Trimmed estimators for robust averaging of event–related potentials

Zbigniew Leonowicz [1,*] Juha Karvanen and Sergei L. Shishkin [2]

Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Saitama 351-0198, Japan

Abstract

Averaging (in statistical terms, estimation of the location of data) is one of the most commonly used procedures in neuroscience and the basic procedure for obtaining event-related potentials (ERP). Only the arithmetic mean is routinely used in the current practice of ERP research, though its sensitivity to outliers is well-known. Weighted averaging is sometimes used as a more robust procedure, however, it can be not sufficiently appropriate when the signal is nonstationary within a trial. Trimmed estimators provide an alternative way to average data. In this paper, a number of such location estimators (trimmed mean, Winsorized mean and recently introduced trimmed L–mean) are reviewed, as well as arithmetic mean and median. A new robust location estimator *tanh*, which allows the data–dependent optimization, is proposed for averaging of small number of trials. The possibilities to improve signal-to-noise ratio (SNR) of averaged waveforms using trimmed location estimators are demonstrated for epochs randomly drawn from a set of real auditory evoked potential data.

Key words: averaging, event–related potentials, evoked potentials, mean, median, trimmed mean, robust estimators of location, trimmed estimators

1 Introduction

Averaging is probably the most common basic statistical procedure in experimental science. It is used for estimating the location of data (or "central tendency") in the presence of random variations among the observations, which can be removed by this procedure. Data variations can be the result of variations in the phenomenon of interest or of some unavoidable measuring errors. In signal processing terms, this can be considered as contamination of useful "signal", such as event–related brain activity, by useless "noise", such as artifacts and ongoing activity, both not repeatedly associated with the event. In the case of linear summation of signal and noise ("additive model"), the fact that only the event–related signal is time–locked to the event and noise is not time–locked, allows the cancellation of the noise by averaging the data separately for each time point relative to the event. Averaging is typically done using arithmetic mean, which is the most widely known estimator of the location of the data.

Event–related averaging is important for various techniques from single neuron firing recording to optical imaging, but is most essential for the old and

* Corresponding author. Address: Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, 2–1 Hirosawa, Wako–shi, Saitama 351-0198, Japan. Phone: +81–48–4679765 Fax: +81–48–4679694

Email addresses: Zbigniew.Leonowicz@pwr.wroc.pl (Zbigniew Leonowicz), juha.karvanen@hut.fi (Juha Karvanen), shishkin@brain.riken.jp (Sergei L. Shishkin).

URL: http://www.bsp.brain.riken.jp.

[1] on leave from Wroclaw University of Technology, Wroclaw, Poland

[2] Dept. of Human and Animal Physiology, Faculty of Biology, M. V. Lomonosov Moscow State University, Russia

still important technique of event–related potentials (ERP). ERPs are voltage fluctuations repeatedly associated in time with some physical or mental events, which can be extracted from the ongoing electroencephalogram (EEG) using signal averaging (Picton et al., 2000). The term "evoked potentials" (EPs) is used for an important subset of ERPs which are "evoked" by a certain event, usually a sensory one, but not by independent endogenous processes. Additive model is most commonly accepted for event-related activity and ongoing EEG, allowing the use of averaging for the extraction of ERP. Though a number of studies cast doubts on the additive model by demonstrating that event–related modifications of the ongoing EEG may also contribute to ERP (Makeig et al., 2002; Jansen et al., 2003), averaging has been proved to be practically a very efficient procedure and is used, therefore, as the basic procedure in ERP analysis (Picton et al., 1995; Lopes da Silva, 1999; Picton et al., 2000).

Temporal resolution in the order of milliseconds makes ERPs an important tool for estimating of the timing for information processing in the human brain (Picton et al., 2000). Thus, the improvement of signal–to–noise ratio by averaging, which is the essence of computing the ERP, should receive a serious attention. In particular, an important task is the development of efficient methods for averaging of small sets of single–trial ERPs, where data may strongly deviate from Gaussian distribution. Due to deviation from Gaussian distribution, their location may not be estimated correctly by arithmetic mean. The number of data epochs (which correspond to experimental trials) used for averaging should be as low as possible because of various reasons: fatigue, learning and other factors affecting the brain response of the subject; even the way the subjects perform the task may change if the task is repeated

many times. In some applications such as Brain–Computer Interface (BCI) only a few trials can usually be used each time the averaging is computed (Wolpaw et al., 2002).

In some cases the improvement of ERP averaging can be obtained with various techniques which compensate for the trial–to–trial variability, mainly the latency jitter. Such approaches, however, require that the ERP could be recognizable in single trials, which is often impossible, and involve the risk that the outcome can be merely the result of lining up the background noise (Picton et al., 1995, 2000).

Improving the noise reduction by averaging can be obtained with a technique called weighted averaging. In this method, each epoch is given a weight depending on estimated noise in this epoch (Hoke et al., 1984; Lütkenhöner et al., 1985; Davila and Mobin, 1992; Łęski, 2002). Though quite many variations of weighting averaging have been developed (for review see, e.g., Łęski (2002)), this method is still rarely used for averaging of ERP. As stated in Özdamar and Kalayci (1999), there are still unsolved issues related to computation of weights and their influence on the result of averaging. An important limitation of weighted averaging comes from the noise model which it assumes. According to this model, noise varies between epochs, but is stationary within each epoch (Lütkenhöner et al., 1985). However, many types of noise (not only artifacts but also waves of ongoing EEG) can strongly vary within an epoch, having high amplitude in some time points and low amplitudes in other ones. Thus, the weights can be underestimated for the part of epoch where a strong noise occurs and overestimated for another part of epoch where the noise is low.

Median averaging is another approach suggested for minimizing the influence of noise in ERPs (Borda and Frost, 1968; Yabe et al., 1993; Picton et al., 1995; Özdamar and Kalayci, 1999; Fox and Dalebout, 2002). It is similar to conventional averaging on the basis of arithmetic mean, with only difference that median is used instead of mean. Computation of median includes ordering of samples according to their amplitudes of all epochs for each time point relative to stimulus, independently from other time points. Due to this independence, median averaging cannot be affected by the nonstationarity of noise within an epoch. It was shown that median averaging can improve averaging of endogenous ERPs using small number of trials (Yabe et al., 1993). A detailed study by Özdamar and Kalayci (1999) demonstrated the advantages of median averaging over conventional averaging for auditory brain stem responses which have low SNR due to very low amplitude and thus requires high number of epochs to be averaged.

The disadvantage of median averaging is that it does not only remove the outliers but also uses the rest of data only in the sense of the order of the values. It is evident that some useful information can be lost by this procedure, comparing to conventional averaging, which employs the data values themselves instead of their order. In practice, median averaged waveforms include a rather strong high frequency noise (though it seems to be easily removable by filtering (Özdamar and Kalayci, 1999)), and the results are not always improved relative to conventional mean averaging (Fox and Dalebout, 2002). One should also consider a possibility of unpredictable effects arising from the "over–robustness" of median. For example, the value of median will not change at all if we add a very large value to each of data values above median (Streiner, 2000).

Is it possible to combine advantages of mean and median averaging? In fact, an estimator of data location which lies between those two extremes already exists and its name is "trimmed mean". In this method (see the exact definition in the next section), a part of extreme values is discarded or modified, but all other values are used for averaging in the same way as in conventional mean averaging.

It is important to note that rejection of extreme values is quite different from the procedure of artifact rejection. The latter procedure typically implies removing not only extreme values but all trials including extreme values or some other signs of artifacts and in this sense it is closer to weighted averaging, which defines the impact of an epoch by estimating it as a whole. Surprisingly, though trimmed mean and its modified version, Winsorized mean, are efficient robust location estimators (Stuart, 1994), averaging of ERPs on the basis of these estimators has never been reported, to the best of our knowledge.

The goal of this paper is to demonstrate the efficiency of trimmed estimators of data location for computing ERPs and to propose some ways to optimize their parameters, such as trimming parameters and parameters for weights related to order of averaged amplitudes.

In this paper, we:

(1) briefly review a number of locator estimators, namely mean, median, and three trimmed estimators: trimmed mean, Winsorized mean and recently proposed trimmed L–mean,
(2) report results of their testing using real auditory evoked potential (AEP) data and their resampling,
(3) propose a new adaptable location estimator.

2 Statistical estimators of location

2.1 Problem of robustness of estimation of the data location

The problem of sensitivity of an estimator to the presence of outliers, i.e. "the data points that deviate from the pattern set by the majority of the data set" (Hampel et al., 1986), has lead to the development of *robust* location measures. Robustness of an estimator is measured by the *breakdown value*, which tells us how many data points need to be replaced by arbitrary values in order to make the estimator explode (tend to infinity) or implode (tend to zero). For instance: arithmetic mean has 0% breakdown whilst median is very robust with breakdown value of 50% (Hampel et al., 1986).

2.2 Arithmetic mean

The most widely used statistical measure and the best known estimator of location is the *arithmetic mean* (see Figure 1(a)).

$$\mu_{mean} = \sum_{i=1}^{N} \frac{1}{N} x_i \tag{1}$$

The arithmetic mean is a standard location estimator used for averaging of ERPs and for many other purposes, however, it is not robust. In the case of arithmetic mean, only one outlier may make the estimate infinitely large or small. The breakdown value becomes here $\lim_{N\to\infty} \frac{1}{N} = 0$.

2.3 Robust location measures

Many location estimators can be presented in unified way by ordering the values of the sample as $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(N)}$ and then applying the weight function w_i (Stuart, 1994)

$$\mu_r = \sum_{i=1}^{N} w_i x_{(i)}, \tag{2}$$

where w_i is a function designed specifically to reduce the influence of certain observations (data points) in form of weighting and $x_{(i)}$ represents the ordered data. For the arithmetic mean it holds $w_i = \frac{1}{N}$.

To make the comparison between different estimators easier, we present all weighting functions (as in (2)) plotted in Figure 1.

2.4 Median

Suppose that the data have the size of $(2M+1)$, where M is a positive integer, then the median is the value of the $(M+1)^{th}$ ordered observation. In the case of even data size $2M$ the median is defined as the value of the mean of the samples M and $M+1$. According to the framework of equation (2) the weight one is applied to the $(M+1)^{th}$ sample in the case when the number of samples is odd and weights equal to $\frac{1}{2}$ to both M^{th} and $(M+1)^{th}$ samples when the number of samples is even (Stuart, 1994) (see Figure 1(b)).

2.5 Trimmed mean

For the α–trimmed mean (where $p = \alpha N$) the weights w_i as in (2) can be defined as:

$$w_i = \begin{cases} \frac{1}{N-2p}, & p+1 \leq i \leq N-p \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Thus, the trimmed mean correspond to the mean value of data samples where p highest and p lowest samples are removed (see Figure 1(c)).

Application of trimming lowers the influence of extreme data values on the result of averaging. However, unlike in median, substantial part of data can be included into average.

2.6 Winsorized mean

In the case of trimmed mean, the tails of the distribution of the data are simply ignored. It can lead to the loss of information and should be avoided when the sample size is small. Winsorized mean is similar to trimmed mean with the exception that it replaces each observation in each α fraction ($p = \alpha N$) of the tail of the distribution by the value of the nearest unaffected observation. Weight w_i becomes here (see Figure 1(d))

$$w_i = \begin{cases} 0, & i \leq p \text{ or } i \geq N-(p-1) \\ \frac{p+1}{N}, & i = p+1 \text{ or } i = N-p \\ \frac{1}{N}, & p+2 \leq i \leq N-(p+1) \end{cases} \tag{4}$$

Usually, the values in the range $0 \leq p \leq 0.25N$ are considered, depending on the heaviness of the tails of the distribution.

An interesting observation is that the *median* can be viewed as an extreme case of the *trimmed mean* or *Winsorized mean* when only one or two central data points are retained (Stuart, 1994).

2.7 Trimmed L–mean (TL–mean)

Recently, Elamir and Seheult (2003) proposed the trimmed *L*–moments (*TL*–moments) as a generalization of *L*–moments (Hosking, 1990). The *TL*–mean can be estimated from a sample as a linear combination of order statistics. Using the formulation of equation (2) the weight function can be calculated as follows ($p = 0$ for arithmetic mean)

$$w_i = \begin{cases} \frac{\binom{i-1}{p}\binom{N-i}{p}}{\binom{N}{2p+1}}, & p+1 \leq i \leq N-p \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $\binom{i}{p} = \frac{i!}{p!(i-p)!}$. Equation (5) reveals the connection between the *TL*–mean and the trimmed mean. In the calculation of the both statistics, the extreme observations are ignored. The main difference is that the trimmed mean applies the equal weight to the remaining observations whereas the *TL*–mean uses higher weights for the observations near the median (see Figure 1(e)). The variance of various location estimators was compared in Elamir and Seheult (2003) for samples originating from normal, logistic, double–exponential and normal distribution with outliers. The conclusions of this simulations are that TL–mean is suitable for general use and performs reasonably well for normal and heavy–tailed distributions. Such information is not precise enough for the purpose of averaging of ERPs and further investigations are necessary.

2.8 New empirical estimator: tanh mean

We propose a new estimator to alleviate the problem of enhancing noise by robust estimators such as trimmed mean or median. The weights are calculated using the hyperbolic tangent functions (see Figure 1(f)) in such way that the signal samples are weighted by

$$w_i = \begin{cases} \tanh[k(i)] - s, & i < \frac{N}{2} \\ -\tanh[k(\frac{i}{2} - N)] + s, & i \geq \frac{N}{2} \end{cases} \tag{6}$$

where k is the factor controlling the slope of the weights for extreme values in data and s determines the vertical shift. The robustness is achieved by setting all weights with negative values to zero so the extreme observations will be ignored, like for trimmed mean, Winsorized mean or TL-mean. Such formulation allows the control over the robustness (cancelling of extreme values) and over the influence of extreme values on the final estimate of the averaged trial. The shape of the weighting function depends on parameters k and s. This estimator belongs to the group of trimmed estimators by its ability to cancel of extreme values. However, its main feature is a flexible and optimized adjusting of the degree of influence of the values close to the extremes on the result of averaging.

Both parameters k and s can be optimized to achieve the best possible value of signal–to–noise ratio (SNR) (or of another computable performance index) of the averaged trials. In order to avoid that the estimator becomes biased it is necessary to estimate the parameters on a separate data set ("hold–out" data set) different from that for which the SNR is computed. For the results reported below in this paper the optimization was accomplished using

the Nelder–Mead (Simplex) method of unconstrained nonlinear optimization (Nelder and Mead, 1965). The Nelder–Mead method does not require the objective function (such as SNR) to be differentiable. Therefore it is well–suited to problems involving a non–differentiable objective function of a small number of decision variables. Optimization of the new estimator involves the function of only two variables. The Nelder–Mead (Simplex) method is usually embedded in standard optimizing packages.

In order to address the small sample performance (which shows how the performance of the estimator degrades when decreasing the number of available data) (Gastwirth and Cohen, 1979) an experiment was carried out (see the section 4 for description and results).

3 Experimental setup

Three healthy male subjects with normal hearing (age 28–40) participated in the study. They seated in a chair in a dimly illuminated electrically shielded room with closed eyes and were asked to be relaxed and listen to the sequence of sound stimuli, namely 800 Hz (frequent) and 1200 Hz (rare) tones, do not pay attention to frequent tones but press a button with right index finger as soon as possible when they hear the rare high–pitch tone.

Stimuli were presented using Neuroscan (*http://www.neuro.com*) STIM system binaurally through earplugs (Tubephone Insert Earphones ER-3 ABR, by Etymotic Research *http://www.etymotic.com*) at 70 dB SPL. Each sound stimulus was a pure sine wave with duration of 35 ms (first and last 5 ms were linearly rising and falling). After a short practice session, experimental sessions

(3 sessions for subjects 1 and 2, 2 sessions for subject 3) with breaks between them were run, with 689 stimuli (both rare and frequent) presented in each of them. Stimulus onset asynchrony (for both types of stimuli) was random numbers uniformly distributed between 0.4 and 1.6 s. Rare tones were approximately 10% of the total number of stimuli. The number of frequent stimuli between rare stimuli was random, however, two constrains were set: stimulus onset asynchrony between a rare stimulus and the next frequent stimuli could not be below 1.2 s; and at least 3 frequent stimuli were placed between each two random stimuli.

The EEG was recorded with Neuroscan Data Acquisition System (software version 4.3.1, SynAmps 5083) from C_3 and C_4 electrodes referenced to electrically linked earlobes. Amplifier gain was 1000, A/D bit size 12 (providing the resolution of 0.084 μV/bit), amplitude range 5.5 mV, sampling rate 500 Hz, analog filter bandpass 0.1-70 Hz (slopes 12dB/octave), notch filter was not used. Impedance was below 5 kΩ for subjects 1 and 2. For the subject 3 it was possible to achieve the impedance between 20 and 40 kΩ and the EEG seemed to be slightly distorted, but no line noise was visible (this EEG was used as an example of a recording of lower quality).

Only the data recorded from C_4 electrode was used. Before segmenting into epochs, the EEG was low–pass filtered (with zero phase shift, 96 dB/oct. digital filter) below 30 Hz. For the analysis, $\pm$500 ms epochs related to frequent stimuli (not requiring the response) were extracted. The 16 frequent stimuli from the beginning of each session up to 2^{nd} such stimulus after the first rare stimulus in the session and all frequent stimuli immediately following any rare stimulus were not used in the further analysis. For a few cases of missed or wrong responses (not related to rare stimulus), 2 frequent stimuli

immediately preceding and 2 immediately following the related stimuli were also excluded from the analysis. This resulted in 1787, 1796 and 1193 valid epochs for subjects 1, 2 and 3, respectively (below: raw data). Some of these epochs contained artifacts, such as electromyogram (EMG) or artifacts caused by movements, but no clipped data.

A "cleaned" subset of the data was formed by rejecting all epochs in which the amplitude in any time point (within the interval of ± 500 ms relative to the stimulus) exceeded a threshold of 50 μV. Visual inspection confirmed that the majority of these epochs contained some artifacts and that no visible artifacts (except, in rare cases, low amplitude EMG) were present in the remaining epochs. The number of epochs in "cleaned" subsets was 1752, 1776 and 1104 for subjects 1, 2 and 3, respectively.

4 Results

Examples of averaged waveforms for all three subjects are given in Figure 2. Median averaging produced more noisy averaged ERPs, which is more evident in the prestimulus time interval (where, ideally, the averaged signal should be close to zero). The difference between the other estimators is, however, not clearly visible in these plots.

We estimated the efficiency of different averaging methods and different trimming parameters by estimating SNR in randomly chosen subsets of epochs. It is commonly assumed (e.g., Davila and Mobin (1992)) that signal is contained in the poststimulus part of the averaged data, while noise is included in both poststimulus and prestimulus part. Prestimulus variance usually was

much lower than poststimulus variance in our data, thus the impact of noise variance to poststimulus interval was insignificant and we considered the poststimulus part (approximately) as "signal". The following formula was used, therefore, for the calculation of signal–to–noise ratio:

$$\mathrm{SNR} = 10 \cdot \log_{10} \frac{\sigma^2_{poststim.}}{\sigma^2_{prestim.}} \ [\mathrm{dB}] \tag{7}$$

where $\sigma^2_{poststim.}$ is signal plus noise variance, defined as variance of the poststimulus interval (in our data, 2 – 400 ms relative to stimulus onset) of the averaged ERP, and $\sigma^2_{prestim.}$ is "noise variance", defined as variance of the prestimulus interval (in our data, –400 – 0 ms relative to stimulus onset) of the averaged ERP.

The SNR was estimated according to (7) for each subject after averaging, using one of the location estimators, a subset of 31 epochs randomly drawn from the whole "cleaned" set of epochs. The procedure was repeated 100 times and SNR values were averaged (using the arithmetic mean). The results obtained for different estimators and different trimming parameters are presented in Figures 3(a–c). For two subjects (Figure 3(a) and 3(c)), mean averaging produced better SNR not only comparing to median, but also comparing to trimmed estimators almost for all values of trimming parameters. For one subject, best SNR was obtained with trimmed estimators (Figure 3(b)).

The fact that mean performed better than any other estimator in two cases can be related to the high homogeneity of our data, which could be the result of the absence of strong artifacts, stable attention provided by the experimental procedure and other factors. Results were similar for raw data sets

(not shown in the figure) which contained only small number of epochs with artifacts. To investigate effects of stronger inhomogeneity of the data, we simulated the alpha rhythm, an EEG component which often becomes a serious problem for ERP averaging but was low in our recordings for all three subjects. White noise (random normally distributed numbers) was generated and filtered by Butterworth filter of 2^{nd} order with the passband of 9–11 Hz. Randomly selected non-overlapping intervals of this simulated signal were added to randomly selected 20% of epochs of each subject's cleaned data set after multiplying by a constant computed for each epoch in such a way that the resulted maximum of absolute amplitude of added "alpha rhythm" in the epoch was 30 μV. This procedure resulted in the increase of data variance from 108 to 150 for subject 1, from 79 to 122 for subject 2 and from 172 to 212 for subject 3 (computed for all epochs together). The procedure of estimating the SNR described above was repeated for these new "alpha" data sets. The results are presented in Figure 3(d-f). Now, 2 of 3 subjects had the best SNR for the after a trimmed averaging procedure.

The worst results for all three subjects and both without and with adding simulated alpha rhythm were found for median. In the case of subject 3 after addition of the simulated alpha rhythm (Figure 3(d)) SNR was similar for mean and median, but in this case trimmed estimators provided especially prominent improvement of SNR.

It is possible that averaging using trimmed estimators can provide best results if their parameters (trimming parameter for trimmed mean, Winsorized mean and TL-mean, and parameters k and s for the tanh mean) are optimized for a given type of data and/or given sample size. This is illustrated by the example presented in Figure 4. One hundred randomly chosen epochs from subject 1

(raw data set) were averaged using all location estimators discussed in this paper after optimizing the parameters of trimming estimators using a separate set of the same number of randomly chosen 100 epochs, and the SNR was estimated according to (7). Optimization was made using Nelder–Mead simplex method of optimization (Bock, 1998) for the tanh estimator and by choosing the parameter p which maximized the SNR for trimmed, Winsorized and TL-mean. The procedure was repeated after successive removal of an increasing number of randomly selected epochs. Averaged results for 100 repetition of the above procedure are plotted in Figure 4. It was possible for every number of removed epochs to obtain better SNR for trimmed estimators comparing to median and mean (which cannot be optimized). The best performance, especially for smaller number of remaining epochs, was obtained for the tanh mean. The relation between the slope parameter k of the tanh mean and the value of the criterion to be maximized was also investigated. The results (not presented in this paper) show that the global maximum is easily obtained for the new estimator with small risk of stuck into local minima, due to applied method of optimization (Bock, 1998).

5 Discussion

Robust trimmed estimators of data location gradually gain popularity in various statistical applications but have not been adopted for the ERP research yet. The examples presented above demonstrate that trimmed estimators may improve the results of averaging, the procedure which is crucial for ERP analysis. The evidence they provide for such improvement is preliminary due to limited number of tests and subjects in the study. However, we can claim that

if improvement of averaging of ERP is especially important, trimmed estimators should be considered as a possible alternative to conventional averaging.

As alternatives to conventional averaging, weighted averaging and median averaging were considered in the ERP literature so far. In the introduction, we already argued that weighted averaging assumes quite unrealistic noise model. Median averaging seems to be too robust estimator which may discard too large part of information presented in the data. Trimmed estimators are more robust than conventional mean but not as "over–robust" as median. Of course, when artifacts are few or can be easily removed and data are very close to normal, there is no need to use other estimators than mean. But in many cases when strong deviations from Gaussianity occurs (e.g., when a small number of trials is averaged), averaging based on arithmetic mean can be not sufficient and trimmed estimators become a reasonable choice. Additional opportunities to improve averaging are given by weighting of the amplitude values from different epochs according to their rank, which is provided by TL-mean and by tanh mean proposed in this paper. Note that this type of weighting inside the averaging procedure is different from such procedure in usual weighted averaging, which utilizes epoch (trial) characteristics rather than the amplitudes only at the given time point; this is important for efficient processing of highly nonstationary data. Trimming can be also understood in terms of weighted averaging, as giving zero weight to extreme values (Figure 1 gives an idea of how this viewpoint can be applied to all estimators studied in this paper, including median and mean.). Because the trimming itself is evidently a rather rough procedure, more advanced weighting, as in TL-mean and tanh mean, can probably provide additional improvements of the results of averaging. As our current results show, averaging using trimmed estimators may

provide much smoother ERP waveforms than median averaging and at least in some cases provide better SNR of ERP than both median and arithmetic mean. However, no clear difference between trimmed mean, Winsorized and TL-mean was found in the cases of the best choice of trimmed parameters for each of them.

Optimization of the choice of the specific location estimator and its trimming parameters or any other parameters can be done for each specific data set. This procedure requires, of course, additional efforts and expertise. However, if an increase of SNR is strongly desirable, it can be worthwhile at least to compare the results of averaging with, for example, usual arithmetic mean, trimmed mean (e.g., with trimming of 25% of data samples from each tail of the distribution) and TL-mean (with a small trimming parameter). Note that the trimming parameter in TL-mean and k and s for the tanh mean influence not only trimming but also the weights of non-trimmed samples (see (5), (4) and Figure 1). This fact explains why TL-mean's dependence on trimming parameter is quite different from such dependence for trimmed and Winsorized mean (Figure 3). More precise optimization, which is especially important in the case of tanh mean, can be done with Nelder–Mead (Simplex) method of unconstrained nonlinear optimization (Nelder and Mead, 1965). The parameter to be optimized should be chosen carefully according to the objectives of the specific study (note that different approaches to estimation of SNR in ERP exist; see, e.g., Hoke et al. (1984); Coppola et al. (1978); Davila and Mobin (1992); Özdamar and Delgado (1996)). For unbiased optimization, a separate set of data with characteristics similar to the analyzed data should be used.

We considered in this paper only symmetrically trimmed estimators, because

they are appropriate for amplitude distributions without high asymmetry, and amplitude distributions of non–averaged ERP data typically are not very asymmetrical. Asymmetric trimmed mean estimators allowing different proportion of trimming at lower and higher tails of the distribution (e.g., (Lee, 2003)) probably can be also applied for averaging of ERPs, especially for small samples where the asymmetry of the distribution may be high.

Trimmed estimators are a class of robust estimators of data locations which can help to improve averaging of ERPs when number of trials is small, the data are highly nonstationary and include outliers. Their advantages can be understood as a reasonable compromise between median which is very robust but discard too much information and arithmetic mean conventionally used for averaging which use all data but, due of this, is sensitive to outliers. Additional improvement of averaging can be gained by introducing weighting of ordered data, as in newly introduced TL-mean and tanh mean proposed in this paper.

6 Acknowledgement

The authors would like to thank Dr. Pando Gr. Georgiev for his valuable suggestions about the tanh averaging.

References

Bock RK. Simplex Method: Nelder–Mead. CERN European Organization for Nuclear Research, Geneva, 1998:online, http://abbaneo.home.cern.ch/rkb/AN16pp/node262.html

Borda RP, Frost JD Jr. Error reduction in small sample averaging through

the use of the median rather than the mean. Electroencephalogr Clin Neurophysiol. 1968;25(4):391–2.

Coppola R, Tabor R, Buchsbaum MS. Signal to noise ratio and response variability measurements in single trial evoked potentials. Electroencephalogr Clin Neurophysiol. 1978;44(2):214–22.

Davila CE, Mobin MS. Weighted averaging of evoked potentials. IEEE Trans Biomed Eng. 1992;39(4):338–45.

Elamir EAH, Seheult AH. Trimmed L-moments. Comput Stat Data An 2003;43:299–314.

Fox LG, Dalebout SD. Use of the median method to enhance detection of the mismatch negativity in the responses of individual listeners. J Am Acad Audiol 2002;13(2):83–92.

Gastwirth JL, Cohen ML. Small Sample Behavior of Some Robust Linear Estimators of Location. J Am Stat Assoc 1970;65(30):946–73.

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel W. Robust Statistics: The Approach based on Influence Functions. Wiley: Toronto, 1986.

Hoke M, Ross B, Wickesberg R, Lütkenhöner B. Weighted averaging – theory and application to electric response audiometry. Electroencephalogr Clin Neurophysiol. 1984;57(5):484–9.

Hosking JRM. L–moments: Analysis and Estimation of Distributions using Linear Combinations of Order Statistics. J Royal Stat Soc B 1990;52(1):105–24.

Jansen BH, Agarwal G, Hegde A, Boutros NN. Phase synchronization of the ongoing EEG and auditory EP generation. Clin Neurophysiol. 2003;114(1):79–85.

Lee JY. Low bias mean estimator for symmetric and asymmetric unimodal distributions. Comput Methods Programs Biomed 2003;72(2):99–107.

Łęski JM. Robust weighted averaging. IEEE Trans Biomed Eng 2002 49(8):796–804.

Lopes da Silva F. Event–related potentials: Methodology and quantification. In Niedermeyer E, Lopes da Silva F, editors. Electroencephalography: Basic Principles, Clinical Applications, and Related Fields. Williams and Wilkins: Baltimore, 1999:947–57.

Lütkenhöner B, Hoke M, Pantev C. Possibilities and limitations of weighted averaging. Biol Cybern. 1985;52(6):409–16.

Makeig S, Westerfield M, Jung TP, Enghoff S, Townsend J, Courchesne E, Sejnowski TJ. Dynamic brain sources of visual evoked responses. Science 2002;295(5555):690–4.

Nelder JA, Mead R. A Simplex Method for Function Minimization. Comput J 1965;7:308–13.

Özdamar Ö, Delgado RE. Measurement of signal and noise characteristics in ongoing auditory brainstem response averaging. Ann Biomed Eng. 1996;24(6):702–15.

Özdamar Ö, Kalayci T. Median averaging of auditory brain stem responses. Ear Hearing 1999;20:253–64.

Picton TW, Lins OG, Scherg M. The recording and analysis of event–related potentials. In F. Boller F, Grafman J, editors. Handbook of Neuropsychology, Elsevier: Amsterdam, 1995;10(14):3–73.

Picton TW, Bentin S, Berg P, Donchin E, Hillyard SA, Johnson R Jr, Miller GA, Ritter W, Ruchkin DS, Rugg MD, Taylor MJ. Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. Psychophysiology 2000;37(2):127–52.

Rice JR. Mathematical Statistics and Data Analysis. Duxbury Press: Belmont, 1995.

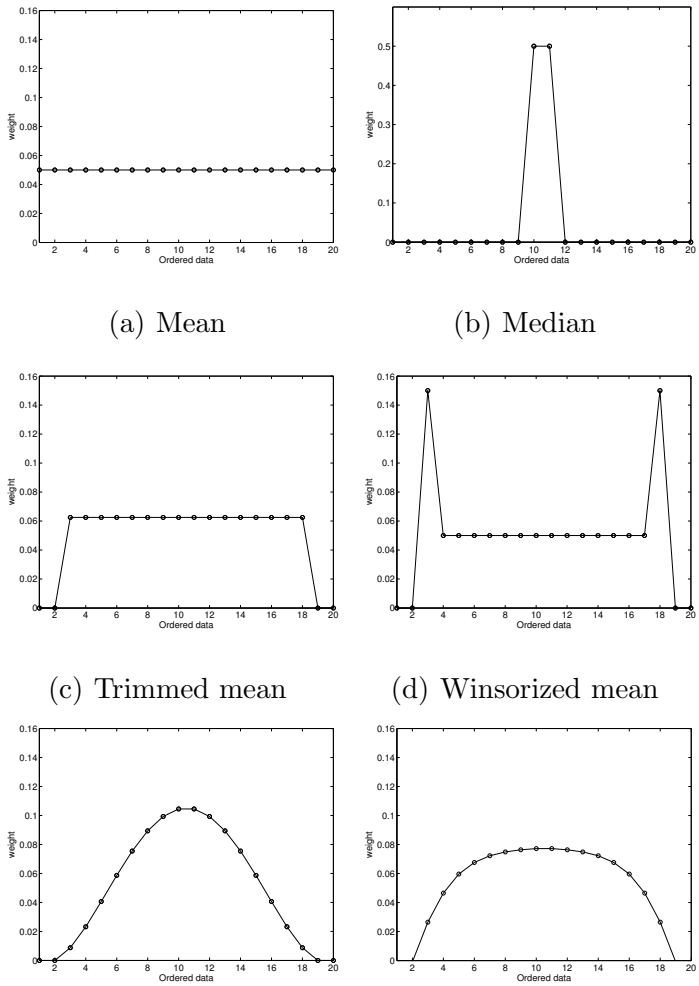
Streiner DL. Do you see what I mean? Indices of central tendency. Can J Psychiatry 2000 45(9):833–6.

Stuart A, Keith Ord J. Kendall's Advanced Theory of Statistics. Edward Arnold: London, 1994.

Tukey JW: A survey of sampling from contaminated distributions. In Olkin I, editor. Contributions to Probability and Statistics. Stanford University Press:Stanford, 1960:448–485.

Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain-computer interfaces for communication and control. Clin Neurophysiol 2002;113:767–91.

Yabe H, Saito F, Fukushima Y. Median method for detecting endogenous event-related brain potentials. Electroencephalogr Clin Neurophysiol. 1993;87(6):403–7.

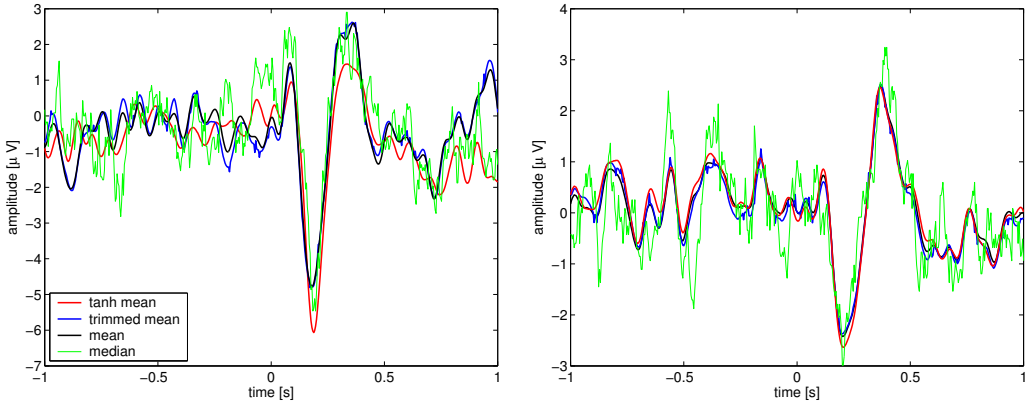


(a) Mean (b) Median

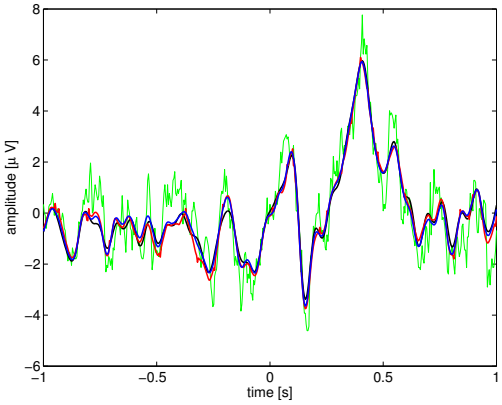
(c) Trimmed mean (d) Winsorized mean

(e) TL-mean (f) Tanh mean

Fig. 1. Weights for the different location estimators. Data ($N = 20$) are sorted in ascending order. The arithmetic mean uses the same weight $1/N$ for all observations. The median uses only the middle observation (odd number of observations) or the two middle observations (even number of observations). Note that the scale is different in figure (b). The trimmed mean applies a zero weight for extreme observations and an equal weight for all other observations. The Winsorized mean concentrates the weights of the ignored extreme observations to the last accepted data points. The TL-mean applies higher weights for the middle observations, while the new **tanh** mean applies smoothly changing weights to the values close to extreme. The trimming parameters for the trimmed mean and the Winsorized mean are set in this example to ignore two minimum and two maximum values. Trimming parameter in TL-mean and parameters of **tanh** mean are set to ignore one minimum and one maximum values.

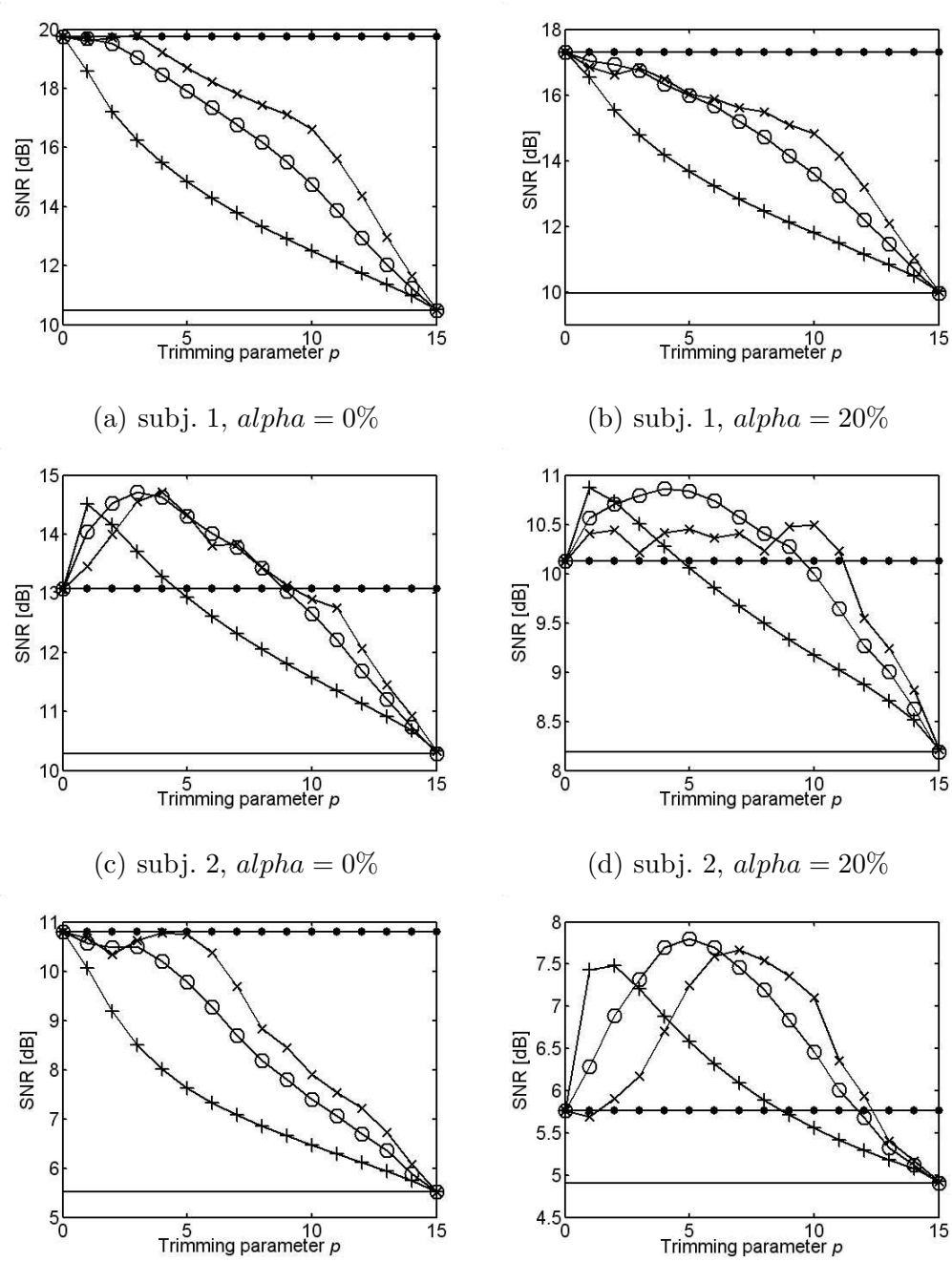


(a) subject 1 (b) subject 2



(c) subject 3

Fig. 2. Comparison of the averaged ERPs (100 trials, raw data) using the mean averaging (black line), median averaging (green line), trimmed mean averaging ($p = 25$, blue line) and **tanh** averaging ($k = 0.1$, $s = 0$, red line). The stimulus is presented at zero time point. Median averaging enhanced strongly the noise. Some high frequency low amplitude noise appeared also in trimmed mean average. Conventional mean and **tanh** mean provided most smooth waveforms while **tanh** mean enhanced N1 peak comparing to mean averaging.



(a) subj. 1, $alpha = 0\%$

(b) subj. 1, $alpha = 20\%$

(c) subj. 2, $alpha = 0\%$

(d) subj. 2, $alpha = 20\%$

(e) subj. 3, $alpha = 0\%$

(f) subj. 3, $alpha = 20\%$

Fig. 3. Comparison of SNR of the averaged ERPs using different averaging methods (mean (dots), median (solid line), trimmed mean (circles), Winsorized mean (X's) and TL–mean (crosses)) and different values of trimming parameter p. SNR was computed 100 times for averaged independent subsets of 31 epochs randomly drawn from "cleaned" data set. Horizontal lines shows the values obtained for mean and median (correspond to lowest and highest possible p). *alpha* parameter shows the percentage of epochs with artificially added "alpha rhythm" (see text for details).

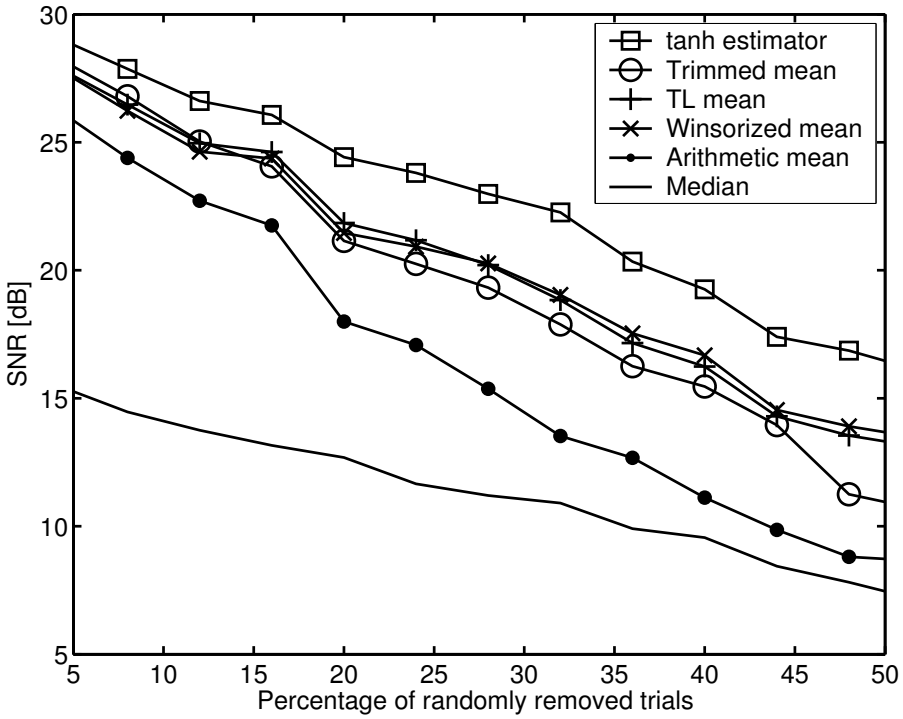


Fig. 4. Small sample performance of the estimators. SNR was estimated for averaged 100 epochs which were randomly chosen from subject 1 raw data set and for averaged epochs remained after randomly removing increasing number of epochs. Parameters of trimmed estimators were optimized using independent sets of the same number of randomly chosen epochs (see details in the text). SNR was averaged for 100 repetitions of the procedure. Note that in the case of **tanh** mean 2 parameters were optimized instead of 1 parameter in three other trimmed estimators, while mean and median have no parameters which could be optimized; this could at least partly contribute to SNR variations amongst the estimators.