A combined bootstrap / histogram analysis approach for computing a lateralization index from neuroimaging data

Marko Wilke [1,2] and Vincent J. Schmithorst [3, 4]

[1] *Department of Pediatric Neurology and Developmental Medicine, Children's Hospital, and*

[2] *Section for Experimental MR of the CNS, Dept. of Neuroradiology,*

University of Tübingen, Germany

[3] *Department of Pediatrics, University of Cincinnati, and* [4] *Imaging Research Center, Cincinnati*

Children's Hospital Medical Center, Cincinnati, OH

Running title: Bootstrapped lateralization index

Corresponding author:

Marko Wilke, MD
Department of Pediatric Neurology and Developmental Medicine
Children's Hospital, University of Tübingen
Hoppe-Seyler-Str. 1
72076 Tübingen, Germany
Phone: + 49 7071 – 29 83416
Fax: + 49 7071 – 29 5473
e-mail: Marko.Wilke@med.uni-tuebingen.de

Abstract

Cerebral hemispheric specialization has traditionally been described using a lateralization index (LI). Such an index, however, shows a very severe threshold dependency and is prone to be influenced by statistical outliers. Reliability of this index thus has been inherently weak, and the assessment of this reliability is as yet not possible as methods to detect such outliers are not available. Here, we propose a new approach to calculating a lateralization index on functional magnetic resonance imaging data, by combining a bootstrap procedure with a histogram analysis approach. Synthetic and real functional magnetic resonance imaging data was used to assess performance of our approach. Using a bootstrap algorithm, 10.000 indices are iteratively calculated at different thresholds, yielding a robust mean, maximum and minimum LI and thus allowing to attach a confidence interval to a given index. Taking thresholds into account, an overall weighted bootstrapped lateralization index is calculated. Additional histogram analyses of these bootstrapped values allow to judge reliability and the influence of outliers within the data. We conclude that the proposed methods yield a robust and specific lateralization index, sensitively detect outliers and allow to assess the underlying data quality.

Introduction

Hemispheric specialization of the brain has been the focus of a large number of studies, mainly using imaging methods like positron emission tomography (PET) or functional magnetic resonance imaging (fMRI; for review, see Cabeza & Nyberg, 2000, and Hugdahl & Davison, 2002). In this note, we propose to apply the bootstrap concept to calculating lateralization indices from neuroimaging data.

Background: bootstraps and lateralization indices

The term “bootstrap” reportedly comes from the legendary German figure Baron of Münchhausen, who dragged himself out of a swamp by pulling on his bootstraps. In statistics, the term describes a technique that tries to find the sampling distribution of a sample, by repeatedly re-sampling, with replacement, the original sample. In other words, several resamples are generated from a given original sample in order to estimate a bootstrap distribution that allows approximating the “real” distribution of the original sample. It is important to note that a bootstrap does not add or replace data from the original distribution, but only uses multiple resamples of the original (Davison & Hinkley, 1997; Hesterberg *et al.*, 2005; Janssen & Pauls, 2003, Moore *et al.*, 2002). This is illustrated in equation 1

$$\begin{array}{cccccccccccc} 1 & & 2 & & 5 & & 4 & & 5 & & 1 & \\ 2 & & 5 & & 4 & & 1 & & 5 & & 5 & \\ 3 & \Rightarrow & 5 & \& & 3 & \& & 3 & \& & 3 & \& 1 & \& \ldots \\ 4 & & 1 & & 3 & & 5 & & 4 & & 4 & \\ 5 & & 3 & & 3 & & 5 & & 3 & & 4 & \\ & & & & & & & & & & & \\ i & & r_1 & & r_2 & & r_3 & & r_4 & & r_5 & \ldots\ r_n \end{array} \qquad \text{(Equation 1)}$$

It is obvious that both computationally and statistically, the procedure is mainly influenced by the size of the resample (size $_r$) and the number or resamples (n). This size $_r$ can be variable: while it is in most cases identical to size $_i$, it can be less than that (the special case of it being size $_{i-1}$ is called a jackknife procedure; Davison & Hinckley, 1997). The number of resamples is mainly limited by the computational demand when dealing with larger sample sizes; typically, several hundred to thousand resamples are used (Davison & Hinkley, 1997; Hesterberg *et al.*, 2005). Bootstrap approaches have been used before in neuroimaging, mainly for fMRI (Auffermann, Ngan & Hu, 2002; Prohovnik *et al.*, 2004) and diffusion tensor imaging (Jones & Pierpaoli, 2005; Lazar & Alexander, 2005).

Hemispheric specialization is a common question in functional and structural neuroimaging as well as in the cognitive neurosciences in general (for review, see Hugdahl & Davison, 2002). Approaches to describe this asymmetry mostly aim at presenting a single number in order to allow for the comparison of results. Akin to classical approaches to describing handedness (Oldfield, 1971), lateralization has traditionally been described using a lateralization index LI, computed as

$$LI = \frac{\sum activation_{left} - \sum activation_{right}}{\sum activation_{left} + \sum activation_{right}} \qquad \text{(Equation 2)}$$

Resulting from such an equation is a value between 1 (complete left-lateralization) and -1 (complete right lateralization). However, issues concerning such an index include vulnerability to outliers, a strong threshold dependency and the lack of immediate inference on data quality (Adcock *et al.*, 2003; Deblaere *et al.*, 2004; Gaillard *et al.*, 2002; Holland *et al.*, 2001).

Approach

In this manuscript, we describe the application of the bootstrapping concept to the calculation of lateralization indices in (functional) neuroimaging data. The main aims are (1) to allow for the assessment of data quality by detecting (statistical or artifactual) outliers; (2) to increase the stability of a given index by broadening the base of underlying information and by restricting outlier influence; (3) software implementation should allow easy usage within a publicly available imaging analysis tool, and (4) to remove the necessity of defining a cutoff threshold for interpreting lateralization in functional neuroimaging data.

Regarding No. 1, factors possibly making the algorithm more sensitive to outliers shall be systematically explored. Moreover, parameters allowing for the detection of outliers should routinely be derived from the analysis of the data and presented as part of the results.

In order to increase stability as defined in No. 2, the bootstrap algorithm is ideally suited to broaden the available data basis by providing multiple resamples. Our approach for robustly excluding outliers from this data is described below.

Considering No. 3, available software suites and the underlying solutions, we implemented our algorithms in MATLAB (The Mathworks, Natick, MA) and designed a graphical user-interface based on routines available within the spm-software environment (SPM2, Wellcome Department of Imaging Neuroscience, University College London, UK).

As to No. 4, we decided to adopt the concept of threshold-dependent laterality curves, iteratively exploring increasing thresholds (Deblaere *et al.*, 2004), as the basis for further calculations. The image under investigation (for example, a t-map; Figure 1) is

thresholded at regularly-spaced intervals. The original implementation of the laterality curves submitted the surviving voxels on the left and the right to equation 1 to yield a lateralization index for each threshold. The resulting diagrams show the obtained lateralization index (in y-direction) versus the threshold (in x-direction).

For our approach here, we used the surviving voxels as input samples for a bootstrap procedure in such a way that from the original single input sample, a multitude of bootstrapped re-samples was generated. From these resulting *n* samples from each side, all possible lateralization indices were calculated. This procedure is repeated during each thresholding step. See also Figure 1 for an overview of the steps.

Issues: speed, specificity, accuracy & outlier detection

Speed: As fMRi deals with large datasets, the computational steps easily become very time consuming. Limiting steps within the framework of the bootstrap are not only the size of the input sample ($size_i$), but also the number of resamples (*n*) and the size of the resample ($size_r$). We believe that limiting both the upper and the lower size of this resample makes sense in this setting, although for different reasons. First, when analyzing large input samples (typically occurring at lower thresholds), the stability of the resulting values should be very high, owing to the large number of contributing voxels. We therefore postulated that it is justified to specify an upper size limit *max* for the resulting bootstrap sample in order to speed up processing at low thresholds. Samples smaller than this will be sampled completely (if complete sampling is chosen, see below). Secondly, very small samples pose additional dangers: a single remaining voxel on one side will lead to a lateralization index of +/-1, which is not a plausible scenario, biologically, statistically, or computationally. We therefore suggest specifying a lower boundary, i.e. the algorithm aborts if a minimum number *min* of

voxels is not found. As a further safeguard against scattered single voxels, a warning is issued if a certain minimum cluster size is not met (default is ET = 5 voxels).

Sensitivity vs. Specificity: In order to punish outliers and restrain their influence on the ensuing results, we used a "trimmed mean" value when analyzing the resulting LI-values from each threshold (along the y-direction of a laterality curve diagram, i.e. for each single iteration). A trimmed mean$_{25}$ only uses the mean 50% of data while disregarding the upper and lower 25% of datapoints (Hesterberg *et al.*, 2005). In a sample skewed by outliers, such a trimmed mean is a more representative measure of the "true" center of the distribution. On the other hand, as the range of resulting value is an important indicator of data homogeneity, we opted to retain this information by plotting the minimum and maximum LI-values from each step. These values reflect the range of results obtained at this threshold, which will be small (in the case of homogenous data) or large (in the case of inhomogeneous data).

Regarding the specificity of obtained results, (functional) neuroimaging data is usually thresholded to ensure the significance of obtained results. Classically, a lateralization index is computed from these "significant" voxels only, which are obtained after additionally accounting for multiple comparisons. However, a number of different correction methods exist which will, while statistically equally legitimate, yield different thresholds (Marchini & Presanis, 2004). Among them are approaches favoring specificity (family-wise error correction, FWE) or sensitivity (false discovery rate, FDR; Nichols & Hayasaka, 2003); the definition of voxels to exclude is therefore not straightforward. Previous attempts did not threshold the statistical images at all prior to lateralization analyses (Holland *et al.*, 2001). Consequently, however, many voxels showing no correlation with the task will also be included, which invariably introduces noise in the calculations. We therefore opted to use an

approach that allows to account for the “meaningfulness” of values obtained from different thresholds by attributing different weights to them. While no attempt was made to adopt any given scheme for determining significance, it is immediately obvious that a voxel showing a higher correlation with the task should have a greater impact on the ensuing results. We therefore decided to employ a *weighted mean* (along the x-direction of a laterality curve diagram, i.e. over the results from all thresholds). A weighted mean computes a mean of a given sample ($x_1 - x_n$), but takes into account a weighting factor w_i for each datapoint x:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i * x_i}{\sum_{i=1}^{n} w_i} \qquad \text{(Equation 3)}$$

In our case, the obvious choice for such a weighting factor is the threshold at which the image was thresholded in order to generate the value of x, which will result in a progressively stronger weighting of the lateralization index values obtained at higher thresholds. Note that, with a constant weighting factor w, a weighted mean is equivalent to the standard arithmetic mean. Therefore, a weighted mean of the results from all thresholding results (i.e., along the x-direction of a laterality curve diagram) can be based on the trimmed means obtained at each threshold, thus combining stability and specificity. Accordingly, all trimmed mean values (together with the respective thresholds as weighting factors) are submitted to equation 3 to yield an overall weighted mean.

Accuracy & outlier detection: As mentioned above, the size $_r$ of the resample usually is equivalent to the size $_i$ of the input sample (see equation 1). In this scenario, each bootstrapped resample r is a complete resample (with replacement) of i, and with large sample sizes, only a limited number of datapoints will not be sampled. However,

(Equation 4)

when decreasing $size_r$ in relation to $size_i$, the number of datapoints not sampled will increase accordingly, as shown in equation 4:

$$size_r = k * size_i$$

Obviously, modifying the resample ratio *k* within a range of 0-1 will result in a resample *r* that is smaller than the input sample *i*. Such a scenario is illustrated in Figure 2 (left panel): a sample of 100 points (Y-axis, left diagram) is randomly sampled 100 times (X-axis, left diagram) with a frequency of $k = .25$. This results in 100 samples of 25 datapoints each, and the resulting random frequency of each original datapoint in the resamples is illustrated in the histogram (Figure 2, right panel): on *average*, each point occurs 25 times in the resamples. While the bootstrapped sample is still based on 2500 datapoints as opposed to the original 100, the resulting samples will, on the whole, be much more likely to detect data inhomogeneity than a sample based on $k = 1$. This is simply because potential outliers will not be present in each sample, thus widening the spread of resulting averages. Considering the example of datapoint 50 (circled, Figure 2, left panel) being an outlier, it will only be present in 20/100 resamples (right panel). In order to increase an algorithms' sensitivity to outliers, the value of *k* can thus be systematically decreased. This might result in a decrease of specificity as the correspondence between each individual resample *r* and the input sample *i* will naturally also decrease. On the other hand, this is outweighed by a large number *n* of resamples. We therefore hypothesized that varying *k* should have discernible effects on the ability to detect outliers while, due to the large number of resamples and the outlier protection implemented above, accuracy should not be severely affected.

Methods

In order to compare the results from our algorithm, we randomly picked imaging data from a previous study on language lateralization (Wilke *et al.*, 2006; see also for details on data acquisition and processing). Data from 5 healthy subjects (3 boys, 2 girls, mean age 12.8 ± 2 years, right-handed with an average Oldfield-score of .77, range .64-1 [Oldfield, 1971]), performing a robust left-lateralized language task (Fernandez *et al.*, 2001) were selected. The resulting t-maps from each subject were analyzed with regard to lateralization within the frontal lobe, defined using masks based on anatomical definitions (Tzourio-Mazoyer *et al.*, 2002).

The bootstrapping algorithm makes use of the random number generation capabilities of the Matlab software environment. For example, consider a 3-dimensional image volume which, after thresholding and masking, consists of 500 non-zero voxels; it is then converted to a 1 x 500-vector and analyzed with a re-sample ratio of $k = .25$. Consequently, 125 random integer numbers are generated in the range from 1-500 which are then used to randomly, with replacement, sample the input data. This procedure is repeated 100 times, resulting in 100 bootstrapped samples from each side.

User interaction is completely GUI-driven, using a toolbox plug-in for spm2. Results from our approach should be compared to the classical approach to only assess voxels surviving a given threshold of significance, either defined using the family-wise correction for multiple comparisons or the false-discovery rate as implemented in spm2. To this effect, fMRI-data from the earlier study (Wilke *et al.*, 2006) is analyzed using all approaches, and results are assessed qualitatively.

To illustrate the influence of outliers on lateralization index calculations, the value of one voxel in a t-map (subject #1 in table 1) was iteratively varied to be between 1 and 50 times the maximum value in the individual image (outlier weight *ow* = 1-50). This serves to investigate robustness of the ensuing lateralization indices. To further assess stability of results from the bootstrap algorithm, 100 re-runs of the same (unaltered) dataset were analyzed, with different resample ratio settings ($k = 1$, $k = 0.5$ and $k = 0.25$).

Presets

We used a default of 20 thresholding intervals (equally-sized steps from 0 to the maximum value in the masked image) and to generate 100 bootstrap samples for each side (coded to be five times the number of thresholding steps to allow for a combined more exhaustive assessment). Owing to the open nature of MATLAB-script files, both values can easily be adjusted within the code. From these 100 samples from each side, a total of 10.000 lateralization indices for each thresholding step results, and an overall maximum of 200.000 indices (with default values: doubling the number of iterations will lead to an overall matrix of 1.6 million LIs). When default settings are used, results are stored in a matrix of 20 (thresholds, in the X-dimension) by 10.000 (lateralization indices, in the Y-dimension.)

By default, the upper sample size limit was set to $max = 1000$ voxels, i.e., from samples larger than this, only 1000 voxels will randomly be drawn. As discussed above, a lower boundary should also be set in order to avoid a lateralization index based on a very small number of voxels: we suggest using $min = 5$ voxels as the minimum bootstrap resample size. Of note, when using a smaller resample size in relation to the input $size_i$ ($k < 1$), in order to keep the bootstrapped sample $size_r$

constant, *k* needs to be taken into account by extending the minimum number of voxels to *min* = *min*/*k*, such that, with a sampling size of 50% (*k* = .5) the minimum input sample size would be 10 voxels.

The histogram of the overall lateralization index-matrix is plotted to allow assessing the resulting distribution at each threshold, giving an impression of how normally distributed each underlying sample is. Additionally, the matrix of lateralization indices from each side was converted to normalized z-scores according to

$$z_{LI} = \frac{LI - Mean_{LI}}{SD_{LI}} \qquad \text{(Equation 5)}$$

These were plotted as a function of iteration in order to allow for the visual assessment of data homogeneity over all thresholding steps. A histogram of these z-scores was also generated to further assess the (normal or skewed) distribution of the z-scores.

Results

For the fMRI-data of the 5 healthy subjects, our bootstrapped results and lateralization indices from significant voxels only are shown in Table 1. It is apparent that no results are obtained from some “classical” calculations if a minimum number of voxels is required (as in our algorithm; if this would be tolerated, a value of 1 resulted). Our lateralization indices are concordant with the results from the other approaches, with the weighted mean indicating stronger lateralization than the overall and trimmed mean. The excellent agreement for unaltered fMRI data is also evident in the stepwise comparison of results from a single dataset (subject #1), comparing results from a classical lateralization curve with our bootstrap approaches (Table 2).

Overall lateralization is clearly influenced by even a single outlier (Figure 3): from an outlier weight $ow \geq 7$, all negative values result in the classical lateralization curves, and with $ow > 8$, the result is almost uniformly -1 (average of weighted means: -.99). This effect becomes more pronounced with increasing outlier weights and increasing thresholds (as voxel numbers decline). An abated effect is seen in the bootstrapping approach with complete resampling ($k = 1$): an average weighted mean lateralization index of -.3 results if $ow > 8$. The outlier influence strongly decreases when using a smaller resample size ($k = .5$ and $k = .25$). With $k = .5$ and $ow > 8$, an average weighted lateralization index of .4 is returned. In the last case, the correct left lateralization is retained in all 50 runs (average weighted lateralization index for $ow > 8$: .68). Note absence of minimum/maximum bounds in these graphs for better accessibility.

When repeating our bootstrapping procedure 100 times on actual fMRI-data (Figure 4, upper panels), the stability of the resulting mean lateralization indices is

extremely high. In all cases, the sampling ratio *k* has no discernible influence on either the mean or the detected range if high-quality fMRI-data is examined (Figure 4, lower panels).

The actual output of the algorithm for a single dataset (with 2 outliers, *ow* = 10, on the right, *k* = 1) is shown in Figure 5 & 6: note much wider range of detected lateralization indices on the "artifacted" (right) side (Figure 5, top panel). The histogram of all lateralization indices shows several smaller, irregular peaks from outlier-influenced samples (Figure 5, lower panel). The z-score histogram for the left side (Figure 6, upper panels) shows an evenly distributed pattern over the whole range of thresholds, and the z-scores are nicely normally distributed. For the right side (Figure 6, lower panels), containing the outliers, a much less homogenous z-score sample results, with extreme values (in "colder" colors) present over all thresholds and a consecutively skewed z-score histogram.

Discussion

In this work, we applied the bootstrapping concept to the calculation of lateralization indices. Special emphasis was put on robust calculations and on sensitive outlier detection in order to both avoid and detect possible outlier influences.

When comparing our results with the “classical” approaches, Table 1 shows no results for the FWE-corrected approaches in 3/5 cases as no voxels survive thresholding on the right. While this will result in a very “clear-cut” lateralization index of LI = 1, it seems (mathematically and biologically) dangerous to accept no voxels on one side of the equation. Due to our minimum size criterion, our algorithm will abort here, as suggested before (Deblaere *et al.*, 2004). Our weighted mean results, designed to combine robustness and specificity, are in every case concordant with results from classical approaches. The false discovery rate only aims at controlling the rate of false positives, it is therefore more sensitive than the stricter family-wise error correction (Nichols & Hayasaka, 2003). With our values consistently indicating stronger lateralization than the values obtained from the FDR-corrected thresholds, a more specific analysis could be postulated, without reaching the rigidity of the FWE-correction. Of note, the user-defined optional input of a lower threshold (e.g., an FDR-corrected cutoff-value) will also allow to individually explore “significant voxels” only with our algorithm (see below).

The calculation of a lateralization index even in the form of a lateralization curve is strongly susceptible to outliers (Figure 3). If only a single voxel has a value of more than 7 times the maximum image value, the expression of an opposite laterality

artifactually results. While our simulation of outliers may not be a very realistic scenario, it was designed to demonstrate the algorithm's outlier detection capabilities. Only this sensitive outlier detection enables an informed choice with regard to whether a dataset is usable or not. In the case of our bootstrap approaches, a sampling ratio of $k = 1$ already shows an increased stability; lower values of k (.5 and .25, respectively), further improve stability against these outliers, with the smallest sampling ratio yielding the best results: in all cases, correct strong left lateralization is retained despite the influence of the outlier. Three additional points seem worth mentioning: one, in every case the widening spread of minimum/maximum LI-values and the consecutively skewed z-score histograms would have alerted the user as to the presence of strong data inhomogeneity even when a "nice-looking" mean curve results. This is a decisive advantage over the classical lateralization curve approach. The effect is not shown in Figure 3, but is illustrated in Figures 5 & 6. Two, it is interesting to notice that the outlier effect in the bootstrapped samples consistently only becomes apparent after the second or third iteration, while it is present from the first iteration on in the classical lateralization curves. As it is independent from the sampling ratio, it must be a consequence of the maximum sample size restriction: only then (approaching the upper bootstrap size limit of $max = 1000$ voxels) is the input sample sampled completely (in the case of $k = 1$) and outliers start to influence results. This further argues in favor of using smaller resamples that are, individually, less likely to contain a small number of outliers in many samples. Third, the amazing stability of the last approach (with $k = .25$) is likely to be enhanced by the "trimmed mean$_{(25)}$" we employed: if an outlier is only present in 25% of samples and the upper and lower 25% of the data are not used, the combined robustness against outliers must be expected to be very high.

However, as a decrease in *k* will require a greater input sample size $_I$ if *k* is constant, there is a tradeoff on how low *k* can be without having to prematurely abort iterations (note 11 iterations in the classical lateralization curve and only 10 iterations with $k = .25$, in Figure 3). Therefore, we suggest to use $k = .25$ by default, with lower values to be used when exploring inhomogeneous datasets (see below for user options). To avoid premature abortion of iterations when using low resample ratios, the algorithm now automatically adjusts the resample ratio when encountering low voxel counts in order to keep the minimum bootstrap sample size *min* constant (e.g., if a dataset with size $_i = 9$ voxels is explored with $k = .5$, $min = 5$ is not met; therefore, *k* is adjusted to be $k = .55$ such that size $_i * k \geq min$; this adaptation is naturally restricted by $k \leq 1$). Even using very low *k*-values (e.g., an exploratory $k = .05$), results in an unaltered dataset remain virtually indistinguishable from a straightforward lateralization index (Table 2). It should also be noted that even with such low resample ratios, the input data will still be oversampled (if $k = .05$, an input sample of size $_i = 100$ will still yield 100 resamples of size $_r = 5$, totaling 500 datapoints).

For unaltered fMRI-data the variability of the resulting lateralization indices reaches a maximum of .2%, strongly decreasing with increasing thresholds (Figure 4, lower panels). The range of detected values (difference between minimum and maximum LI) shows the same pattern, at low thresholds a maximum range of 17% is detected. This indicates that, for “normal” fMRI data, virtually identical results can be expected even with lower resample ratios ($k < 1$; Table 2). Interestingly, this range is higher at lower thresholds, which is again most likely due to the upper size limit we exposed on the bootstrap sample size in order to speed up processing (in these calculations

max = 1000 voxels). Alternatively, a higher data inhomogeneity at lower thresholds could be responsible for this effect (as both low and high values enter the sampling, instead of only high values at higher thresholds). Ultimately, as overall processing of a typical fMRI dataset is completed in less than a minute on a standard PC workstation, this upper limit can be adjusted. Based on these results, the algorithm now uses a maximum sample size of *max* = 10.000 voxels. Of note, this suggestion is based on the simulations conducted here, using typical fMRI data. For high-resolution data, a higher limit or even no limit may be more adequate (to this effect, "inf" can be entered when prompted for the upper bootstrap size)

We used a trimmed mean for averaging the lateralization indices at each threshold in order to only assess the central and most representative parts of the lateralization index-matrix from the bootstrapped samples, emphasizing robustness by being much less vulnerable to outliers (Hesterberg *et al.*, 2005). The excellent agreement in all cases with straightforward LI-calculations (Table 2) constitutes a validation of the approach. The positive effect on stability is also apparent in Figure 5: although the outliers severely influence the range of obtained LI-values on the side of the artifacts, the resulting trimmed mean is much closer to the upper limit of detected lateralization indices. Such an imbalance between the trimmed mean and the distance to the upper and lower bounds is a further criterion for uneven data homogeneity between the two sides. While it is currently only a visible indicator, a mathematical marker could be derived and implemented (e.g., the difference between a trimmed $_{(25)}$ and an arithmetic [regular] mean); at this point, however, we believe that an additional parameter would only over-complicate the already complex graphical output. As to the rigid control of upper and lower bounds, relaxing the criterion of how much data to be "trimmed"

may be an option to increase sensitivity (e.g., using a trimmed mean$_{(10)}$ instead). However, a detailed examination of this effect was not done here and remains a question for future research.

The disadvantage of using a trimmed mean along the x-direction (i.e., to assess an overall lateralization index over all thresholds) is that the higher values obtained at higher thresholds will also be discarded. This is counterproductive as voxels surviving a higher threshold in functional MR-imaging data do this due to their stronger (and ultimately, significant) correlation with the task at hand (Holland *et al.*, 2001; Marchini & Presanis, 2004; Nichols & Hayasaka, 2003). In this case, therefore, it seems justified to give more weight to lateralization indices obtained from such voxels, without specifying a hard cut-off. To this effect, we implemented a weighted mean (see equation 3) to calculate an overall lateralization index from the (trimmed) mean values obtained at all thresholds, therefore increasing specificity. Ultimately, the decision on which value to use is again one of sensitivity versus specificity: a straight mean will weigh all lateralization indices the same way, whether they come from high or low thresholds. A trimmed mean will effectively exclude outliers, but will, if applied to all values, exclude low as much as high values in order to yield a robust mean. Lastly, a weighted mean (based on the trimmed means from all thresholds) will be rather immune to outliers and will give proportionately more weight to values obtained from higher thresholds. Considering the multitude of possible scenarios in which the assessment of lateralization is of interest, the decision on which value to choose cannot be expected to be the same for all cases (therefore, all values are reported, see Figure 5). We believe that, if no indicator suggests significant outlier influence, a weighted mean over all thresholds is a good compromise.

A number of variables influence the results from this algorithm: the lower threshold cutoff, the resample ratio, the minimum and the maximum bootstrap sample size. While we suggest to explore the whole range of thresholds in an image, a lower cutoff may be specified by the user, so that the option to only investigate “significant” voxels, however defined, remains. To allow flexible explorations, the user is requested to confirm the defaults (lower cutoff = 0, k = .25, minimum bootstrap sample size = 5, maximum bootstrap sample size = 10.000) or to specify his own settings. Additionally, the open nature of Matlab script files allows changing all other relevant settings within the (thoroughly documented) code.

To summarize, our algorithm not only supplies a lateralization curve, describing lateralization at different thresholds, but also a comprehensive, single lateralization index based on large body of data. At the same time, several parameters allow for the assessment of the underlying data quality, thereby offering a decisive advantage over previous approaches. We therefore conclude that the application of the bootstrapping concept to calculating lateralization indices from imaging data is fast, robust, and powerful.

References

Adcock JE, Wise RG, Oxbury JM, Oxbury SM, Matthews PM (2003)
Quantitative fMRI assessment of the differences in lateralization of language-related brain activation in patients with temporal lobe epilepsy
NeuroImage 18: 423-438

Auffermann WF, Ngan SC, Hu X, 2002
Cluster significance testing using the bootstrap
NeuroImage 17: 583-591

Cabeza R, Nyberg L, 2000
Imaging cognition II: An empirical review of 275 PET and fMRI studies
J Cogn Neurosci 12: 1-47

Davison AC, Hinkley DV, 1997
In Davison AC, Hinkley DV: Bootstrap Methods and their Application, 1st Edition, Cambridge University Press, Cambridge

Deblaere K, Boon PA, Vandemaele P, Tieleman A, Vonck K, Vingerhoets G, *et al*., 2004
MRI language dominance assessment in epilepsy patients at 1.0 T: region of interest analysis and comparison with intracarotid amytal testing
Neuroradiology 46: 413-420

Fernandez G, de Greiff A, von Oertzen J, Reuber M, Lun S, Klaver P, *et al.*, 2001
Language mapping in less than 15 Minutes: real-time functional MRI during routine clinical investigation
NeuroImage 14: 585–594

Gaillard WD, Balsamo L, Xu B, Grandin CB, Braniecki SH, Papero PH, *et al.*, 2002
Language dominance in partial epilepsy patients identified with an fMRI reading task
Neurology 59: 256–265

Hesterberg T, Moore DS, Monaghan S, Clipson A, Epstein R, 2005
Bootstrap Methods and Permutation Tests. In: Moore DS, McCabe GP (eds.): Introduction to the Practice of Statistics, 5th Ed., WH Freeman & Co, 14.1-70

Holland SK, Plante E, Byars A, Strawsburg RH, Schmithorst VJ, Ball WS Jr. (2001)
Normal fMRI brain activation patterns in children performing a verb generation task
NeuroImage 14: 837-843

Hugdahl K, Davison RJ, 2002
The Asymmetrical Brain, 2nd ed.
MIT Press

Janssen A, Pauls T, 2003
How do bootstrap and permutation tests work?
Ann Statist 31: 768–806

Jones DK, Pierpaoli C, 2005

Confidence mapping in diffusion tensor magnetic resonance imaging tractography using a bootstrap approach

Magn Reson Med 53: 1143-1149

Lazar M, Alexander AL, 2005

Bootstrap white matter tractography (BOOT-TRAC)

NeuroImage 24: 524-532

Marchini J, Presanis A, 2004

Comparing methods of analyzing fMRI statistical parametric maps

NeuroImage 22: 1203-1213

Moore DS, McCabe GP, Duckworth WM II, Sclove SL, 2002

In: Moore DS, McCabe GP, Duckworth WM II, Sclove SL (eds.): The Practice of Business Statistics: Using Data for Decisions, 1st Ed., WH Freeman & Co, 18.1-73

Nichols T, Hayasaka S, 2003

Controlling the familywise error rate in functional neuroimaging: a comparative review

Stat Methods Med Res 12: 419-446

Oldfield RC, 1971

The assessment and analysis of handedness: the Edinburgh inventory

Neuropsychologia 9: 97-113

Prohovnik I, Skudlarski P, Fulbright RK, Gore JC, Wexler BE, 2004

Functional MRI changes before and after onset of reported emotions

Psychiatry Res 132: 239-250

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, *et al.*, 2002

Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain

NeuroImage 15: 273–289

Wilke M, Lidzba K, Staudt M, Buchenau K, Grodd W, Krägeloh-Mann I, 2006

An fMRI task battery for assessing hemispheric language dominance in children

NeuroImage (in press)

Acknowledgements

We would like to thank Ingeborg Krägeloh-Mann, MD, PhD, Wolfgang Grodd, MD, PhD, and Scott K. Holland, PhD, for their continued support. This work has been supported by the *Deutsche Forschungsgemeinschaft* DFG (SFB550/C4). The algorithm is part of our LI-toolbox and is available free of charge for scientific use. Interested Researchers are encouraged to contact the authors at Marko.Wilke@med.uni-tuebingen.de

Subject	***Classical lateralization indices***		***Bootstrapped lateralization indices (k = 1)***		
	FWE	**FDR**	**Overall mean**	**Trimmed Mean $_{(25)}$**	**Weighted Mean**
1 (m, 12y)	0,65	0,4	0,44	0,41	0,5
2 (m, 15y)	N/A	N/A	0,74	0,76	0,85
3 (m, 9y)	N/A	N/A	0,87	0,86	0,92
4 (f, 8y)	N/A	0,59	0,53	0,46	0,67
5 (f, 13y)	0,97	0,56	0,58	0,55	0,73

Table 1: Demographic details and lateralization indices from the unaltered fMRI example datasets (for details, see Wilke *et al.*, 2006).

Threshold	Classical lateralization index	Bootstrapped lateralization indices			
		Mean [k = 1] (Range)	Mean [k = .5] (Range)	Mean [k = .25] (Range)	Mean [k = .05] (Range)
0	.355	.353 (.299-.409)	.358 (.306-.410)	.354 (.293-.414)	.355 (.299-.412)
.29	.361	.360 (.313-.401)	.361 (.296-.405)	.361 (.310-.401)	.358 (.313-.427)
.58	.376	.375 (.339-.411)	.375 (.336-.419)	.376 (.335-.406)	.377 (.338-.422)
.87	.395	.394 (.361-.428)	.394 (.365-.431)	.394 (.362-.426)	.394 (.349-.434)
1.16	.410	.411 (.385-.433)	.410 (.384-.432)	.410 (.387-.435)	.410 (.372-.444)
1.45	.410	.410 (.388-.431)	.410 (.386-.430)	.410 (.391-.430)	.410 (.372-.438)
1.74	.380	.380 (.358-.398)	.380 (.360-.397)	.381 (.362-.400)	.381 (.34-.414)
2.03	.344	.344 (.329-.361)	.345 (.329-.361)	.346 (.324-.363)	.346 (.315-.373)
2.32	.307	.306 (.291-.322)	.307 (.294-.322)	.307 (.288-.324)	.307 (.279-.34)
2.61	.293	.293 (.279-.306)	.293 (.281-.308)	.293 (.280-.306)	.294 (.256-.322)
2.9	.331	.331 (.321-.343)	.331 (.319-.341)	.332 (.319-.347)	.331 (.296-.358)
3.19	.396	.396 (.387-.404)	.396 (.386-.405)	.396 (.378-.408)	.395 (.365-.421)
3.48	.467	.467 (.460-.475)	.467 (.457-.476)	.467 (.454-.482)	.467 (.439-.493)
3.77	.562	.562 (.555-.568)	.562 (.552-.570)	.562 (.548-.572)	.562 (.53-.589)
4.06	.608	.609 (.603-.615)	.608 (.600-.616)	.609 (.594-.621)	.609 (.585-.637)
4.35	.614	.614 (.607-.620)	.614 (.605-.624)	.614 (.602-.626)	.614 (.582-.641)
4.64	.645	.645 (.637-.653)	.645 (.635-.654)	.645 (.627-.658)	N/A
4.93	.703	.703 (.695-.710)	.703 (.693-.713)	N/A	N/A
5.22	.497	.497 (.485-.507)	N/A	N/A	N/A

Table 2: Comparison of results form a classical lateralization index calculation from different thresholds with bootstrapped results (k = 1/.5/.25/.05). N/A: not available with constant k due to decreasing voxel numbers.

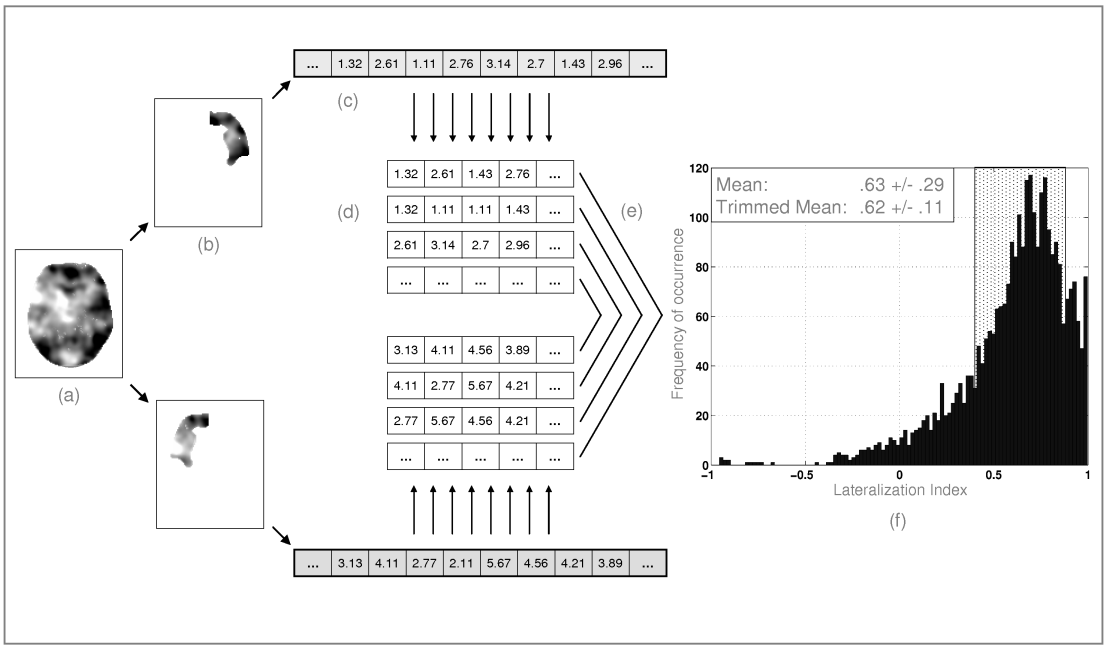

Figure 1: Overview of the bootstrapping procedure: an input image (a) is thresholded and masked, yielding data for the left and the right side (b), which is then converted to a vector containing all voxel values (c). From this, *n* bootstrapped resamples are generated (d; default: 100) from which all possible lateralization index combinations are calculated (e; default: 10.000). All resulting values are then plotted in a histogram (f), from which only the central 50% are used (shaded area), "trimming" the upper and lower 25% of datapoints. This procedure is repeated at each thresholding step.

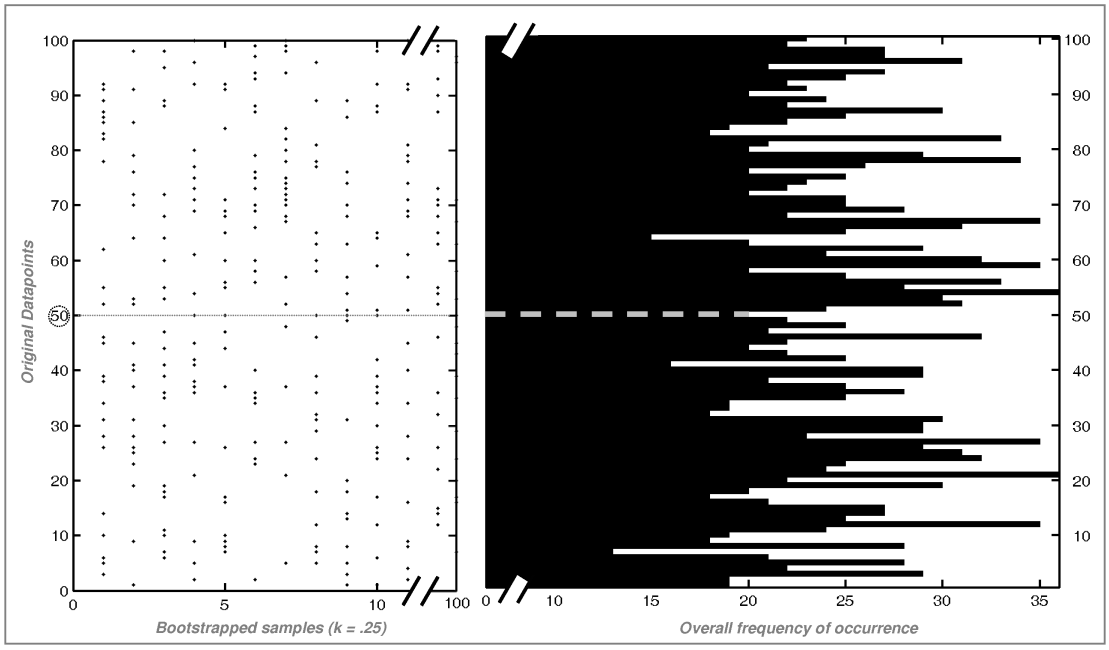

Figure 2: Left panel: illustration of the bootstrap approach: an input sample with 100 datapoints is randomly sampled, with replacement, 100 times with a resample ratio of $k = .25$. Right panel: resulting histogram of the presence of each point in the resulting resamples.

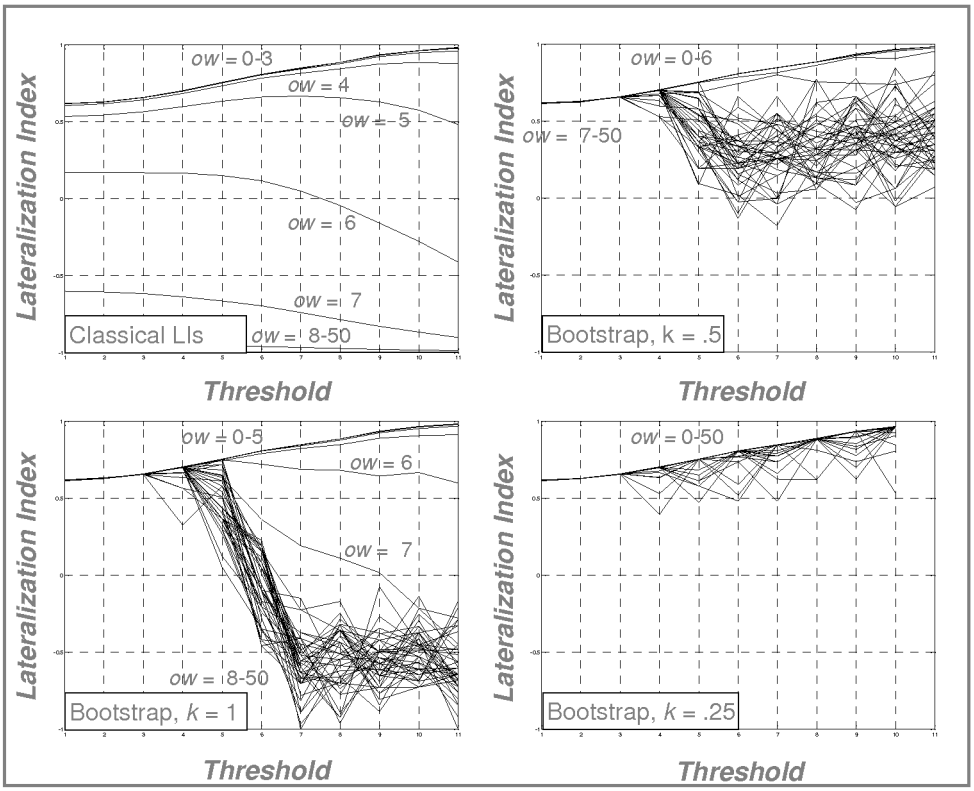

Figure 3: Effect of outliers of different values (outlier weight *ow*) on classical lateralization curves (top left) and on bootstrapped lateralization curves with different resample ratios. Note increasing stability towards outliers with decreasing resample ratio.

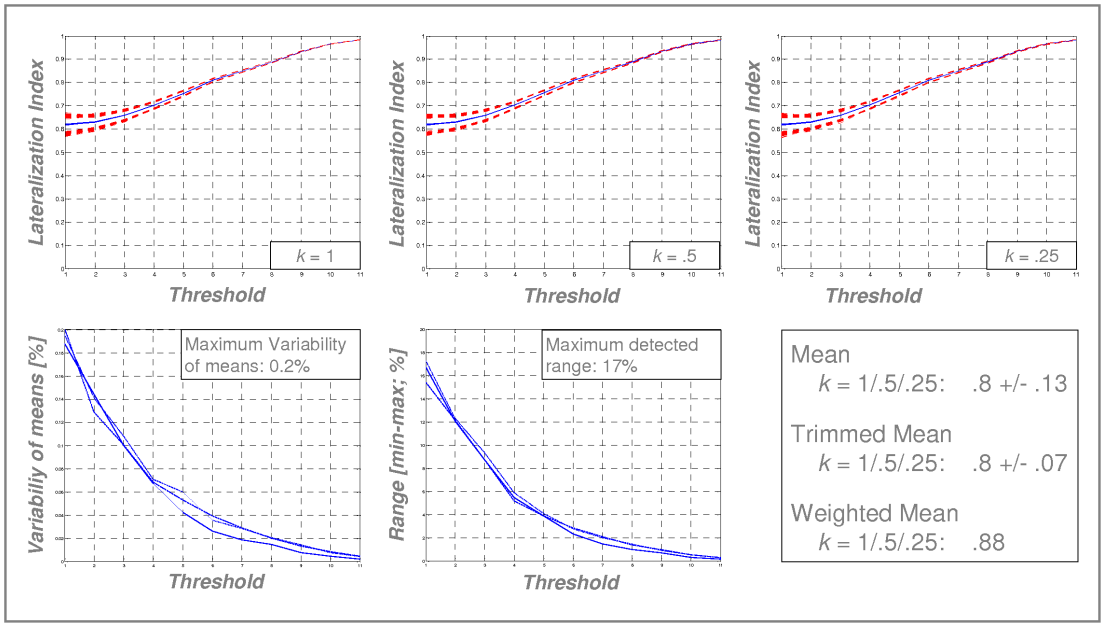

Figure 4: Upper panels: illustration of re-run stability of the bootstrapped lateralization curves: 100 re-runs of the same dataset with different resample ratios *k*. Solid lines: mean lateralization indices; slashed lines: minimum/maximum detected values. Note virtually identical results independent of *k*. Lower panels: Variability and range for different resample ratios (solid line: $k = 1$; slashed line: $k = .5$; dotted line: $k = .25$). Lower right panel: identical overall results independent of *k*.

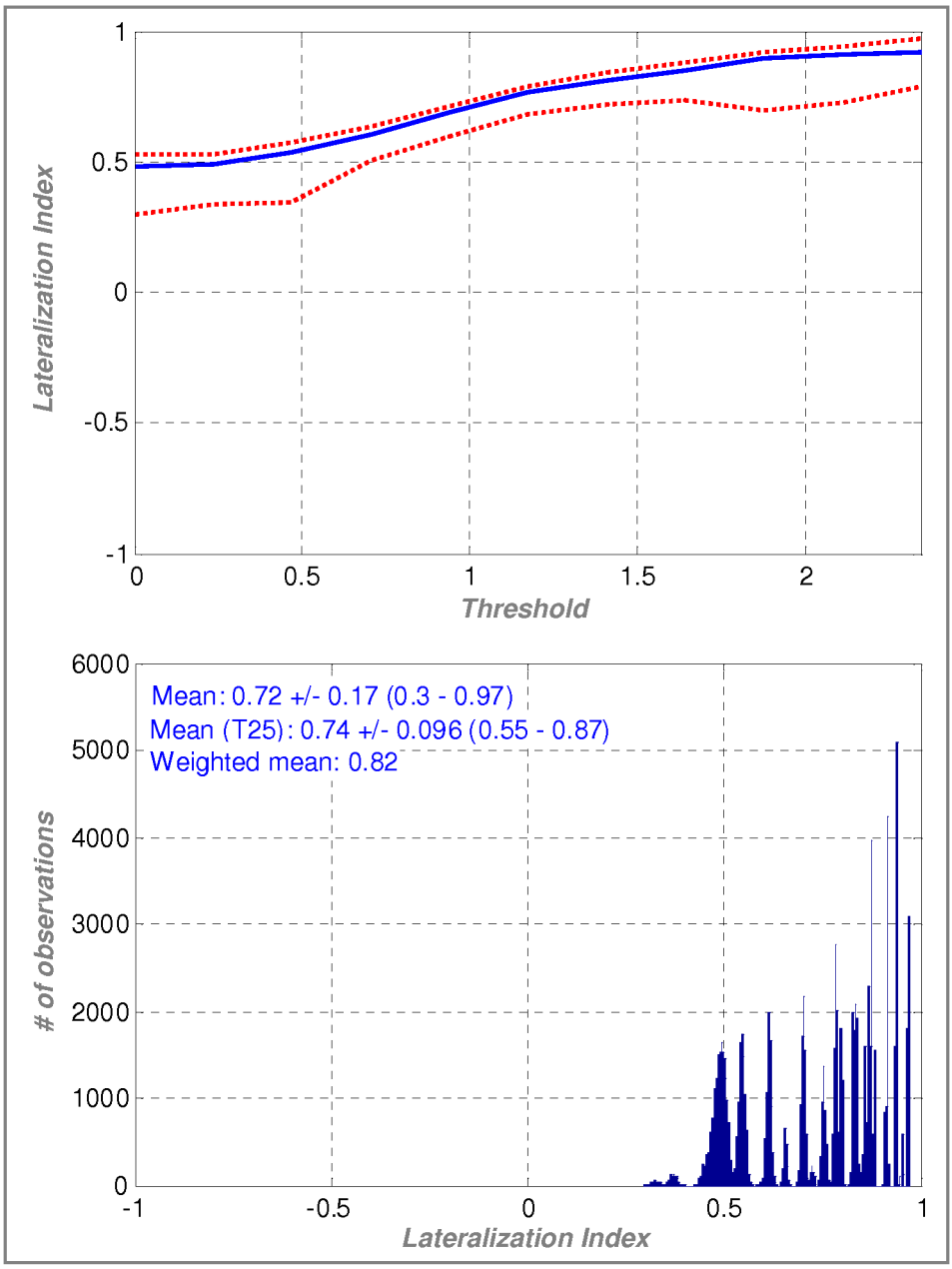

Figure 5: Actual algorithm output for a single dataset (lateralization within the frontal lobe), containing two outliers on the right (p1/2, see also Figure 6). Upper panel: note much wider spread of detected minimum lateralization indices on the right side (lower dotted line) and closer adherence of the trimmed mean (solid line) to the detected maximum values (upper dotted line). Also note absence of routinely included number of voxels at each threshold for better visibility. Lower panel: histogram of the overall lateralization index-matrix: note several small, irregularly shaped histogram contributions from outliers.

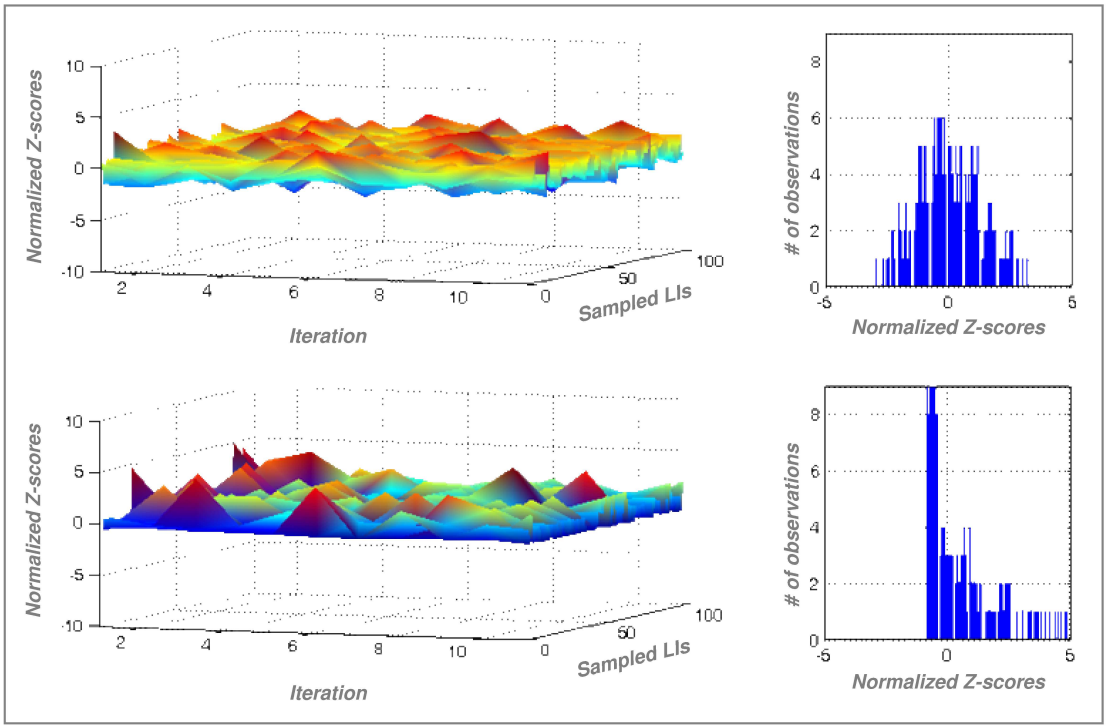

Figure 6: Actual algorithm output for a single dataset (lateralization within the frontal lobe), containing two outliers on the right (p2/2, see also Figure 5). Upper panels: surface representation of all z-scores from all iterations and the resulting histogram for the LEFT side; lower panels: corresponding data from the RIGHT side, clearly showing the outlier influence in both the surface plot and the skewed histogram (see text for details)