Statistica Sinica **18**(2008), 1519-1534

SCOPE OF RESAMPLING-BASED TESTS IN fNIRS NEUROIMAGING DATA ANALYSIS

Archana K. Singh[1,2], Lester Clowney[1], Masako Okamoto[1], James B. Cole[2] and Ippeita Dan[1]

[1]*National Food Research Institute and* [2]*University of Tsukuba, Japan*

Abstract: Functional near infrared spectroscopy (fNIRS) is an emerging non-invasive optical technique to monitor the cortical hemodynamic response. Generally, parametric statistical methods are used to analyze fNIRS data, requiring certain strong assumptions that may fail in fNIRS data. This paper illustrates the application of non-parametric alternatives, such as permutation and bootstrap methods, which require fewer and weaker assumptions. We demonstrate that the proposed methods can increase the statistical significance of results when compared to the equivalent parametric methods in controlling familywise error rate in fNIRS group studies.

Key words and phrases: Adjusted p-values, maximum t correction, multiple comparison, multiple testing problem, non-parametric test, exchangeability, Type I error, optical imaging.

1. Introduction

Functional near infrared spectroscopy (fNIRS) is an emerging optical imaging technique that non-invasively monitors the brain activity using near infrared light (NIR). A neuronal activity in the brain evokes the regional cerebral blood flow (neurovascular coupling). As a result, there is a localized change in the concentration of oxygenated and deoxygenated hemoglobins (HbO_2 and HbR), which are the dominant absorbers of NIR light in the brain tissue (Cope et al. (1988)). In fNIRS, a pair of illuminator (light emitting diode) and detector are affixed to the scalp. The illuminator emits NIR light through the scalp, which scatters and penetrates the underlying head and brain tissue, and the detector detects the light that reflects back to the scalp. fNIRS measures the relative changes in hemoglobin components as a function of the change in light intensity during emission and detection. The midpoint of an illuminator-detector pair defines a channel, which is used as a reference to locate the brain activity. Multichannel fNIRS is used to monitor many brain locations simultaneously by setting up a probe holder on the head surface that contains several pairs of light illuminators and detectors. fNIRS has several advantages over other non-invasive functional

neuroimaging techniques, e.g., functional magnetic resonance imaging (fMRI) and positron emission tomography (PET). It is compact, portable, and relatively more tolerant of body movement, enabling a wide range of experimental applications in neuropsychological and diagnostic situations (reviewed in Koizumi et al. (2003), Obrig and Villringer (2003) and Strangman, Boas and Sutton (2002)).

As in the case of fMRI and PET studies, a two-level summary statistics approach based on the random effects general linear model (RFX) is commonly used in fNIRS (Schroeter et al. (2004)). This model accounts for both within-subject and between-subject variability, and thus extrapolates the inference at the population level (Friston et al. (1995)). At the first level, subjects' averages are computed for each condition to summarize within-subject effect, and then a t-test is performed to detect whether the subjects respond differently under experimental and control conditions. The family wise error rate (FWER) control in multiple channel testing is often applied using Bonferroni correction.

The above model relies on parametric tests, and thus requires certain distributional assumptions that may be too strong in fNIRS. The distribution of the measured hemoglobin signal is associated with many uncertainties. The mechanism of neurovascular coupling that causes the change in hemoglobin concentration is not well understood (Steinbrink et al. (2006) and Villringer and Dirnagl (1995)). These changes also depend on certain unknown factors, such as optical path length, and thickness of the brain tissues (Hoshi (2003) and Okada and Delpy (2003)). In addition, neuroimaging studies often employ only a few subjects. In fNIRS, their response data vary a lot even if they are sampled from the same head location under identical conditions. In the case of the two-sample t-test, we test the equality of sample means assuming independent and identical error variances. Since this test is performed on the subjects' averages at the second level of GLM, where the subjects are conventionally assumed to be independent, the assumption of independence is tenable. However, the assumption of identical variances cannot be ascertained. In addition, fNIRS allows the monitoring of subjects who cannot be prohibited from moving (e.g., in the case of awake infants), which adds motion-related noise. Removal of this noise results in unbalanced datasets with unequal sample sizes between conditions, which may further aggravate the possibility of failure to meet the assumption of identical variances.

Thus, it is important to explore statistical options that can deal with such possible failures of assumptions in practical fNIRS datasets. In this paper, we examine the potential scope of resampling-based methods, such as the permutation and bootstrap tests on multichannel and multisubject fNIRS data, to relax such assumptions, focusing primarily on the independent two-sample problem. We also explore the FWER controlling property of a resampling-based multiple

testing correction, maximum t (max t) correction (Westfall and Young (1993)), which can deal with any underlying spatial correlation structure. We applied the two-sample permutation and bootstrap tests, including the max t correction, using an unbalanced dataset acquired from an fNIRS group experiment. In the experiment, we performed the standard neuropsychological task of word generation that is known to activate the motor speech (Broca's) area in the left ventrolateral prefrontal cortex. In addition, we also used simulation to compare the power of the two resampling-based methods. The results obtained from real and simulated data show that the proposed two-sample permutation and bootstrap tests can detect 50% more channels than traditional parametric tests after applying familywise error correction.

2. Conventional Parametric Framework

Suppose the subjects' averages for task and baseline conditions are represented by two samples, $X = (X_1, \ldots, X_{N_1})$ and $Y = (Y_1, \ldots, Y_{N_2})$, with sizes N_1 and N_2, from normal populations with means and variances $\mu_X,\ \mu_Y,\ \sigma_X^2\ and\ \sigma_Y^2$, respectively. The null hypothesis is, $H_0 : \mu_X = \mu_Y$. For paired samples, $N_1 = N_2 = N$, and a one-sample t-test is used on the difference $(X - Y)$ between pairs with $N - 1$ df. If the samples are independent and have homogeneous variances, a two-sample t-test is used. The p-value is obtained using Student's t distribution with $N_1 + N_2 - 2$ df. The validity of the test depends on how well its assumptions are met.

$$t = \frac{(\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{N_1} + \frac{\sigma_Y^2}{N_2}}} \sim \quad t_{dist}(N_1 + N_2 - 2). \tag{2.1}$$

3. Resampling-Based Framework

In a resampling-based test, the inference is based upon repeated sampling of the observed data. The resampling is done without replacement in a permutation test and with replacement in a bootstrap test. While the resampling-based methods have been around since permutation tests were first reported (Fisher (1935) and Pitman (1937)), bootstrap methods have been developed more recently (Efron (1979)). A permutation test is exact but requires the assumption of exchangeability (for definition refer to Section 3.1), which is not applicable in certain situations, e.g., for testing the equality of sample means without assuming identical distributions in a two-sample t-test. A bootstrap test is more flexible. It can accommodate a broader range of situations depending on the null hypothesis being tested, including those that fail to justify exchangeability (ref. Section 3.3). In the following sections, we discuss several resampling-based alternatives for emulating two-sample t-test.

3.1. Permutation tests

In the permutation test, the null distribution is generated by random permutation of the data. Suppose that subjects' averages for task, $X = (X_1, \ldots, X_{N_1})$, are distributed as F, while for baseline, $Y = (Y_1, \ldots, Y_{N_2})$, are distributed as G, with $N_1 \geq N_2$. If $Z = \{X_1, \ldots, X_{N1}, Y_1, \ldots, Y_{N2}\}$ represents the union of the two samples, the observations in Z can be permuted in $NP = (N_1 + N_2)!/N_1!N_2!$ possible ways. If the null hypothesis is true, the equality $F = G$ will assure an equal probability $1/NP$ to the observations in any subset of the common sample space of observations in X and Y. The labels X and Y will not affect their joint outcome, therefore we can label the first N_1 of the pooled observations as X, the rest as Y, and compute the t-statistic using a large number of permutations to find the null distribution. The p-value is obtained as the proportion of the resampled t^* that are as extreme as or larger than the observed t (Efron and Tibshirani (1986)). A sufficient condition for the inference obtained from the permutation test to be exact and unbiased is the exchangeability of observations with respect to the observations in the pooled sample Z. The observations in Z are exchangeable if the probability of any joint outcome of these observations, such as a t-statistic, is invariant to their permutations. In a two-sample comparison problem, exchangeability holds if the observations are independent and identically distributed.

However, there is a special case for comparing means of paired observations, in which the exchangeability is granted under a milder assumption of symmetry. In the one-sample (paired) t-test on the difference between the pairs, $D = X - Y$, and $N_1 = N_2 = N$, the null hypothesis is that the $D's$ are symmetrically distributed around zero, i.e., there is no activation. Under the null hypothesis, $(X-Y)$ and $(Y-X)$ are the same, and changing the signs of observations in D will not affect the t-statistic. For a sample size of N, we can re-arrange observations in D by randomly prefixing them with a plus or a minus sign in 2^N equally possible ways (Good (2000), Holmes et al. (1996) and Nichols and Holmes (2002)). If X and Y are drawn from different symmetric distributions and they have the same mean (but possibly different variances) then $X - Y$ and $Y - X$ will have the same distribution, and hence the permutation test that flips signs on the values of $D's$ is exact.

3.2. Bootstrap tests under exchangeability

This bootstrap test is used for testing the null hypothesis $Ho : F = G$. The resampling procedure is similar to that of the permutation test as described in the previous section, except that bootstrap resamples of X and Y are drawn with replacement from the pooled sample Z. The bootstrap null distribution can be obtained by computing t-statistic from these bootstrap resamples (Efron

and Tibshirani (1986)). Like the permutation test, validity relies on how well exchangeability holds under the null hypothesis. If the exchangeability condition applies, the null distributions from the two-sample permutation and bootstrap test are asymptotically equivalent (Romano (1989, Figure 1A 2B)). In this article, we use the abbreviation 'exchangeable bootstrap' for this test.

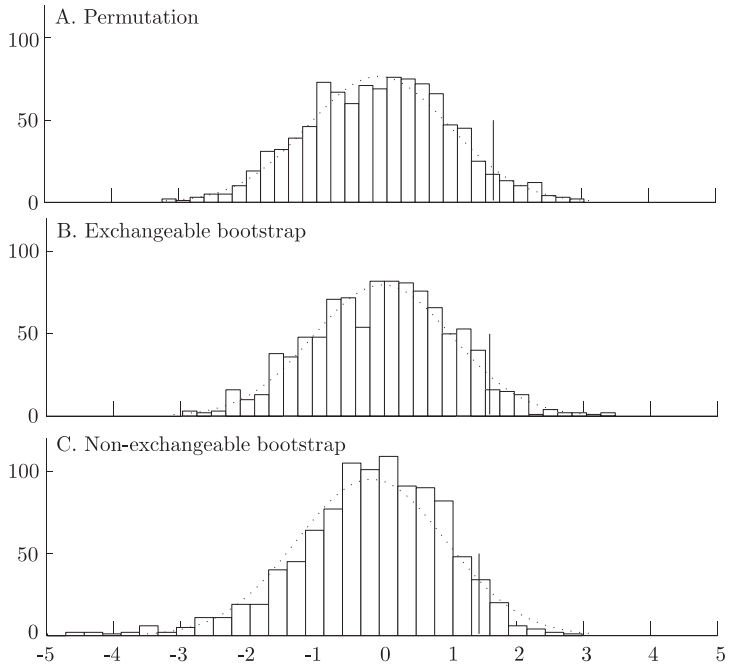

Figure 1. Histograms for the distribution of t-statistics obtained from the two-sample t-test statistic using permutation (A), exchangeable bootstrap (B), and non-exchangeable bootstrap (C) methods. All these tests used 1,000 resamples. The vertical line indicates the corresponding 5% threshold.

3.3. Bootstrap tests under possible non-exchangeability

This test is used for testing the equality of the sample means, without assuming equal distribution. Suppose the two samples X and Y are mutually independent, and the observations are independent and identically distributed within each sample. We construct two independent bootstrap resamples, X^* and Y^*, by randomly drawing N_1 values from X and N_2 values from Y, respectively, with replacement. Repeating this B times and computing the t-statistic from the resampled sets each time generates the bootstrap null distribution of t^*. In order to reflect the null hypothesis, $H_0 : \mu_X - \mu_Y = 0$ in the bootstrap scheme,

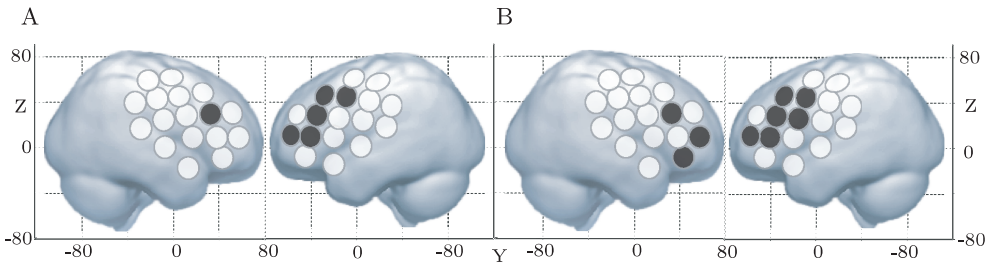

Figure 2. Analysis of unbalanced WG data. The spatial visualization of the results was done by registering the datasets in MNI space using our fNIRS registration program (Okamoto et al. (2004), Okamoto et al. (2005) and Singh et al. (2005)). The channels that are detected active are shown in black. The left column (A) depicts the results obtained from the parametric t-test after Bonferroni correction at 5% threshold. The permutation, exchangeable bootstrap, and non-exchangeable bootstrap t-tests, after max2 t FWER control detected the same channels, as shown in the right column (B).

we subtract μ_X^* from μ_X and μ_Y^* from μ_Y in (2.1), as

$$t^* = \frac{(\mu_X^* - \mu_X) - (\mu_Y^* - \mu_Y)}{\sqrt{\frac{\sigma_X^{*\,2}}{N_1} + \frac{\sigma_Y^{*\,2}}{N_2}}}, \tag{3.1}$$

where μ_X^*, μ_Y^*, $\sigma_X^{*\,2}$, and $\sigma_Y^{*\,2}$ are the mean and variance of the bootstrapped samples X^* and Y^* respectively. This method has been referred to as the shift and pivot method in Westfall and Young (1993) in the one-sample situation. The algorithm for the two-sample test is summarized in Efron and Tibshirani (1986). We use the abbreviation 'non-exchangeable bootstrap' to refer to this test. In case of a paired t-test with samples of equal size ($N = N_1 = N_2$), we can resample with replacement the difference between pairs after shifting their means, and compute a one-sample t-statistic to obtain the bootstrap null distribution.

3.4. Resampling-based FWER corrected p-values

The permutation or bootstrap distribution of the test statistic, t^*, for the null hypothesis H_{0j} ($j = 1, \ldots, C$), is given by the empirical distribution $t_{j,1}^*, \ldots, t_{j,B}^*$. For the right-tail alternative, the resampling-based uncorrected p-value corresponding to H_{0j} is calculated as $p_j^* = (1/B) \sum_{b=1}^{B} \varphi(t_{j,b}^* \geq t_j))$, where $\varphi(\cdot)$ is 1 if the condition in parenthesis is true, and 0 otherwise. The resampling-based FWER threshold can be obtained from the distribution of the maxima of resampling-based t-values using a single-step procedure (max1), or a step-wise procedure (max2) (Westfall and Young (1993, Algorithm 2.8), Holmes et al. (1996) and Dudoit et al. (2002)). First, we compute the t-statistic $t_{1,b}^*, \ldots, t_{C,b}^*$

for each of B bootstrap or permutation sets. In the single step max t procedure, a global threshold is computed from the maximal t-statistic among the channels for each b, $U_b = \max(t^*_{j,b})$, for $j = 1, \ldots, C$. The adjusted p-values are $p^*_j = (1/B) \sum_{b=1}^{B} \varphi(U_b \geq t_j)$. In the stepwise max t procedure, we compute successive maxima for each of the test-statistics

$$U_{C,b} = t^*_{r_C}; \quad U_{j,b} = \max\left(U_{j+1,b}\ ,\ t^*_{r_j,b}\right),\ \textit{for } j = C-1, \ldots, 1, \tag{3.2}$$

where rj denotes the ordering of the observed test statistic such that $t_{r_1} \geq \cdots \geq t_{r_C}$. The adjusted p-value is $p^*_{r_j} = (1/B) \sum_{b=1}^{B} \varphi\left(U_{j,b} \geq t_{r_j}\right)$. The monotonicity constraints can be enforced as

$$p^*_{r_1} \leftarrow p^*_{r_1}, \quad p^*_{r_j} \leftarrow \max\left(p^*_{r_j}, p^*_{r_j-1}\right), \qquad \textit{for } j = 2, \ldots, C. \tag{3.3}$$

4. Application to fNIRS Data

We acquired the data from an fNIRS word generation (WG) experiment that was expected to activate Broca's area. We monitored 15 healthy, right handed subjects using fNIRS topography system OMM-2000 optical multi-channel monitor (Shimadzu, Kyoto, Japan) with 34 channels that covered inferior frontal, middle frontal, precentral and postcentral gyri. Paired illuminators and detectors were located 3 cm apart. The subjects sat in a quiet room with eyes closed during the measurements. The tasks were 20 s in duration (block design) with an inter-stimulus interval of 20-24 s. During the task period, a category (e.g., fruit, countries etc.) was dictated, and the subjects were asked to silently think of possible nouns from the selected category. During the rest period, the subjects were asked not to think about nouns. Each time series was filtered with a band pass filter to remove temporal dependencies due to baseline drifts and physiological noises, using 0.01 Hz and 0.8 Hz as the high and low pass cutoff, respectively. The HbO_2 signal values for the 5th to the 20th second of the task period and those from 5 seconds prior to the onset of the task period were averaged across 10 trials for each subject to generate task and baseline samples, respectively. We focused only on the HbO_2 signal.

We are interested in a statistical method that is robust to unequal numbers of observations in the task and baseline samples, though this was not the case here. Therefore, we randomly omitted some task observations, which resulted in twelve task observations for four of the channels, and fifteen for all others. The two-sample t-test at (2.1) can be used with such unpaired data, with different sample sizes and possibly different variance. We used the two-sample permutation and bootstrap t-test, and applied single-step and stepwise max t correction. For comparison, we also applied the conventional model with Bonferroni correction. As fNIRS does not provide structural brain information, we previously

developed a probabilistic data registration method for presenting fNIRS studies in MNI (Montreal Neurological Institute) standard brain co-ordinate system (Singh et al. (2005)). First we constructed our fNIRS template, NFRI-CB17 (National Food Research Institute Canonical Brain based on 17 brains), in MNI space by averaging the MR brain images from 17 subjects (mongoloid; 9 males, 8 females; aged 22 to 51 years) that are normalized in MNI space (Okamoto et al. (2004)). Using this template, our method obtains the most probable anatomical source of activation, and the associated positional errors in the standard MNI space. The detailed algorithms are discussed in (Okamoto et al. (2004), Jurcak, Okamoto, Singh and Dan (2005), Okamoto and Dan (2005) Singh et al. (2005), Jurcak, Tsuzuki and Dan (2007) and Tsuzuki et al. (2007)).

5. Simulation Analysis

In order to evaluate and compare the power as a function of the sample size (number of subjects) and examine the influence of the multiplicity level (number of channels) on the resampling-based FWER control methods, we synthesized datasets from the full fNIRS WG dataset. It is rare to find studies using more than 100 channels in fNIRS, though there are fNIRS devices that offer up to 200 channels. Therefore, we synthesized datasets for up to 204 channels by replicating the original WG dataset with 34 channels. Then, from this synthesized dataset, we randomly selected subsets of data from a range of TN = [7 9 11 13 15] subjects. We generated 30 randomly selected unique combinations of subjects for all subsets, except for the 15-subject subset in which we had only one possible combination. In each simulation, we systematically varied the number of channels, TC, in the range of [34 68 102 204] and, for each combination of TN and TC, we repeated the permutation or bootstrap test 1,000 times (i.e., $B = 1,000$). We averaged the results obtained from these simulations and derived the null distributions of the max t statistic for the chosen sample sizes, then computed max t and Bonferroni correction thresholds and the number of channels detected by these thresholds (Figures 4, 5).

6. Results

Table 1 shows p-values and t-value thresholds acquired from various combinations of statistical methods using the unbalanced WG data before and after FWER correction. These results are overlaid on the fNIRS reference brain for spatial visualization of the activated brain areas (Figure 2). As expected for the word generation task, activation was detected in the ventro-lateral prefrontal area with most of the activated channels in the left hemisphere (language dominance hemisphere). The observations can be summarized as follows. (1) The final outcome from the three resampling-based options for FWER control in terms of

detection of channels remains the same (Figure 2), though the p-values in the non-exchangeable bootstrap test tend to be relatively larger than those of the permutation and exchangeable bootstrap tests (Table 1). (2) The max t stepwise correction offers the most lenient threshold among the FWER controlling methods used in these examples, and the max t single-step procedure has produced more conservative estimates than even the Bonferroni correction in two cases (Table 1). For the results of the full dataset (as compared to the unbalanced dataset), refer to simulation analysis with 15 subjects 34 channels, using simulation parameters TN=15 and TC=34 (Figures 3, 4 and 5).

In the simulation analysis, the t-value thresholds in non-exchangeable bootstrap tests tend to be more conservative than the exchangeable bootstrap or permutation tests, particularly for small sample sizes (Figure 3) and larger numbers of channels (Figures 4 and 5). We have noted that, despite the visible loss in power with the decrease in sample size (Figure 5), both bootstrap and permutation tests can detect a few channels in the language area of the brain with a small sample having just six degree of freedom. Since, the parametric inference may be questionable at such low degrees of freedom, resampling-based inference indeed offers a good escape from this issue.

Table 1. Comparison of results from different combinations of statistical tests and FWER controls using the unbalanced data example. The first column indicates the rank of the channel in the descending order of original t-value given in second column. The subsequent columns show the p-values obtained from parametric, permutation, exchangeable bootstrap, and non-exchangeable bootstrap methods. The sub-columns, Unc, Bonf, Max1, and Max2 denote uncorrected p-value, and pvalues after Bonferroni, max t (single step), and max t (stepwise) correction, respectively. The corrected p-values in bold indicate that they are significant at 5% threshold.

		Parametric		Permutation				Exchangeable Bootstrap				Non-exchangeable Bootstrap			
Rank	T	Unc	Bonf	Unc	Bonf	Max1	Max2	Unc	Bonf	Max1	Max2	Unc	Bonf	Max1	Max2
22	8.189	0.0000	**0.0000**	0.0000	**0.0000**	**0.0000**	**0.0000**	0.0000	**0.0000**	**0.0000**	**0.0000**	0.0000	**0.0000**	**0.0000**	**0.0000**
19	7.412	0.0000	**0.0000**	0.0000	**0.0000**	**0.0000**	**0.0000**	0.0000	**0.0000**	**0.0000**	**0.0000**	0.0000	**0.0000**	**0.0000**	**0.0000**
23	7.008	0.0000	**0.0000**	0.0000	**0.0000**	**0.0000**	**0.0000**	0.0000	**0.0000**	**0.0000**	**0.0000**	0.0000	**0.0000**	**0.0000**	**0.0000**
6	3.704	0.0005	**0.0157**	0.0005	**0.0170**	**0.0112**	**0.0098**	0.0000	**0.0000**	**0.0088**	**0.0068**	0.0000	**0.0000**	**0.0130**	**0.0120**
21	3.623	0.0006	**0.0194**	0.0008	**0.0255**	**0.0140**	**0.0118**	0.0005	**0.0170**	**0.0108**	**0.0080**	0.0000	**0.0000**	**0.0160**	**0.0150**
25	3.614	0.0006	**0.0199**	0.0002	**0.0085**	**0.0142**	**0.0118**	0.0005	**0.0170**	**0.0108**	**0.0080**	0.0000	**0.0000**	**0.0160**	**0.0150**
26	3.192	0.0017	0.0591	0.0012	**0.0425**	**0.0408**	**0.0325**	0.0010	**0.0340**	**0.0305**	**0.0232**	0.0010	**0.0340**	**0.0350**	**0.0290**
14	3.107	0.0022	0.0732	0.0022	0.0765	**0.0500**	**0.0398**	0.0002	**0.0085**	**0.0370**	**0.0258**	0.0010	**0.0340**	**0.0400**	**0.0320**
7	3.099	0.0022	0.0746	0.0032	0.1105	0.0508	**0.0398**	0.0028	0.0935	**0.0372**	**0.0258**	0.0020	0.0680	**0.0410**	**0.0330**
32	2.755	0.0051	0.1735	0.0068	0.2295	0.1022	0.0752	0.0058	0.1955	0.0908	0.0612	0.0030	0.1020	0.0910	0.0680
2	2.564	0.0084	0.2845	0.0080	0.2720	0.1520	0.1120	0.0082	0.2805	0.1310	0.0925	0.0020	0.0680	0.1400	0.1070
3	2.547	0.0087	0.2957	0.0150	0.5100	0.1562	0.1120	0.0108	0.3655	0.1358	0.0925	0.0090	0.3060	0.1460	0.1090
20	2.545	0.0084	0.2844	0.0110	0.3740	0.1568	0.1120	0.0080	0.2720	0.1370	0.0925	0.0050	0.1700	0.1470	0.1090
29	2.541	0.0084	0.2870	0.0075	0.2550	0.1575	0.1120	0.0042	0.1445	0.1378	0.0925	0.0030	0.1020	0.1480	0.1090
17	2.067	0.0241	0.8184	0.0292	0.9945	0.3572	0.2508	0.0232	0.7905	0.3180	0.2155	0.0180	0.6120	0.3140	0.2340

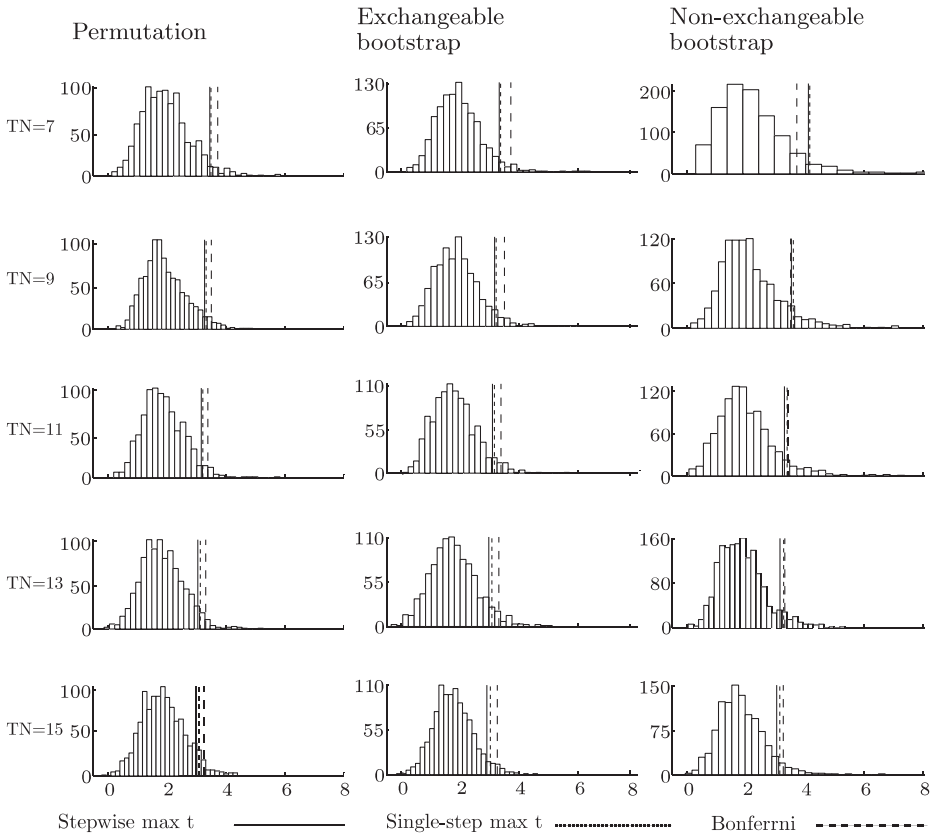

Figure 3. Histograms for resampling-based max t distribution. The simulation analysis is depicted in the histograms of max t distribution obtained from synthetic data with different number of subjects, TN=[7 9 11 13 15], and TC=34 channels. The stepwise max t, single-step max t, and Bonferroni thresholds at 5% level are shown as black, dotted and dashed lines, respectively. The stepwise max t threshold corresponding to the least t-value among the channels that survived the FWER correction at 5% level was used as the common threshold for providing an objective comparison of all the thresholds. All these tests used 1,000 resamples.

7. Discussion

Our results show that both permutation and bootstrap methods in two-sample testing, along with max t correction, can increase the significance of the inference from fNIRS group analysis. Unlike permutation tests, bootstrap tests are not exact and therefore are not guaranteed to preserve the Type I error. The guidelines to optimize the power and specificity of the bootstrap test, highlighted by and Hall and Wilson (1991), recommend that resampling should be done in a manner so that it reflects the null hypothesis, and that a pivotal statistic should

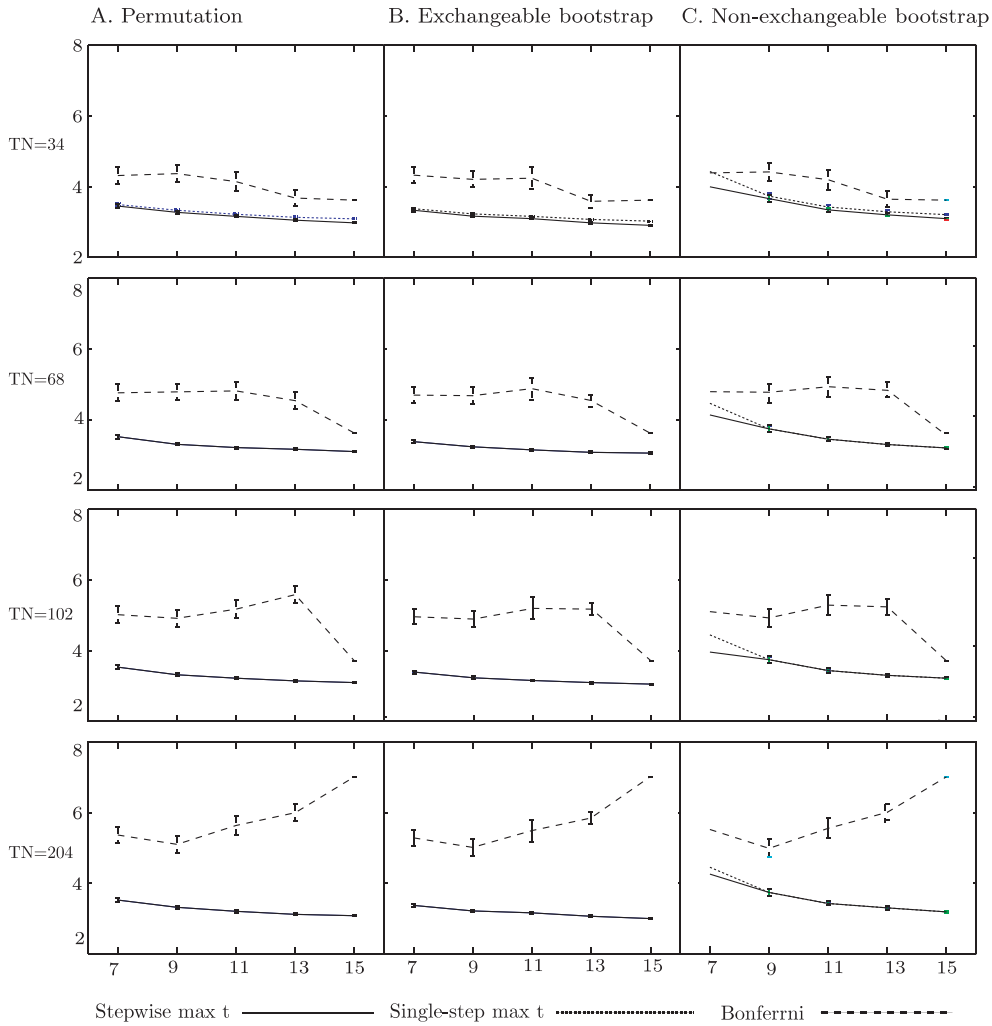

Figure 4. Comparison of FWER thresholds. The graphs show the t-value thresholds (y-axis) obtained from simulation analysis after systematically varying the multiplicity level (TC) from 34 to 204, and the number of subjects (TN) from 7 to 15 (x-axis). The stepwise max t threshold corresponding to the least t-value among the channels that survived the FWER correction at 5% level was used as the common threshold for providing an objective comparison of all the thresholds. All these tests used 1,000 resamples over 30 simulations. The error bars indicate the simulation error in the thresholds.

be used. The null hypothesis of equality of distributions in exchangeable bootstrap test is reflected by pooling the observations before resampling, and the null hypothesis of equality of means in non-exchangeable bootstrap test is reflected by shifting the means before resampling. A statistic is said to be pivotal if its

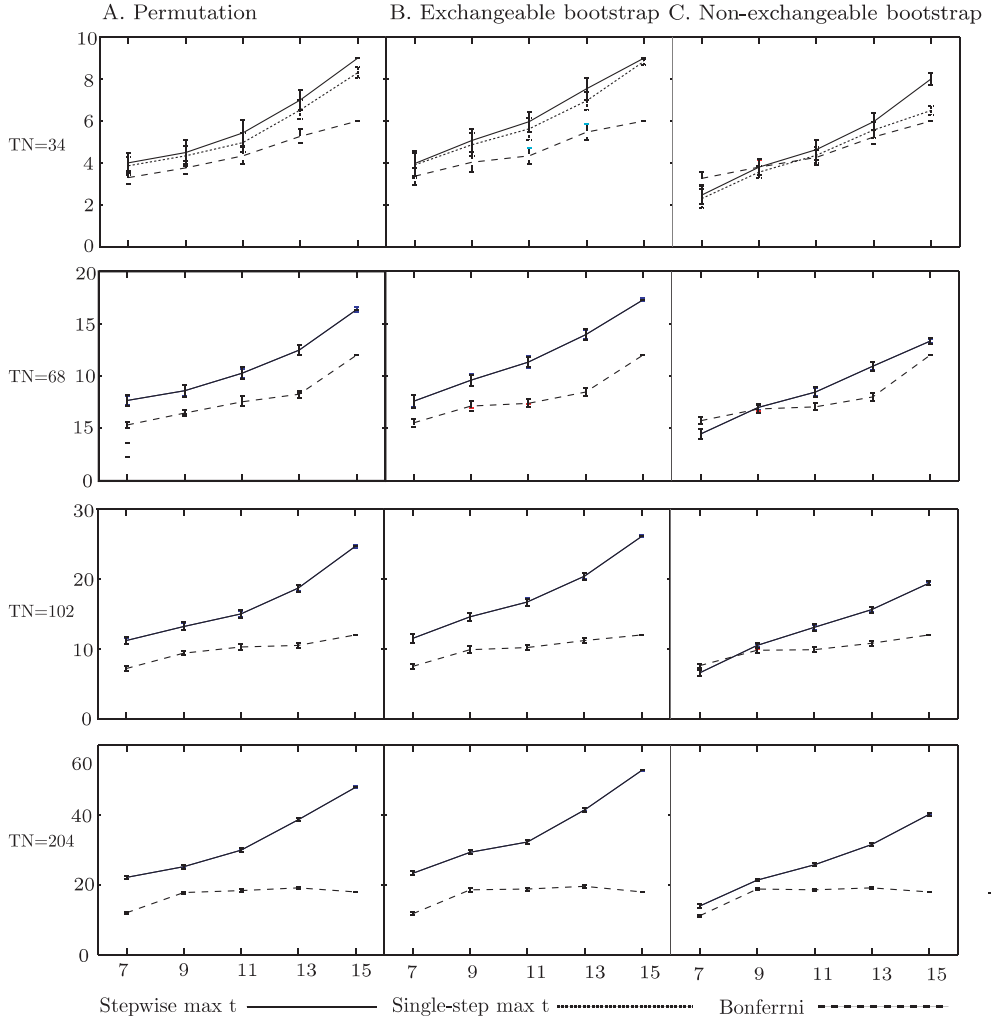

Figure 5. Power and sample size. The graph shows the influence of multiplicity level (TC) and sample size (x-axis) on the power in terms of the number of the detected active channels along (y-axis) for each resampling-based test option applied in simulation analysis. The error bars indicate the simulation error in the number of channels.

sampling distribution is independent of the the distribution of the data that generated it. The t-statistic used in the bootstrap tests (3.1), is pivotal for normal data, and asymptotically pivotal for random and non-normal data (Westfall and Young (1993, Sec. 2.2.2)). Therefore, in order to preserve the pivotality of the statistic asymptotically, the bootstrap test should be avoided with very small samples. Interestingly, a previous study on theoretical comparisons of the exact

permutation and bootstrap tests reported that both preserved the Type I error, but the bootstrap tests were more conservative in the specific single test examples that were used (Corcoran and Mehta (2002)). Subsequently these guidelines were extended to the multiple testing cases under the assumption of subset pivotality (Westfall and Young (1993)), which is granted by max t correction, and is assumed in most neuroimaging experiments (Holmes et al. (1996)). In the WG example with unbalanced data, the power of permutation and bootstrap tests was equivalent in terms of detection of active channels, when the max t stepwise correction was used (Table 1 and Figure 2).

The complete enumeration of a resampling-based test including all possible resamples is computationally difficult in large samples. Instead, an approximate resampling-based p-value can be enumerated from a random subset of all possible resamples at the cost of a small bias (Dwass (1957)) that depends on the number of resamples included in the enumeration, B. There is no single rule of thumb for selecting the value of B, but a histogram can give an indication whether the null distribution is well approximated at the chosen B or not (e.g., Figure 1) (Good (2000)). In a multiple testing situation, it is complicated to express the deviation in the p-values of all the channels. Therefore, instead of the deviation in p-values, we estimated the deviations in single-step (max1) and step-down (max2) thresholds in the simulation study, at $B = 1,000$, over 30 repetitions (Figure 4).

The issue of selecting an ideal Type I error control procedure is common to most neuroimaging techniques, where Bonferroni correction tends to be too conservative because of positive spatial correlation. Several alternatives have been suggested to overcome this in functional neuroimaging, e.g., random field theory (RFT) (Worsley et al. (1992)). and false discovery rate (Benjamini and Hochberg (1995), Genovese, Lazar and Nichols (2002) and Singh and Dan (2006)). However RFT requires good lattice assumptions that are problematic with fNIRS data, and FDR requires a certain kind of positive spatial correlation structure (Benjamini and Yekutieli (2001)). Fortunately, resampling based max t correction methods do not require such assumptions.

Some authors have explored the appropriateness of several permutation tests for fMRI/PET data (Nichols and Holmes (2002)). These tests should be applicable to any data, including fNIRS data, but the scope and applicability might vary. For example, due to the massive volumes of fMRI data and their processing requirements (e.g., the smoothed variance t-test), stepwise correction is not recommended because its power gain over single-step correction has not been found worth the additional computational burden imposed by its iterative nature (Holmes et al. (1996)). This is not the case for fNIRS. Sometimes the distribution of the permutation statistic will not be identical among channels, which is

more likely in case of the unbalanced data with different number of observations between channels. The stepwise max t threshold is successively adjusted among channels, and therefore provides greater sensitivity, particularly at channels not having the maximal distribution. As fNIRS data is less bulky and does not require complicated processing, stepwise correction is easily applicable, and we have found it to be more powerful than Bonferroni correction in our examples.

In functional neuroimaging, the application of bootstrap tests is rare and has been limited to time series analysis (Bullmore, Breakspear and Suckling (2003)); their application in FWER control, e.g., max t, remains unexplored. Our results indicate that they may be useful for fNIRS when the experimenter may not want to assume equal variance under the null hypothesis, due to possible failure of homogeneity, e.g., in the case of the unbalanced (missing) data example (Figure 3). The multiplicity level in fNIRS is moderate. Our simulation example shows that, although a non-exchangeable bootstrap test becomes conservative compared to the other resampling-based tests with increasing multiplicity, it is still able to detect some active channels in the language area. A similar result was observed in microarray analysis, where non-exchangeable bootstrap tests were found to be very conservative compared to permutation tests in their specific examples (Troendle et al. (2004)). We hope that the bootstrap test will also be examined in the context of other techniques with large multiplicity issues, e.g., fMRI.

The results described in this paper are implemented in matlab, using in-house software that is developed for analyzing fNIRS data. It may be downloaded from our website http://brain.job.affrc.go.jp/.

Acknowledgement

We would like to thank Prof. Ajit Tamhane (Northwestern University, Evanston) and Dr. Thomas Nichols (GlaxoSmithKline Clinical Imaging Centre) for their valuable comments and suggestions. This work was supported by the Grant-in-Aid program from the Japan Society for Promotion of Scinece (JSPS) (19650079) and the Program for Promotion of Basic Research Activities for Innovative Biosciences (PROBRAIN).

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165-1188.

Bullmore, E., Breakspear, M., and Suckling, J. (2003). Wavelets and statistical analysis of functional magnetic resonance images of the human brain. *Statist. Meth. Medical Res.* **12**, 375-399.

Corcoran, C. D. and Mehta, C. R. (2002). Exact level and power of permutation, bootstrap, and asymptotic tests of trend. *J. Modern Applied Statistical Methods* **1**, 42-51.

Cope, M., Delpy, D. T., Reynolds, E. O., Wray, S., Wyatt, J. and van der Zee, P. (1988). Methods of quantitating cerebral near infrared spectroscopy data. *Adv. Exp. Med. Biol.* **222**, 183-189.

Dudoit, S., Yang Y. H., Callow, M. J. and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica* **12**, 111-139

Dwass, M. (1957). Modified randomization tests for non-parametric hypotheses. *Ann. Math. Statist.* **28**, 181-187.

Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *Ann. Statist.* **7**, 1-26.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* **1**, 54-75.

Fisher, R. A. (1935). *Design of Experiments*. Oliver and Boyd, Edinburgh.

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

Genovese, C. R., Lazar, N. A. and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* **15**, 870-878.

Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. 2nd edition. Springer.

Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757-762.

Holmes, A. P., Blair, R. C., Watson, J. D. and Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. *J. Cereb. Blood Flow Metab.* **16**, 7-22.

Hoshi, Y. (2003). Functional near-infrared optical imaging: utility and limitations in human brain mapping. *Psychophysiology* **40**, 511-520.

Jurcak, V., Okamoto, M., Singh, A. and Dan, I. (2005). Virtual 10-20 measurement on MR images for inter-modal linking of transcranial and tomographic neuroimaging methods. *Neuroimage* **26**, 1184-1192.

Jurcak, V., Tsuzuki, D. and Dan, I. (2007). 10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems. *Neuroimage* **34**, 1600-1611.

Koizumi, H., Yamamoto, T., Maki, A., Yamashita, Y., Sato, H., Kawaguchi, H. and Ichikawa, N. (2003). Optical topography: practical problems and new applications. *Appl. Opt.* **42**, 3054-3062.

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

Obrig, H. and Villringer, A. (2003). Beyond the visible–imaging the human brain with light. *J. Cereb. Blood Flow Metab.* **23**, 1-18.

Okada, E. and Delpy, D. T. (2003). Near-infrared light propagation in an adult head model. II. Effect of superficial tissue thickness on the sensitivity of the near-infrared spectroscopy signal. *Appl. Opt.* **42**, 2915-2922.

Okamoto, M., Dan, H., Sakamoto, K., Takeo, K., Shimizu, K., Kohno, S., Oda, I., Isobe, S., Suzuki, T., Kohyama, K. and Dan, I. (2004). Three-dimensional probabilistic anatomical cranio-cerebral correlation via the international 10-20 system oriented for transcranial functional brain mapping. *Neuroimage* **21**, 99-111.

Okamoto, M. and Dan, I. (2005). Automated cortical projection of head-surface locations for transcranial functional brain mapping. *Neuroimage* **26**, 18-28.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any population. *Roy. Staist. Soc. Supplement* **4**, 119-130.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.* **17**, 141-159.

Schroeter, M. L., Bucheler, M. M., Muller, K., Uludag, K., Obrig, H., Lohmann, G., Tittgemeyer, M., Villringer, A. and von Cramon, D. Y. (2004). Towards a standard analysis for functional near-infrared imaging. *Neuroimage* **21**, 283-290.

Singh, A. K., Okamoto, M., Dan, H., Jurcak, V. and Dan, I. (2005). Spatial registration of multichannel multi-subject fNIRS data to MNI space without MRI. *Neuroimage* **27**, 842-851.

Singh, A. K. and Dan, I. (2006). Exploring the false discovery rate in multichannel NIRS. *Neuroimage***33**, 542-549.

Strangman, G., Boas, D. A., and Sutton, J. P. (2002). Non-invasive neuroimaging using near-infrared light. *Biol. Psychiatry* **52**, 679-693.

Steinbrink, J., Villringer, A., Kempf, F., Haux, D., Boden, S. and Obrig H. (2006), Illuminating the BOLD signal: combined fMRI-fNIRS studies *Magn Reson Imaging*, **24**, 495-505

Troendle, J. F., Korn, E. L. and McShane, L. M. (2004). An example of slow convergence of the bootstrap in high dimensions. *The American Statistician* **58**, 25-29.

Tsuzuki, D., Jurcak, V., Singh, A. K., Okamoto, M., Watanabe, E. and Dan, I. (2007). Virtual spatial registration of stand-alone fNIRS data to MNI space. *Neuroimage* **34**, 1506-1518.

Villringer, A. and Dirnagl, U. (1995). Coupling of brain activity and cerebral blood flow: basis of functional neuroimaging. *Cerebrovasc. Brain Metab. Rev.* **7**, 240-276.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing : Examples and Methods for p-Value Adjustment*. Wiley-Interscience.

Worsley, K. J., Evans, A. C., Marrett, S. and Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *J. Cereb. Blood Flow Metab.* **12**, 900-918.

Sensory and Cognitive Food Science Lab., National Food Research Institute, 2-1-12, Kannondai, Tsukuba, 305-8642, Japan.

Department of Computer Science, Graduate School of Systems and Information Engineering, University of Tsukuba, Japan.

E-mail: archana@affrc.go.jp, sine.arc@gmail.com

Sensory and Cognitive Food Science Lab., National Food Research Institute, 2-1-12, Kannondai, Tsukuba, 305-8642, Japan.

E-mail: les.clowney@affrc.go.jp

Sensory and Cognitive Food Science Lab., National Food Research Institute, 2-1-12, Kannondai, Tsukuba, 305-8642, Japan.

E-mail: masakoo@affrc.go.jp

Department of Computer Science, Graduate School of Systems and Information Engineering, University of Tsukuba, Japan.

E-mail: cole@is.tsukuba.ac.jp

Sensory and Cognitive Food Science Lab., National Food Research Institute, 2-1-12, Kannondai, Tsukuba, 305-8642, Japan.

E-mail: dan@affrc.go.jp

(Received April 2007; accepted April 2008)