\alpha=0.05$, $\alpha=0.01$, and $\alpha=0.0016$ (the conservative Bonferroni-corrected alpha level for 31 tests). For each column representing a test at a specific threshold, each row entry is shaded and shows the channel number for channels included in the chosen subset by the test. The top two panels of Fig. 6 show the intersection rates and standard deviations for the partial ranking method applied to this data—both of which show a clear peak at $I=8$. This gives the subset of 11 channels show in the partial ranking (standard—"Strd") column Table 1. In the last column, we present the "true" channel values, i.e. which channels had signal on them. We include fractions in this column indicating the presence of randomly added signal on channels numbered >8, with the fraction indicating how many of the 10 subjects had randomly added signal on that particular channel Table 2.

In terms of choosing the correct channels 1–8, partial ranking (Strd) dominates both of the other methods at all thresholds, with the Bonferroni-corrected *t*-test coming in a close second with the 12 channel subset shown in the table. While decreasing the significance level for the *t*-test might make it drop further noise channels, it is very uncommon for researchers to use *p*-values correction methods more conservative than the Bonferroni, and thus without knowing the answer in advance as we do here, it is likely the best possible result one could obtain in practice using the *t*-test. A Bonferroni correction on the Wilcoxon test actually thresholds below any of the obtained *p*-values, choosing none of the channels.

We also include partial ranking results corrected for the small possible bias resulting from the rigid truncation rate threshold (discussed above) in column "partial ranking (Corr)". As channels 11, 12, and 19 all only appear once above the truncation rank $I=8$ and appear as the very last rank above this cut-off (i.e., they rank 8th), applying our heuristic correction from above improves the subset prediction, yielding a subset of only channels 1–8.

Simulated data set 2: physical model

For the second, physically based simulation, we used a four-sphere model of the head (brain, CSF, skull and scalp) with the spherical shells at radii r(brain)=8 cm, r(CSF)=8.1 cm, r(skull)=8.6 cm, and r(scalp)=9.2 cm (Nunez and Sirinivasan, 2006). The conductivity ratios were assumed to be brain to CSF=0.2 and brain to scalp=1 for all subjects, and brain to skull uniform randomly distributed with average 40 and variance 4. Two radial dipole sources were placed 4.2 cm below the outer sphere in the triangle areas defined by electrodes Fp1, Fp2, and Fz (dipole 1), and Pz, O1, and O2 (dipole 2). This placement was such that we would expect these six channels to be the "best" subset across subjects; however, the random variation in the dipoles would

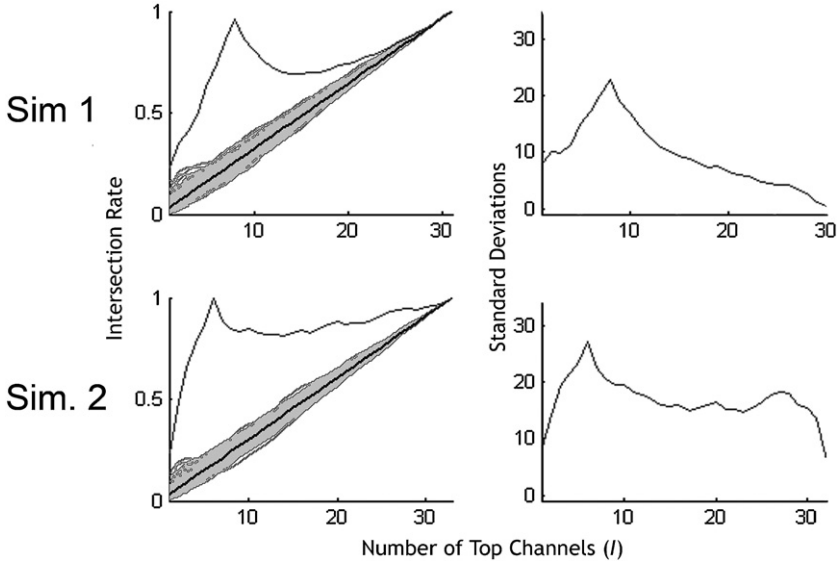


Fig. 6. Intersection method results for the two simulated data sets. Results of applying the intersection method are shown here for both the "basic" and "physical" data simulations. Plots in the upper row refer to the first simulation while those in the lower row correspond to the second. In the two figures on the left, the black line is n_{exp}, corresponding to the observed data, and the other 1000 light grey lines are the simulated null distribution n_{sim}. In both we show the intersection rate as a function of I, the number of "top" channels being compared. In the figures on the right, we show the number of standard deviations the intersection rate of the observed data was from the expectation under the null hypothesis, also as a function of channel number. The plots yield unambiguous maxima.

Table 2

t-test			Wilcoxon			Partial ranking		True*
α=.05	α=.01	α=.0015	α=.05	α=.01	α=.0015	Strd	Corr	
Fp1	Fp1	Fp1	Fp1	Fp1		Fp1	Fp1	Fp1
Fp2	Fp2	Fp2	Fp2	Fp2		Fp2	Fp2	Fp2
F9	F9	F9	F9	F9				
F7	F7	F7	F7	F7				
F3	F3	F3	F3	F3				
Fz	Fz	Fz	Fz	Fz		Fz	Fz	Fz
F4	F4		F4	F4				
F8	F8	F8	F8	F8				
F10	F10	F10	F10	F10				
FC5	FC5	FC5	FC5	FC5				
FC2	FC2		FC2					
FC6	FC6	FC6	FC6	FC6				
T9	T9	T9	T9	T9				
T7	T7	T7	T7	T7				
C3	C3		C3	C3				
Cz	Cz		Cz					
C4	C4	C4	C4	C4				
T8	T8	T8	T8	T8				
T10	T10	T10	T10	T10				
CP5	CP5	CP5	CP5	CP5				
			CP2					
CP6	CP6	CP6	CP6	CP6				
P9	P9	P9	P9	P9				
P7	P7	P7	P7	P7				
P3	P3	P3	P3	P3				
Pz	Pz	Pz	Pz	Pz		Pz	Pz	Pz
P4	P4	P4	P4	P4				
P8	P8	P8	P8	P8				
P10	P10	P10	P10	P10				
O1	O1	O1	O1	O1		O1	O1	O1
O2	O2	O2	O2	O2		O2	O2	O2

prevent any one channel off the six to always be best. The sources were then activated with sinusoidal waves of frequencies of 1 kHz (for class 1 trials) and 1.05 kHz (for class 2 trials). A trial of 1 s (100 frames) was generated for each class by numerically computing the potential at electrode sites at each instant. For each subject, a time series of 200 trials was built by randomly concatenating trials of class 1 and class 2. Noise was generated in a similar way – but independently of trial class – using a set of 10 radially oriented dipoles. The direction of these dipoles as well as their position inside the brain was chosen to be uniformly random at each time instant. This noise was added to the signal trials to generate the final simulated data set. We show scalp maps of signal and noise for the first three simulated subjects in Fig. 7.

Applying the same tests as we did for the first simulation, the results are dramatic. The simulation of data with more realistic spatial correlations results in the *t*-test and Wilcoxon rank sum test failing to capture the "best" channels, while the partial ranking method captures them perfectly with and without the heuristic bias correction. The bottom half of Fig. 6 shows the intersection rates and standard deviations for the partial ranking method applied to this data—which yield a clear peak at $I=6$. This gives the correct subset of 6 channels shown in the table.

It is clear from these results that applying the *t*-test or the Wilcoxon on the rates across subjects to find a best subset is sub-optimal, tending to choose too many channels, even with a conservative family-wise error correction on the *p*-value threshold employed. More important than the poor performance of these methods, however, is the positive performance of the partial-ranking intersection method, which with our simple bias correction chooses both of those subsets set out as the "best" subsets in the simulated data. Without correction, the method chooses the correct subset plus three extra channels in the first simulation and chooses the correct subset in the second simulation.

While these simulated data sets are artificial and obviously far less complex that actual EEG data, they allow us to validate the methods in a context where the "correct answer" is known.

Discussion and conclusions

We report a method for quantifying inter-subject agreement at the channel level using a simple modification of a classical rank correlation test combined with a permutation simulation of the corresponding multivariate null distribution. This "partial-ranking" method (1) allows the user to establish the existence of channel agreement across subjects—even if the agreement is on a few channels in the presence of many noisy channels, (2) gives an optimal subset size by showing at which number of "top" channels the ranks across subjects are furthest from the simulated null distribution, and (3) tells which channels are responsible for agreement across subjects, even if these channels are randomly distributed within the top channels. Additionally, it is a non-parametric approach, making it well suited to brain-imaging investigations, which often have small sample sizes and may suffer from large inter-subject and inter-channel variability (Kherif et al., 2003). As the test makes no prior assumptions about the distribution of "best" channel subsets (e.g., that they will be contiguous), it may act as a conservative subset selection or ROI validation tool, including only those channels that are most concordant across subjects. Our application of the method to two EEG classification data sets yielded very significant and well-localized channel subsets consistent with *a priori* hypotheses about "best" channel subset distribution—in one case reducing subset size markedly.

Classical Spearman ranking successfully detected concordance in both data sets examined here, but as it compares entire channel sets, did not naturally provide information useful for subset selection. Our partial-ranking "intersection method", however, performed well on both data sets, yielding clear peaks in the number of standard deviations from the simulated null distribution at specific subsets sizes (a threshold of $I=6$ top channels yielding a 11 bipolar channel subset for the 48-sentence experiment, and a threshold of $I=3$ top channels yielding 8 monopolar channel subset for the BCI experiment). The method appears well suited to choosing/validating "best" channel subsets and across subjects.

For both experiments, the subsets chosen by the intersection method were spatially localized and closely resembled subsets chosen *a priori* as likely "best" subsets using functional and anatomical criteria (Figs. 5a and c). The relative contiguity of the channels in both experiments is interesting as our method does not require the subset to be contiguous in any way. In the case of the 48-sentence data, the subset chosen by the intersection method was identical to the one chosen *a priori* given functional considerations, while for the BCI experiment a reduced subset of the anatomically chosen channels resulted. For both experi-

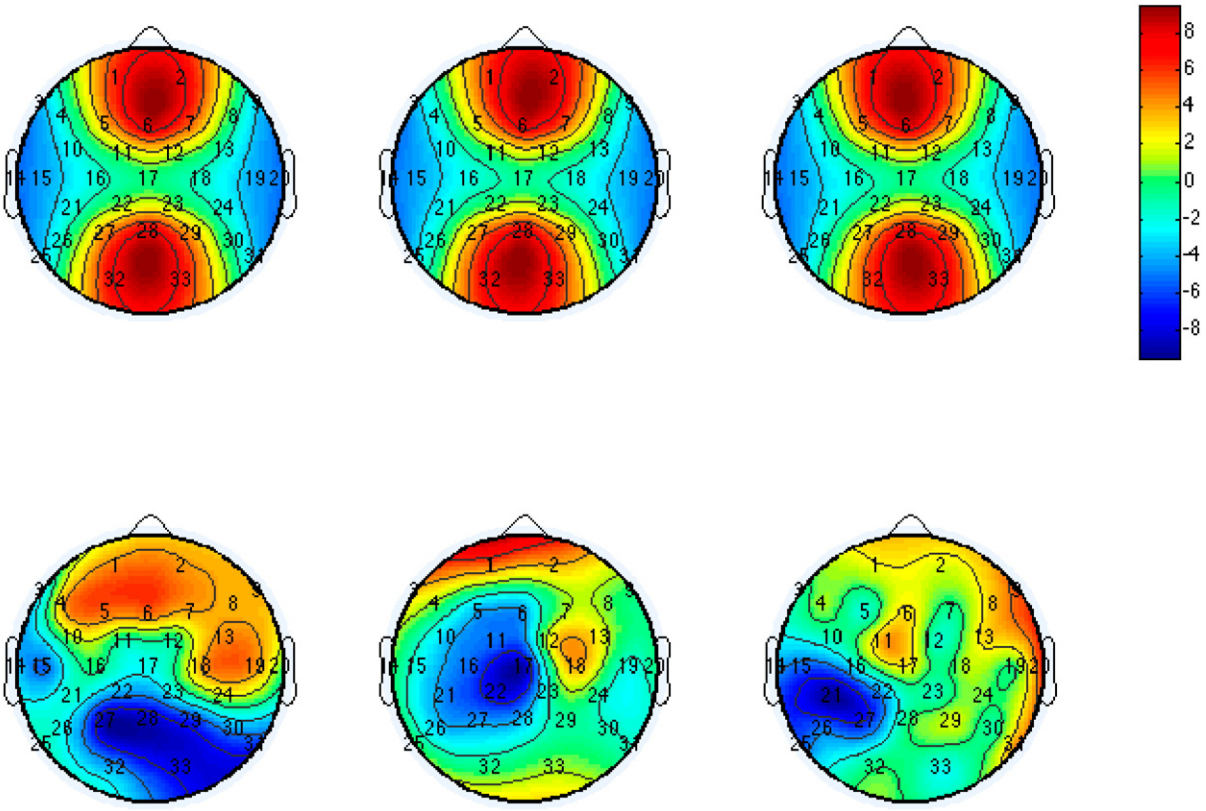


Fig. 7. Scalp Maps for the Simulation 2 ("Physical" data). The top row plots the time average of the "signal" part of the data for the 1st three simulated subjects, with radial dipoles centered under the triangles defined on the scalp surface by channels Fp1, Fp2, and Fz, for the first dipole and Pz, O1, and O2 for the second. Small random movements of the dipoles prevent any one of the six channels from being uniformly best across subjects. The bottom row plots the time average of the additional additive "noise" (beyond the noise already introduced by the four-sphere model) for the 1st three simulated subjects, generated using a set of 10 radially-oriented random dipoles.

ments, these subsets showed highly significant intersection rates, with standard deviations of approximately 15 and 20 from the simulated multivariate null distribution for the 48-sentence and BCI cases, respectively. Such rates indicate a significant concordance in the sensors appearing above the chosen number of channels I across subjects. Thus, for these data, the intersection method confirms that strongly concordant subsets exist across subjects and provides a clear non-parametric approach to the selection of "best" concordant subsets across subjects. The 48-sentence data also provides a good example of a situation in which no single best channel emerges—instead concordance emerges only as we evaluate channel subsets.

Finally, we ran a validation on two simulated data sets where the set of best channels across subjects was known by construction. Our method consistently outperformed simple thresholding approaches using either a t-test or a Wilcoxon rank sums test to look for channels performing at above chance level. Most importantly, our method performed well at "discovering" the known subsets, choosing both perfectly when a slight correction for the rigid cut-off across subject was employed.

This is, to the best of our knowledge, the first statistical confirmation of channel subset concordance across subjects. With its lack of assumptions and subject-by-subject comparisons, this method provides a stringent and robust test that should provide a useful way for investigators using EEG, MEG, and fMRI to choose or validate channel subsets using hypothesis-testing methods.

Acknowledgment

We thank Thomas Nevin Lal for providing permission to use the BCI data.

Appendix A

A.1. Derivation of Eqs. (2) and (3)

For Eq. (3): Given a ranking function $r(\bullet)$ which takes a vector and returns a ranking of its elements, consider the ranks $r(\mathbf{X})=(r(\mathbf{X}_1),\ldots,r(\mathbf{X}_N))$ and $r(\mathbf{Y})=(r(\mathbf{Y}_1),\ldots,r(\mathbf{Y}_N))$, of some observations $\mathbf{X}=(\mathbf{X}_1,\ldots,\mathbf{X}_N)$ and $\mathbf{Y}=(\mathbf{Y}_1,\ldots,\mathbf{Y}_N)$ of length N. Channel subsets up to and including those of rank I can be defined as $S_I[r(\mathbf{X})]$ and $S_I[r(\mathbf{Y})]$, where channel $i\in S_I[r(\mathbf{X})]$ if $1\leq r(\mathbf{X}_i)\leq I$ for $i\in\{1,\ldots,N\}$, and similarly for $S_I[r(\mathbf{Y})]$.

The number of intersections m_I can be computed by the cardinality of the intersection of the sets $S_I[r(\mathbf{X})]$ and $S_I[r(\mathbf{Y})]$,

$$m_I(\mathbf{X},\mathbf{Y}) = \left|S_I[r(\mathbf{X})]\cap S_I[r(\mathbf{Y})]\right|$$

In deriving Eq. (3), there are three terms of interest. The first term ${}_IC_k$ is the number of combinations of k channels chosen from

a set of I. The second term is the number of combinations that $(N-k)$ channels can be chosen from the remaining set of $(N-I)$ channels. The last term is the portion of the restrictive permutations of the two-element partitions of I and $(N-I)$ out of all the possible non-restrictive permutations of N channels. Based on these three terms, the distribution of k for block size I based on the null hypothesis $f_I(k)$ can be computed by:

$$f_I(k) = \frac{I}{k!(I-k)!}\frac{(N-1)!}{(I-k)(N-2I+k)!}\frac{I!(N-I)}{N!}$$

$$= \frac{[I!(N-I!)^2]}{N!k![(I-k)]^2(N-2I+k)!}$$

The p-value for an intersection of fixed size c can then be obtained in the standard way by summing the probabilities of intersections of size c or greater under the null distribution. The extension of this bivariate to the multivariate case is non-trivial (Oja, 1999; Stepanova, 2003) but an approximation of such distribution can be established by simulation. Eq. (2) may now be found by computing the first moment of Eq. (3) (in the case $I=J$).

A.2. Algorithm for simulating the multivariate null distribution for the intersection method

Pseudo-code for creating null distributions:

1. Create Q random matrices $\boldsymbol{R}_{1\ldots Q}$, each of size $N_C \times N_S$, with each column of a particular matrix $\boldsymbol{R}_q$ containing a random permutation of the ranks 1 to N_C (where $q \in \{1,\ldots,Q\}$, N_C is the number of channels, and N_S the number of subjects).
2. For all disjoint, two-element, non-empty groupings of the columns of a particular matrix $\boldsymbol{R}_q$, we get two sub-matrices $\boldsymbol{R}_{q,1}$ and $\boldsymbol{R}_{q,2}$ corresponding to two groupings of the columns. Averaging across the columns of one of the sub-matrices yields an average rank based that sub-matrix ($\boldsymbol{R}_{q,1}$). We renormalize this to a rank to create $\bar{r}$.
3. For all rankings defined by a column vector $r_j$of $\boldsymbol{R}_{q,2}$, compute $M_I(q) = \sum_j m_I(\bar{r}, r) = \sum_j |S_I(\bar{r}) \cap S_I(r_j)|$ (where $S_I(\bullet)$ and $m_I(\bullet,\bullet)$ are the functions defined in the derivations of Eqs. (2) and (3) above, $q \in \{1,\ldots,Q\}$ indexes a particular matrix, and I is the variable cut-off value). As an example, looking at all values of I for a particular $M_I(q)$ would give a single curve like those shown in the null distributions of Fig. 2.
4. Repeat steps 2–3 for each random matrix $\boldsymbol{R}_q$, $q \in \{1,\ldots,Q\}$.
5. For each $q \in \{1,\ldots,Q\}$, we get a $M_I(q)$. Normalizing each value of these by dividing by the maximum intersection size yields the simulated null distribution.

By applying steps 2–3 to a particular matrix $\boldsymbol{R}_d$ of real ranking data (e.g., the matrix of ranks represented in Fig. 1), we get values for $M_I(d)$ which may be compared against the null distribution from step 5 in the standard way.

A.3. Experimental methods

The experimental methods and human subjects permissions specific to the 48-sentence and BCI experiments are readily available in (Suppes et al., 1999a; Lal et al., 2004).

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2007.07.064.

References

Best, D.J., Roberts, D.E., 1975. The upper tail probabilities of Spearman's rho. Appl. Stat. 24, 377–379.

Blankertz, B., Curio, G., Müller, K.-R., 2002. Classifying single trial EEG: toward brain computer interfacing. Proc. Adv. Neural Inf. Process. Syst. (NIPS 01) 14, 157–164.

Brammer, M.J., Bullmore, E.T., Simmons, A., Williams, S.C.R., Grasby, P.M., Howard, R.J., Woodruff, P.W.R., Rabe-Hesketh, S., 1997. Generic brain activation mapping in functional magnetic resonance imaging: a nonparametric approach. Magn. Reson. Imaging 15, 763–770.

Buchner, H., Waberski, T.D., Fuchs, M., Wischmann, H., Wagner, M., Drenckhahn, R., 1995. Comparison of realistically shaped boundary-element and spherical head models in source localization of early somatosensory evoked potentials. Brain Topogr. 8, 137–143.

Chen, Y., Seth, A.K., Gally, J.A., Edelman, G.M., 2003. The power of human brain magnetoencephalographic signals can be modulated up or down by changes in an attentive visual task. Proc. Natl. Acad. Sci. U. S. A. 100, 3501–3506.

Cox, D.D., Savoy, R.L., 2003. Function magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in the human visual cortex. NeuroImage 19, 261–270.

D'Esposito, M., Postle, B.R., Rypma, B., 2000. Prefrontal cortical contributions to working memory: evidence from event-related fMRI studies. Exp. Brain Res. 133, 3–11.

Eickhoff, S.B., Weiss, P.H., Amunts, K., Fink, G.R., Zilles, K., 2005. Identifying human parieto-insular vestibular cortex using fMRI and cytoarchitectonic mapping. Hum. Brain Mapp. 27, 611–621.

Etkin, A., Egner, T., Peraza, D.M., Kandel, E.R., Hirsch, J., 2006. Resolving emotional conflict: a role for the rostral anterior cingulate cortex in modulating activity in the amygdala. Neuron 51, 871–882.

Fieller, E.C., Hartley, H.O., Pearson, E.S., 1957. Tests for rank correlation coefficients. Biometrika 44, 470–481.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.B., Frith, C.D., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210.

Grill-Spector, K., Kanwisher, N., 2005. Visual recognition: as soon as you see it, you know what it is. Psychol. Sci. 16, 152–160.

Grill-Spector, K., Kushnir, T., Edelman, S., Itzchak, Y., Malach, R., 1998. Cue-invariant activation in object-related areas of the human occipital lobe. Neuron 21, 191–202.

Guger, C., Ramoser, H., Pfurtscheller, G., 2000. Real-time EEG analysis with subject-specific spatial patterns for a brain-computer interface (BCI). IEEE Trans. Rehabil. Eng. 8, 447–456.

Hagoort, P., 2003. Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. J. Cogn. Neurosci. 15, 883–899.

Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M., 2004. Integration of word meaning and world knowledge in language comprehension. Science 304, 438–441.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York.

Hotelling, H., Pabst, M.R., 1936. Rank correlation and tests of significance involving no assumption of normality. Ann. Math. Stat. 7, 29–43.

Huettel, S.A., McKeown, M.J., Song, A.W., Hart, S., Spencer, D.D., Allison, T., McCarthy, G., 2004. Linking hemodynamic and electro-physiological measures of brain activity: evidence from functional MRI and intracranial field potentials. Cereb. Cortex 14, 165–173.

Jasper, H., 1958. The 10–20 electrode system of the international federation. Electroencephaologr. Clin. Neurophysiol. 10, 371–375.

Joe, H., 1990. Multivariate concordance. J. Multivar. Anal. 35, 12–30.

Johnson-Frey, S.H., Maloof, F.R., Newman-Norlund, R., Farrer, C., Inati, S., Grafton, S.T., 2003. Actions or hand-object interactions? Human inferior frontal cortex and action observation. Neuron 39, 1053–1058.

Kherif, F., Poline, J.-B., Meriaux, S., Benali, H., Flandin, G., Brett, M., 2003. Group analysis in functional neuroimaging: selecting subjects using similarity measures. NeuroImage 20, 2197–2208.

Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30, 81–93.

Kendall, M.G., Smith, B.B., 1940. On the method of paired comparisons. Biometrika 31, 324–345.

Kendall, M.G., Kendall, S.F.H., Smith, B.B., 1939. The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. Biometrika 30, 251–273.

Kruskal, W.H., 1958. Ordinal measures of association. J. Am. Stat. Assoc. 53, 814–861.

Lal, T.N., Schroder, M., Hinterberger, T., Weston, J., Martin, B., Birbaumer, N., Scholkopf, B., 2004. Support vector channel selection in BCI. IEEE Trans. Biomed. Eng. 51, 1003–1010.

Lazar, N.A., Luna, B., Sweeney, J.A., Eddy, W.F., 2002. Combining brains: a survey of methods for statistical pooling of information. NeuroImage 16, 538–550.

Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation. Springer, New York.

Li, Y., Gao, X., Liu, H., Gao, S., 2004. Classification of single-trial electroencephalogram during finger movement. IEEE Trans. Biomed. Eng. 51, 1019–1025.

Li, Y., Dong, G., Gao, X., Gao, S., Ge, M., Yan, W., 2005. Single-trial EEG classification during finger movement task by using hidden Markov models. Proc. 2nd Int. IEEE EMBS Conference Neural 625–628.

Liu, J., Harris, A., Kanwisher, N., 2002. Stages of processing in face perception: and MEG study. Nat. Neurosci. 5, 910–916.

Meriaux, S., Roche, A., Thirion, B., 2006. Robust statistics for nonparametric group analysis in fMRI. Proc. 3rd IEEE 936–939.

Moran, P.A.P., 1947. On the method of paired comparisons. Biometrika 34, 363–365.

Nashmi, R., Mendonca, A.J., MacKay, W.A., 1994. EEG rhythms of the sensorimotor region during hand movements. Electroencephalogr. Clin. Neurophysiol. 91, 456–467.

Nunez, P.L., Sirinivasan, R., 2006. Electric Fields of the Brain: The Neurophysics of EEG, 2nd ed. Oxford Univ. Press.

O'Craven, K.M., Kanwisher, N., 2000. Mental images of faces and places activates corresponding stimulus-specific brain regions. J. Cogn. Neurosci. 12, 1013–1023.

Oja, H., 1999. Affine invariant multivariate sign and rank tests and corresponding estimates: a review. Scand. J. Statist. 26, 319–343.

Oostenveld, R., Praamstra, P., 2001. The five percent electrode system for high-resolution EEG and ERP measurements. Clin. Neurophysiol. 112, 713–719.

Perreau-Guimaraes, M., Wong, D.K., Uy, E.T., Grosenick, L., Suppes, P., 2007. Single-trial classification of MEG recordings. IEEE Trans. Biomed. Eng. 54, 436–443.

Phelps, E.A., Delgado, M.R., Nearing, K.I., LeDoux, J.E., 2004. Extinction learning in humans: role of the amygdala and vmPFC. Neuron 43, 897–905.

Polich, J., 1997. EEG and ERP assessment of normal aging. Electroencephalogr. Clin. Neurophys. 104, 244–256.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Schendan, H.E., Searl, M.M., Melrose, R.J., Stern, C.E., 2003. An fMRI study of the role of the medial temporal lobe in implicit and explicit sequence learning. Neuron 37, 1013–1025.

Schroder, M., Bogdan, M., Rosenstiel, W., Hinterberger, T., Birbaumer, N., 2003. Automated EEG feature selection for brain computer interfaces. Proc. 1st Int. IEEE EMBS Conf. Neural Eng. 626–629.

Schwartz, S., Maquet, P., Frith, C., 2002. Neural correlates of perceptual learning: a functional MRI study of visual texture discrimination. Proc. Natl. Acad. Sci. U. S. A. 99, 17137–17142.

Sereno, M.E., Trinath, T., Augath, M., Logothetis, N.K., 2002. Three-dimensional shape representation in monkey cortex. Neuron 33, 635–652.

Shaffer, J., 1995. Multiple hypothesis testing. Annu. Rev. Psychol. 46, 561–584.

Simon, O., Mangin, J., Cohen, L., Bihan, D.L., Dehaene, S., 2002. Topographical layout of hand, eye, calculation, and language-related areas in the human parietal lobe. Neuron 33, 475–487.

Stark, C.E.L., Squire, L.R., 2001. When zero is not zero: the problem of ambiguous baseline conditions in fMRI. Proc. Natl. Acad. Sci. U. S. A. 98, 12760–12766.

Stepanova, N.A., 2003. Multivariate rank tests for independence and their asymptotic efficiency. Math. Methods Stat. 12, 197–217.

Stephens, L.J., Claypool, P.L., Buchalter, B., 1977. Partial ordering of populations. J. Educ. Stat. 2, 41–53.

Student, 1921. An experimental determination of the probable error of Dr. Spearman's correlation coefficients. Biometrika 13, 263–282.

Suppes, P., Han, B., Epelboim, J., Lu, Z., 1999a. Invariance between subjects of brain-wave representations of language. Proc. Natl. Acad. Sci. U. S. A. 96, 12953–12958.

Suppes, P., Han, B., Epelboim, J., Lu, Z., 1999b. Invariance in brain-wave representations of simple visual images and their names. Proc. Natl. Acad. Sci. U. S. A. 96, 14658–14663.

Suppes, P., Han, B., Lu, Z., 1998. Brain-wave recognition of sentences. Proc. Natl. Acad. Sci. U. S. A. 95, 15861–15866.

Tang, A., Sutherland, M.T., Wang, Y., 2006. Contrasting single-trial ERPs between experimental manipulations: improving differentiability by blind source separation. NeuroImage 29, 335–346.

Taylor, J.E., Worsley, K.J., Gosselin, F., 2007. Maxima of discretely sampled random fields, with an application to 'bubbles'. Biometrika 94 (1), 1–18.

Van Boven, R.W., Ingeholm, J.E., Beauchamp, M.S., Bikle, P.C., Ungerleider, L.G., 2005. Tactile form and location processing in the human brain. Proc. Natl. Acad. Sci. U. S. A. 102, 12601–12605.

Wong, D.K., Perreau-Guimaraes, M., Uy, E.T., Suppes, P., 2004. Classification of individual trials based on the best independent component of EEG-recorded sentences. Neurocomputing 61, 479–484.

Yeung, N., Bogacz, R., Holroyd, C.B., Cohen, J.D., 2004. Detection of synchronized oscillations in the electroencephalogram: an evaluation of methods. Psychophysiology 41, 822–832.

Yeung, N., Bogacz, R., Holroyd, C.B., Nieuwenhuis, S., Cohen, J.D., 2007. Theta phase-resetting and the error-related negativity. Psychophysiology 44, 39–49.