Chaos and Complexity Letters
Volume 2, Number 2/3, pp. 259–274
ISSN 1555-3995

ROBUSTIFYING EEG DATA ANALYSIS BY REMOVING OUTLIERS

Matthias Krauledat[1,2]*; ***Guido Dornhege***[2]†;*
Benjamin Blankertz[2]‡ ***and Klaus-Robert Müller***[1,2]§
[1]Technical University Berlin, Str. des 17. Juni 135, 10623 Berlin, Germany
[2]Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany

Abstract

Biomedical signals such as EEG are typically contaminated by measurement artifacts, outliers and non-standard noise sources. We propose to use techniques from robust statistics and machine learning to reduce the influence of such distortions. Two showcase application scenarios are studied: (a) Lateralized Readiness Potential (LRP) analysis, where we show that a robust treatment of the EEG allows to reduce the necessary number of trials for averaging and the detrimental influence of e.g. ocular artifacts and (b) single trial classification in the context of Brain Computer Interfacing, where outlier removal procedures can strongly enhance the classification performance.

1. Introduction

Identifying outlier points in a dataset can enhance our understanding of the data. By removing outliers, it is possible to improve the estimation of intrinsic properties such as mean or covariance matrix, and to analyze the data in single trial analysis. Various definitions of the outlier concept have been suggested, e.g. [1, 2, 3, 4, 5, 6, 7]. We will in the following introduce some model assumptions about the EEG data, and by outliers simply refer to those points not fulfilling these assumptions. We will show how this concept can be used to robustify the analysis of motor-related EEG data. This type of data is often subject to examination by Brain-Computer-Interface research, which aims to allow direct control of, e.g., a computer application or a neuroprosthesis, by human intentions that are reflected by suitable brain signals (e.g. [8]).

An effective discriminability of different brain states used in a BCI paradigm is an important neurophysiological prerequisite to implement a suitable system. Furthermore appropriate features have to be chosen by signal processing techniques such that they can effectively be translated into a control signal, either by simple threshold criteria (cf. [8]), or

*E-mail address: kraulem@first.fhg.de
†E-mail address: dornhege@first.fhg.de
‡E-mail address: blanker@first.fhg.de
§E-mail address: klaus@first.fhg.de

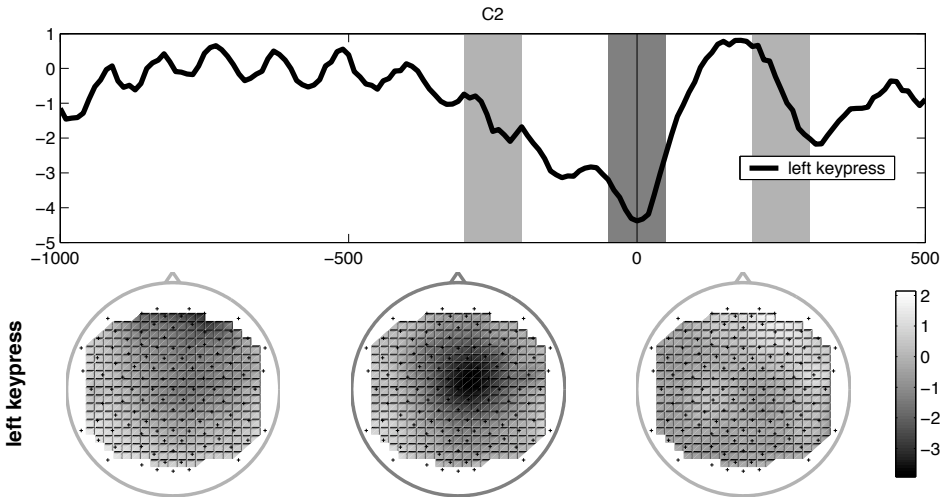

Figure 1. This figure shows the Lateralized Readiness Potential during a finger movement for one subject. The timecourse for electrode C2, averaged over more than 500 trials, is shown above; the spatial distribution corresponding to the timepoints in the grey shaded areas is visible from the three scalp plots below.

by machine learning techniques where the computer learns a decision function from some training data [9, 10, 11, 12].

Typically EEG signals are distorted by artifacts and noise; they are furthermore subject to nonstationarity. If the few training samples that are measured within the 'training' time are contaminated by such artifacts, a sub-optimal or even highly distorted classifier can be the consequence [13]. Since simple linear classifiers like Linear Discriminant Analysis (LDA), Regularized Discriminant Analysis (RDA) or Quadratic Discriminant Analysis (QDA) assume Gaussian distributions of the classes in feature space, every deviation from this assumption can result in poor performance of the discrimination method. We will show that outliers can transform the data to a non-gaussian distribution. Therefore it is important to strive for robust machine learning and signal processing methods that are as immune as possible against such distortions.

2. Robustification Approaches for EEG Data

The literature points out various methods of how to identify outliers [1, 2, 3, 4, 5, 6, 7]. In Section 3., we will use the delta-method ([14], see Section 6. for a short introduction) to identify outliers. This method does not rely on the estimation of parameters such as mean or covariance matrix of the data in feature space, but rather uses the relative distances of each data point to its k nearest neighbors. In Section 4., we will use the Mahalanobis distance [1, 15], which requires to estimate both mean and covariance matrix of the data sample to find points with the largest deviance from the class mean. Points with high distances to all others are really different from the usual data ensemble and should therefore not be considered representative. Furthermore a decision has to be made on how many trials

should be removed based on the outlierness curve. Our tests to automatize the cut point in this curve did not result in significant changes. Thus, for the purpose of this paper we present only results where the [10]%-worst trials were removed.

Apart from the general issue of choosing an outlier detection method, it is also an inherent problem of EEG data that the dimensions of the feature space may have different qualities: usually, data points are given with a certain number of repetitions (trials), and they contain channel informations and the temporal evolution of the signal. A natural approach is to specifically use this information to find outliers within a certain dimension, i.e., removing channels with an increased noise level (due to high impedances at the specific electrode) or removing trials which are contaminated by artifacts from muscular or ocular activity. These approaches will be explained in detail in ection 4..

3. Outliers in LRP-Features

This part of the paper will serve as an introduction on the nature of outliers in neurophysiological data. We will demonstrate exemplarily how outliers can disrupt the estimation of the distribution of certain features of the EEG, which suggests that removing those outlier trials can lead to a more robust estimation of the original LRP signal.

3.1. Experimental Setup

We recorded EEG data in 34 experiments from 17 different subjects who were sitting in a comfortable chair in front of a computer monitor. Brain activity was recorded from the scalp with multi-channel EEG amplifiers using 32–128 channels, at a sampling rate of 1000 Hz. The subjects pressed buttons of a keyboard with their index fingers in a (selfpaced) rhythm of approximately 0.5 Hz, in a selfchosen, random order. Each experiment consisted of 500–1000 repetitions of these movements (“trials”). The data were then stored for training classifiers for online BCI feedback experiments. In the course of these experiments, a cross-shaped cursor was presented to the subjects on the screen, indicating the estimated laterality of the keypress. The results obtained during training and feedback experiments are presented in previous publications, [16, 17, 18]. We will now use the same feature extraction as it was applied for classification purposes in order to demonstrate qualitative differences between in- and outlier trials.

3.2. Neurophysiological Background

According to the model known as homunculus, for each part of the human body there is a respective region in the motor and somatosensory area of the neocortex. The ’mapping’ from the body to the respective brain areas preserves topography, i.e., neighboring parts of the body are represented in neighboring parts of the cortex. While the region of the feet is at the center of the vertex, the left hand is represented lateralized on the right hemisphere and the right hand on the left hemisphere. In preparation of motor tasks, a slow negative shift can be observed in the EEG. Analyzing multi-channel EEG recordings, it has been shown that several brain areas contribute to this shift ([19, 20]). In finger movements, the

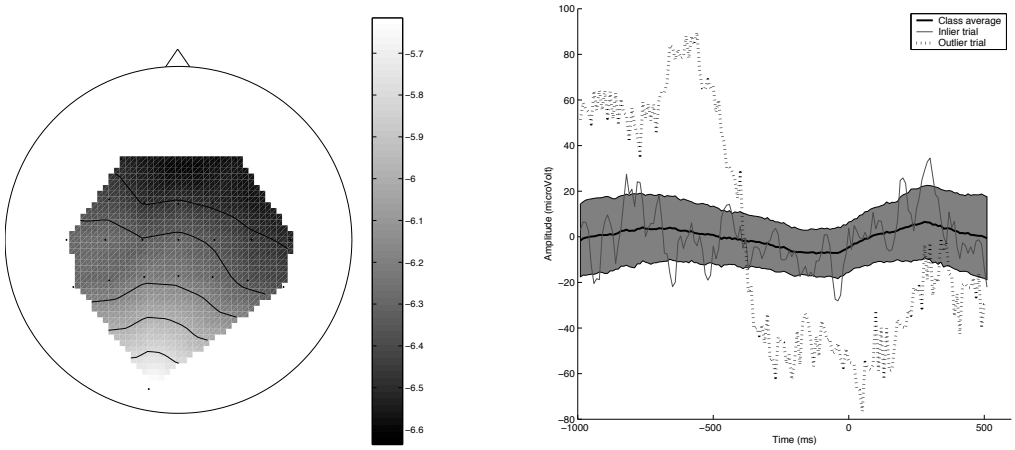

Figure 2. In the left part of this figure, the differences between outlier and inlier trials are presented in terms of the Wilcoxon ranking score, averaged over data from 17 subjects (see text for details). The right part shows the EEG signal of one subject at electrode C2, averaged over more than 500 trials of repeated left index finger keypresses. One trial that has been identified as an outlier trial and a typical inlier trial are shown in the same plot. The gray area depicts the standard deviation of the inlier trials.

focus of this shift is in the frontal lobe of the corresponding motor cortex, i.e., contralateral to the performing hand (see figure 1). It is possible to classify the laterality of an upcoming hand movement with high accuracy based on the spatial distribution of this EEG signal, up to 120 ms prior to the actual execution of the movement, see [16, 17, 18].

3.3. Feature Extraction

First, we select up to 20 central channels that cover the areas corresponding to the motor cortices of the fingers. The data are then bandpass-filtered to 0.8–3 Hz, and the last 150 ms preceding the keypress are subsampled to 20 Hz, such that only three samples per channel remain. The samples are then concatenated over all channels. These steps are explained in detail in [16].

3.4. Outlier Identification

According to the delta-score (see appendix; a more detailed version is given in [14]) obtained by each trial, we label those 10% of the trials with the highest scores as outliers. Figure 2 shows the difference in the power between outlier- and inlier-trials in terms of the w-scores w_{ch} of the average bandpower $\mathrm{fv}_{\mathrm{ch}}$ in the frequency band from 0.8 to 5 Hz. The w-score is used in the Wilcoxon test for the comparison of two random samples for equal distribution. It is computed in the following way:

$$w_{\mathrm{ch}} = \frac{R_{\mathrm{ch,in}} - \frac{n_{\mathrm{in}}(n_{\mathrm{in}}+n_{\mathrm{out}}+1)}{2}}{\sqrt{\frac{n_{\mathrm{in}}n_{\mathrm{out}}(n_{\mathrm{in}}+n_{\mathrm{out}}+1)}{12}}},$$

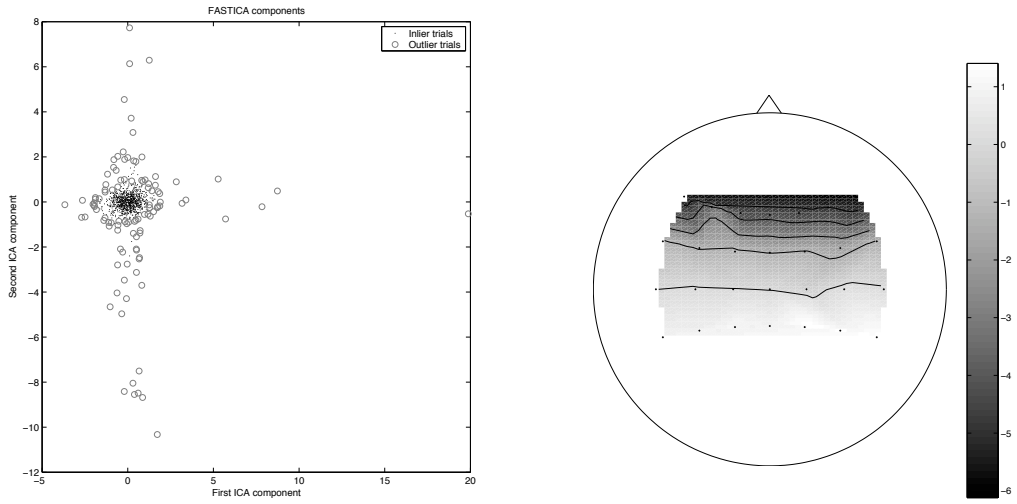

Figure 3. The left part of this figure shows a scatterplot of two normalizations of linear projections of the feature space in one subject. The cross-shape of this plot reveals a non-gaussian structure of the data. The grey circles mark the trials which are identified as outliers. In the right plot, one of the corresponding projection matrices is shown. The spatial distribution suggests that the distribution of this projection is caused by eye movements.

where $n_{\text{in}}, n_{\text{out}}$ are the respective numbers of in- and outliers, and

$$R_{\text{ch,in}} = \sum_{i=1}^{n_{\text{in}}} R(\text{fv}_{\text{ch},i})$$

is the sum of the ranks of all inlier trials in the combined sample of in- and outlier trials. A low w-value indicates that the variance of the outlier trials in this channel is higher than the variance of the inlier trials. The figure shows the spatial distribution of these differences after averaging over all subjects. Since the w-values of all channels are negative, the trials that have been identified by the outlier method have higher variances in this frequency band. By the spatial distribution, it is also apparent that this variance is caused by eye movements, since the influence of eye movements is maximal in the electrodes near the eyes and falls off with increasing distance, see e.g. [21]. In the right part of figure 2, the timecourse of the trials with lowest and highest delta-score (i.e., of an in- and an outlier) at electrode C2 are shown for one subject. This also illustrates the high variance of the outlier trials.

Figure 3 shows a two-dimensional linear projection of the feature space with the most "non-gaussian" components. These projections are found by applying Independent Component Analysis to the feature space for one subject. It has been shown in [16] that this preprocessing converts the data into a feature space where it is safe to assume gaussian distributions for the data. Under this assumption, every projection of the feature space should be normally distributed again, but this figure shows that there is in fact a strong "non-gaussianity" due to the outliers. The gray circles indicate the trials which the delta-method would identify as outliers. After the removal of 10% outlier trials, the projections are no longer significantly different from normal distributions.

3.5. Results

In this section we have illustrated that eye movements are a common source of deteriorating influences on the EEG signal when dealing with slow cortical potentials. In our experiments, there is a significant correlation between eye movements and the identification of the trials as outliers. Note that these trials may also be removed from the data ensemble by simple eye artifact-rejection; however, this rejection method assumes that only the eyes are sources of signal deterioration, while outlier detection methods also capture other types of influences, such as muscular activity or movement artifacts.

It has also been shown that outliers in EEG recordings can deteriorate the data in such a way that basic assumptions about the underlying distribution, e.g. gaussianity, are not met and hence a robust estimation of the parameters can not be guaranteed. Removing outlier trials from the recording can help to remove this detrimental effect of the outliers.

4. Outliers in Bandpower Features

So far, possible effects of outlier identification and outlier removal have been demonstrated only by their effect on the distribution in the EEG feature space. Now we will quantify this effect by applying the presented methods in a single trial classification context with bandpower features.

4.1. Experimental Setup

In this section we investigate data from 22 EEG experiments with 8 different subjects. All experiments included so called training sessions in which the subjects performed mental motor imagery tasks according to visual stimuli. In such a way samples of recorded EEG data are obtained that reflect brain activity during the involved mental tasks. These can be used to train a classifier by machine learning techniques which can be applied in further sessions to produce a feedback signal from (unlabelled) continuous brain activity. Our earlier recordings were only done for investigational purpose, while later experiments included online feedback sessions in which the users could control some simple computer games like *brain pong* or steer a cursor.

All 5.5 (± 0.25) seconds one of three different visual stimuli indicated for 3.5 seconds which mental task the subject should accomplish during that period. The investigated mental tasks were imagined movements of the left hand (l), the right hand (r), and the right foot (f). Besides EEG channels, we recorded the electromyogram (EMG) from both forearms and the right leg as well as horizontal and vertical electrooculogram (EOG) from the eyes. The EMG and EOG channels were exclusively used to make sure that the subjects performed no real limb or eye movements correlated with the mental tasks that could directly (artifacts) or indirectly (afferent signals from muscles and joint receptors) be reflected in the EEG channels and thus be detected by the classifier, which operates on the EEG signals only. Two experiments were only done with the 2 classes l and r. Between 120 and 200 trials for each class were recorded. In this study we investigate only binary classifications, but the results can be expected to safely transfer to the multi-class case, [22, 23].

4.2. Neurophysiological Background

One feature of brain activity, which can be exploited for brain-computer interfacing relies on the following neurophysiological observation: when a subject is not engaged with one of his limbs (movements, tactile senses, or just mental concentration), large populations of neurons in the respective cortex fire in rhythmical synchrony, which can be measured at the scalp in the EEG as a brain rhythm around [10]Hz (μ-) or [20]Hz (β-rhythm). These are so-called idle rhythms that are attenuated when engagement with the respective limb takes place. As this effect is due to loss of synchrony in the neural populations, it is termed event-related desynchronization (ERD), see [24]. The dual effect is called event-related synchronization (ERS). Since the ERD in the motor and/or sensory cortex can be observed even when a subject is only thinking of a movement or imagining a sensation in the specific limb, this feature can well be used for BCI control. The discrimination of the imagination of movements of left hand vs. right hand vs. foot is based on the topography of the attenuation of the μ and/or β rhythm.

The strength of the sensorimotor idle rhythms as measured by scalp EEG is known to vary strongly between subjects. This introduces a high intersubject variability on the accuracy with which an ERD-based BCI system works. There is another feature reflecting imagined or intended movements, the movement related potentials (MRP), denoting a negative DC shift of the EEG signals in the respective cortical regions. This feature can be exploited for BCI use, both as a single feature and in combination with the ERD features. See [22, 23] for an investigation of how this combination strategy was able to greatly enhance classification performance in offline studies. In this paper we focus only on enhancing the ERD-based classification, but all the improvements presented here can as well be used in the combined algorithm.

4.3. The CSP Algorithm

The common spatial pattern (CSP) algorithm is very useful in calculating spatial filters for detecting ERD/ERS effects ([25]) and for ERD-based BCIs ([26]), and has been extended to multi-class problems in [22]. Given two distributions in a high-dimensional space, the (supervised) CSP algorithm finds directions (i.e., spatial filters) that maximize variance for one class and at the same time minimize variance for the other class. After having band-pass filtered the EEG signals to the rhythms of interest, high variance reflects a strong rhythm and low variance a weak (or attenuated) rhythm. Let us take the example of discriminating left hand vs. right hand imagery. According to Section 4.2., the spatial filter that focusses on the area of the left hand is characterized by a strong motor rhythm during imagination of right hand movements (left hand is in idle state), and by an attenuated motor rhythm during left hand imagination. This criterion is exactly what the CSP algorithm optimizes: maximizing variance for the class of right hand trials and at the same time minimizing variance for left hand trials. Furthermore the CSP algorithm calculates the dual filter that will focus on the area of the right hand (and it will even calculate several filters for both optimizations by considering orthogonal subspaces).

Let Σ_i be the covariance matrix of the trial-concatenated matrix of dimension [channels $\times$ concatenated time-points] belonging to the respective class $i \in \{1,2\}$. The CSP analysis

consists in calculating a matrix Q and diagonal matrix D with elements in $[0,1]$ such that

$$Q\Sigma_1 Q^\top = D \qquad \text{and} \qquad Q\Sigma_2 Q^\top = I - D. \tag{1}$$

This can be accomplished in the following way. First we *whiten* the matrix $\Sigma_1 + \Sigma_2$, i.e., determine a matrix P such that $P(\Sigma_1 + \Sigma_2)P^\top = I$ which is possible due to positive definiteness of $\Sigma_1 + \Sigma_2$. Then define $\hat{\Sigma}_i = P\Sigma_i P^\top$ and calculate an orthogonal matrix R and a diagonal matrix D by spectral theory such that $\hat{\Sigma}_1 = RDR^\top$. Therefore $\hat{\Sigma}_2 = R(I-D)R^\top$ since $\hat{\Sigma}_1 + \hat{\Sigma}_2 = I$ and $Q := R^\top P$ satisfies (1). The projection that is given by the i-th row of matrix R has a relative variance of d_i (i-th element of D) for trials of class 1 and relative variance $1 - d_i$ for trials of class 2. If d_i is near 1 the filter given by the i-th row of R maximizes variance for class 1, and since $1 - d_i$ is near 0, minimizes variance for class 2. Typically one would retain some projections corresponding to the highest eigenvalues d_i, i.e., CSPs for class 1, and some corresponding to the lowest eigenvalues, i.e., CSPs for class 2.

Now let us consider the issue of robustness. In single-trial classification of multi-channel EEG signals, one is usually confronted with a bad sample to dimension ratio. Having 100 EEG trials per class of 300 data points (3 seconds at [100]Hz) in 120 channels each gives a ratio of 100 samples to 36000 dimensions in feature space when trying to classify raw EEG trials. This makes the estimation of a covariance matrix (size 36000×36000) of the class distributions in feature space from the 100 samples really hard. Given the bad signal to noise ratio in most EEG classification tasks even careful regularization is a hard task. So feature dimensions have to be reduced first. The CSP algorithm is a supervised method that does this job, but is it robust? At least for the covariance matrices which have to be estimated for calculating the CSPs the situation is much more favorable. Here only spatial covariance matrices are considered, i.e., of size 120×120 in our example, which are calculated from the concatenated trials, i.e., $100 \cdot 300 = 30000$ samples. This ratio (dimension vs. number of samples) is sufficient which is the reason why in the CSP algorithm usually there is no need for regularization (shrinking the estimated covariance matrices towards a sphere).

This consideration shows that the estimation of the covariance matrices in the CSP algorithm is quite robust. Still there are a number of factors that could degrade the performance of the CSP method: (1) outlier trials where the subject either produces artifacts or does not perform the required mental task, (2) unreliable channels, that are partly noisy due to measurement problems. In this paper we investigate two methods that would compensate for (1) in different ways and one method that tries to compensate for (2). Our expectation in this study was that robustifying methods could only improve performance in few experiments because we had well controlled EEG measurements on subjects that were highly motivated for the experiments such that they would canonically try to avoid to produce artifacts.

4.4. Feature Extraction, Classification and Validation

4.4.1. Feature Extraction

There are several parameters in this feature extraction procedure that should be specifically chosen for each subject to obtain optimal results. In our online experiments this is done

semiautomatically by combining machine learning, expert knowledge and visual inspection of some characteristic curves such as spectra and ERD curves, see [27]. In this comparative offline analysis absolute performance does not matter, so we have chosen one fixed setup for all subjects.

After choosing all channels except the EOG and EMG and a few outmost channels of the cap we apply a causal band-pass filter from [7–30]Hz to the data, which encompasses both the μ- and the β-rhythm. The trials we extract are the windows [750–3500]ms after the presented visual stimulus, since in this period discriminative brain patterns are present in most subjects. Afterwards we apply the CSP algorithm (see Section 4.3.) to the data which decreases the number of channels by suitable linear spatial filters which are learned on the training trials. Here we use 3 patterns per class which leads to 6 remaining channels. As a measure of the amplitude in the specified frequency band we calculate the logarithm of the variances of the remaining channels as feature vectors.

4.4.2. Classification

After the presented preprocessing usually between 120 and 200 six-dimensional feature vectors for each class remain. Although we have tested exemplarily non-linear classification methods on these features, so far no significant gain could be observed compared to Linear Discriminant Analysis (LDA), which was also reported in [13, 9, 16]. Therefore we choose LDA as the classifier.

4.4.3. Validation

To explore the performance of an algorithm we apply a 10×10-fold cross-validation to the feature vectors. This means that we randomly split the data set into ten equal parts, use each once as a test set while training on the other 90 percent, and repeat this procedure ten times to get 100 test errors.

Since the CSP algorithm and other techniques presented later on exploit label information, these techniques have to be used only on the training set within the cross-validation procedure. Otherwise the cross-validation error could underestimate the generalization error.

To maintain comparability between algorithms, we keep track of the chosen divisions into training and test sets and apply all algorithms to the same divisions.

4.5. Outlier Removal

4.5.1. Channel Removal

Instead of calculating the covariances, the evaluation of the correlation coefficients gives the opportunity to estimate the certainty for each channel. Here we take the difference of the lower bound and upper bound of the [95]% confidence interval for the estimation of the correlation coefficients. Using this as a measure of the goodness resp. badness, unreliable channels can be removed by a simple threshold criterion.

4.5.2. Outlier-Trial Removal

As a simple and reliable approach we will show here only one way, which performs reasonably well in our studies. The development of improved outlier removal methods is process of ongoing research. For the validation of the presented algorithms outliers were only removed considering the training set, but for the test set all trials without recognizing their outlierness were used. However, the information that a trial is an outlier might also be used in feedback situations, e.g. by freezing the cursor instead of providing the regular feedback. This option would greatly enhance the range of possible application, but since we are here only dealing with training data, we forgo this option.

The presented outlier removal approach is based on the idea to use the Mahalanobis distance of the variance of each trial and channel as measurement of the outlierness of the trials (cf. [1, 15]).

4.5.3. Robustification by Normalization

For the robust estimation of covariance matrices, many different algorithms have been proposed. Other feasible variants include approximating covariances via 1-norm, median absolute deviation (MAD) or using the least informative distribution approach (cf. [2]).

The method we are going to present in this category is to normalize each time point in the filtered EEG signal to have euclidean norm 1 over the channels. With this modified signal we estimate the covariances and the CSPs and apply them to the normalized data which has been processed as before. Different strategies like applying this spatial filter to the original filtered but unnormalized data or normalizing the whole window trialwise results in similar performance. Normalizing the EEG data in this way deletes the absolute amplitude of the signals and retains only the relative amplitudes in their spatial configuration. This is enough information to detect ERD features, and additionally has the effect that outliers have less influence in estimating covariances (of the normalized signals).

4.6. Results

As reported in earlier publications ([28, 22, 23]), one can see that the usual CSP algorithm often performs quite well. Nevertheless there are some experiments in which one or more of the robustification approaches can greatly improve classification. Unfortunately the same new methods can also deteriorate the results in other instances. This means that for the application in BCI feedback experiments, a meta-decision about the robustification method has to be taken, based on the data of the training session for each subject. For the validation of such a procedure on our offline data, we applied two schemes, which used different partitions of each data set.

In the *chron* approach, we have split the data into the (chronological) first and second half of the data. On the first half we calculated the cross-validation error for each of the competing algorithms as described in Section 4.4.3.. We will call the results here the "expected performance" or "expected error" of the algorithm. Based on the expected performance, we now decide on the algorithm which is to be used for the test session. For this decision, we calculate the difference between the expected error of our baseline CSP

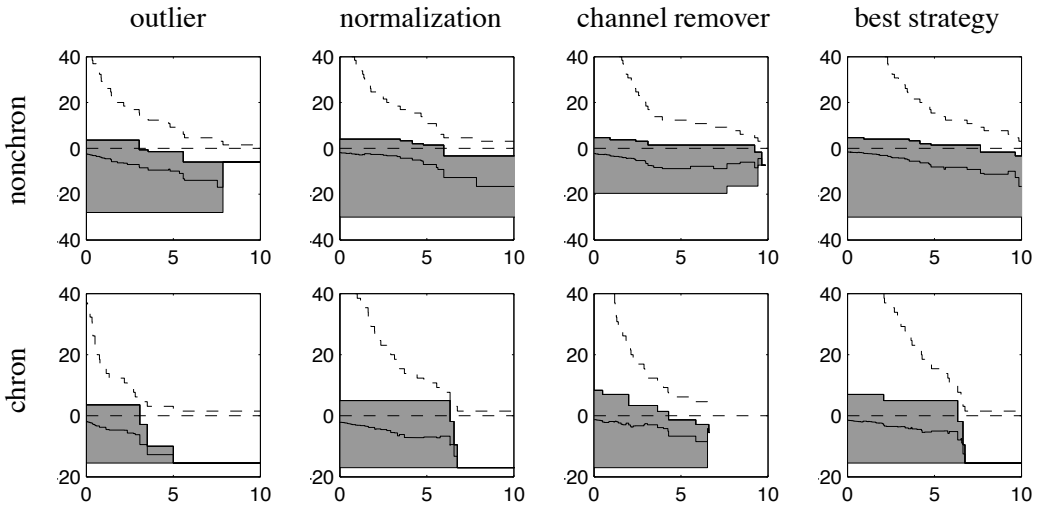

Figure 4. We vary the decision threshold between [0]% and [10]% on the x-axis. Out of the experiments where one of the robustification algorithms increases the expected performance of the CSP by at least the threshold, the mean of the test error gain on the chosen algorithm against CSP is plotted as a black line. The range of all these values is visualized by the gray shaded area. Below zero the change to the robustified method was successful: the lower the solid line, the higher the improvement. The dashed line shows the portion of experiments where the robustified method was chosen. In the first three columns each single robustification approach is compared to CSP whereas in the last column the best of all three robustified methods was used respectively. *chron* (for chronological order) denotes an evaluation mode where the expected error is estimated by cross-validation on the first half of the data and the test error is determined on the second half; the *nonchron* mode splits the data into even and odd trials.

approach and the expected error of each of the algorithms presented in Section 4.5.. Only if this difference exceeds a certain switching threshold, we choose the alternative algorithm instead of the CSP approach for the evaluation of the test set. Once the decision is taken for one of the methods, we train the classifier on this first half and apply it to the other half of the data ("test performance"). This evaluation mode closely resembles an actual feedback situation; a fixed classifier is trained using only data from a preceding training session, and is applied to the following feedback data. Note, however, that this evaluation is prone to be affected by nonstationary behaviour of the EEG data, which is often encountered in this type of experiments.

The *nonchron* approach, the second evaluation method, is to a large extent invariant to these local changes in the EEG; here the training set consists of every even trial and the test set of every odd trial, such that slow trends are always present in both training and test data. The evaluation then proceeds as in the *chron* method.

In Figure 4 we have compared this test performance gain in different switching thresholds for each of the algorithms and for the best of all of them. Furthermore the percentage of experiments are shown where a switch to a robustification algorithm took place. Obviously, this portion decreases with increasing thresholds, i.e., if we choose a more conservative

strategy. On the other hand our mean performance gain increases (i.e., the classification test difference decreases) with increasing threshold, until only few or no false decisions are left. Nevertheless, there are very few experiments where our decision to change was wrong as seen in the figure, but the cases where a change improves the classification accuracy outweigh the others. Between the algorithms there is no substantial difference visible, but as their success lies in different experiments, further improvement by combination strategies can be expected.

In total the *chron* and *nonchron* evaluation strategy lead to similar interpretations. One important difference is that the gray area above the zero line is thinner in the *nonchron* case. That means that in the *chron* evaluation there are several cases in which the result of the chosen robustification method is worse than the baseline CSP result, while in the *nonchron* case there are less severe failures. This gives a clue for the reason of the failure: nonstationarity in the data. If all datapoints were drawn from the same distribution, then *nonchron* and *chron* evaluation should result in similar classification accuracies, but this finding shows that the distributions are undergoing changes throughout the time.

In the end, the figure shows that it can be profitable in some cases to switch to a suitable outlier algorithm for enhancing performance.

5. Concluding Discussion

EEG data recorded in motor-related tasks are highly challenging to evaluate due to noise, nonstationarity and diverse artifacts. Thus BCI provides an excellent testbed for testing the quality and applicability of robust machine learning methods (cf. [29]). In this paper we analyzed the effects that outlier trials may have on the distribution of the data in feature space. It was shown that eye movements are a common source for the outlierness of trials in slow cortical potential data; the result we encountered was a shift of the data cloud towards a non-gaussian distribution, where the removal of outliers may help to restore the model property of gaussianity that is assumed for linear classification. Finally, we showed how outlier removal methods can improve the classification accuracy in the discrimination between different motor actions.

As our BCI system has so far mainly relied on dimension reduction techniques like CSP, this paper has explored directions of their robustification, such as channel removing, outlier and normalization approaches. However in a BCI training protocol it is essential to decide whether to apply one of the robust alternatives or to stick with the conventional baseline algorithm, that obtains better results in some cases. As shown, this meta-decision, if exercised sufficiently conservatively, i.e., only after an expected gain of more than [5]%, can yield significant performance improvements. These encouraging results should nevertheless be carefully put into perspective: (i) we find no overall best robustification strategy and (ii) individualized choices need to be made for each subject. Furthermore we should note that the more conservative our strategy, the less likely it is to switch and also the less likely it is to have erroneously switched. Part of the reason, why the selected algorithm occasionally performs suboptimal is the intrinsic nonstationarity in a BCI experiment. Obviously BCI users are subject to variations in attention and motivation. But this kind of nonstationarity that deteriorates the BCI classifiers still has to be investigated and ways to

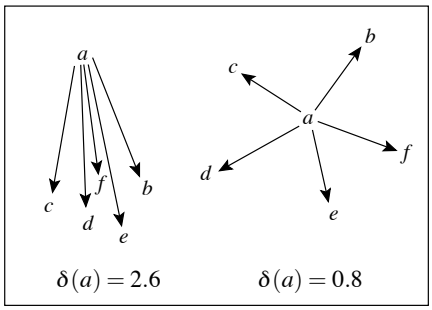

Figure 5. In the left example a is an outlier an thus its δ index is large. In the right example it is part of a larger group so its δ index is small. Both examples assume $k = 5$.

circumvent that problem have to be found [30].
As the generally positive outcome of this offline analysis shows, outlier removal for classificator training is a very useful technique for robustification of BCI classifiers; this finding is also supported by our experience from recent online experiments, [27].

Our conjecture from the above findings is that in order to further improve information transfer rates in BCI we will need to counter the effects of switching dynamics (i.e., nonstationarity in the feature space of the online feedback data) by moving towards online learning by designing algorithms that are adaptive throughout the BCI session.

6. Appendix

Consider n data points $\{x_1,\ldots,x_n\} \subset \Re^d$ in d-dimensional space with a norm, $||x|| = \sqrt{x^\top x}$. We denote the k nearest neighbors of $x \in \Re^d$ among the given set by

$$z_1(x),\ldots,z_k(x) \in \{x_1,\ldots,x_n\} \subset \Re^d.$$

The outlier index $\delta(x)$ is defined to be the length of the mean of the vectors pointing from x to its k nearest neighbors, i.e.,

$$\delta(x) = ||\frac{1}{k}\sum_{j=1}^{k}(x - z_j(x))||.$$

As shown in Figure 5, δ is large if the neighbors are all in the same direction, which is usually the case for outliers.

Acknowledgments

We thank S. Harmeling, C. Schäfer, M. Kawanabe, A. Ziehe and G. Rätsch for comments and helpful discussions. The studies were supported by the *Deutsche Forschungsgemeinschaft* (DFG), FOR 375/B1 and MU 987/1-1, by BMBF-grants FKZ 01IBB02A and FKZ 01IBB02B and by the PASCAL Network of Excellence (EU # 506778).

References

[1] V. Barnett and T. Lewis, *Outliers in Statistical Data*, Wiley, New York, 3rd edn., 1994.

[2] P. Huber, *Robust Statistics*, John Wiley and Sons, New York, 1981.

[3] F. R. Hampel, E. M. Rochetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics*, Wiley, 1986.

[4] G. E. Birch, P. D. Lawrence, and R. D. Hare, "Single-Trial Processing of Event-Related Potentials Using Outlier Information", *IEEE Trans. Biomed. Eng.*, **40**(1): 59–73, 1993.

[5] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution", *Neural Computation*, **13**(7): 1443–1471, 2001.

[6] D. Tax and R. Duin, "Uniform object generation for optimizing one-class classifiers", *Journal for Machine Learning Research*, 155–173, 2001.

[7] P. Laskov, C. Schäfer, I. Kotenko, and K.-R. Müller, "Intrusion detection in unlabeled data with quarter-sphere Support Vector Machines (Extended Version)", *Praxis der Informationsverarbeitung und Kommunikation*, **27**: 228–236, 2004.

[8] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control", *Clin. Neurophysiol.*, **113**: 767–791, 2002.

[9] B. Blankertz, G. Curio, and K.-R. Müller, "Classifying Single Trial EEG: Towards Brain Computer Interfacing", in: T. G. Diettrich, S. Becker, and Z. Ghahramani, eds., *Advances in Neural Inf. Proc. Systems (NIPS 01)*, vol. 14, 157–164, 2002.

[10] L. Trejo, K. Wheeler, C. Jorgensen, R. Rosipal, S. Clanton, B. Matthews, A. Hibbs, R. Matthews, and M. Krupka, "Multimodal Neuroelectric Interface Development", *IEEE Trans. Neural Sys. Rehab. Eng.*, (**11**): 199–204, 2003.

[11] L. Parra, C. Alvino, A. C. Tang, B. A. Pearlmutter, N. Yeung, A. Osman, and P. Sajda, "Linear spatial integration for single trial detection in encephalography", *NeuroImage*, **7**(1): 223–230, 2002.

[12] W. D. Penny, S. J. Roberts, E. A. Curran, and M. J. Stokes, "EEG-Based Communication: A Pattern Recognition Approach", *IEEE Trans. Rehab. Eng.*, **8**(2): 214–215, 2000.

[13] K.-R. Müller, C. W. Anderson, and G. E. Birch, "Linear and Non-Linear Methods for Brain-Computer Interfaces", *IEEE Trans. Neural Sys. Rehab. Eng.*, **11**(2): 165–169, 2003.

[14] S. Harmeling, G. Dornhege, D. Tax, F. Meinecke, and K.-R. Müller, "From outliers to prototypes: ordering data", *Neurocomputing*, 2005, accepted.

[15] A. Stuart and K. Ord, *Distribution Theory*, vol. 1 of *Kendall's Advanced Theory of Statistics*, Wiley, 1994.

[16] B. Blankertz, G. Dornhege, C. Schäfer, R. Krepki, J. Kohlmorgen, K.-R. Müller, V. Kunzmann, F. Losch, and G. Curio, "Boosting Bit Rates and Error Detection for the Classification of Fast-Paced Motor Commands Based on Single-Trial EEG Analysis", *IEEE Trans. Neural Sys. Rehab. Eng.*, **11**(2): 127–131, 2003.

[17] M. Krauledat, G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "The Berlin Brain-Computer Interface For Rapid Response", *Biomed. Tech.*, **49**(1): 61–62, 2004.

[18] M. Krauledat, G. Dornhege, B. Blankertz, F. Losch, G. Curio, and K.-R. Müller, "Improving Speed And Accuracy Of Brain-Computer Interfaces Using Readiness Potential Features", in: *Proceedings of the 26th Annual International Conference IEEE EMBS on Biomedicine, San Francisco*, 2004.

[19] R. Q. Cui, D. Huter, W. Lang, and L. Deecke, "Neuroimage of voluntary movement: topography of the Bereitschaftspotential, a 64-channel DC current source density study", *Neuroimage*, **9**(1): 124–134, 1999.

[20] W. Lang, O. Zilch, C. Koska, G. Lindinger, and L. Deecke, "Negative cortical DC shifts preceding and accompanying simple and complex sequential movements", *Exp. Brain Res.*, **74**(1): 99–104, 1989.

[21] R. J. Croft and R. J. Barry, "Removal of ocular artifact from the EEG: a review", *Neuropsychol. Clin.*, **30**: 5–19, 2000.

[22] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms", *IEEE Trans. Biomed. Eng.*, **51**(6): 993–1002, 2004.

[23] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Increase Information Transfer Rates in BCI by CSP Extension to Multi-class", in: S. Thrun, L. Saul, and B. Schölkopf, eds., *Advances in Neural Information Processing Systems* , vol. 16, 733–740, MIT Press, Cambridge, MA, 2004.

[24] G. Pfurtscheller and F. H. L. da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles", *Clin. Neurophysiol.*, **110**(11): 1842–1857, 1999.

[25] Z. J. Koles and A. C. K. Soong, "EEG source localization: implementing the spatio-temporal decomposition approach", *Electroencephalogr. Clin. Neurophysiol.*, **107**: 343–352, 1998.

[26] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement", *IEEE Trans. Rehab. Eng.*, **8**(4): 441–446, 2000.

[27] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The Berlin Brain-Computer Interface: Report from the Feedback Sessions", *Tech. Rep. 1*, Fraunhofer FIRST, 2005.

[28] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Combining Features for BCI", in: S. Becker, S. Thrun, and K. Obermayer, eds., *Advances in Neural Inf. Proc. Systems (NIPS 02)*, vol. 15, 1115–1122, 2003.

[29] B. Blankertz, K.-R. Müller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schlögl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schröder, and N. Birbaumer, "The BCI Competition 2003: Progress and Perspectives in Detection and Discrimination of EEG Single Trials", *IEEE Trans. Biomed. Eng.*, **51**(6): 1044–1051, 2004.

[30] P. Shenoy, M. Krauledat, B. Blankertz, R. P. N. Rao, and K.-R. Müller, "Towards Adaptive Classification for BCI", *Journal of Neural Engineering* , **3**: R13-R23, 2006.