Classification of independent components of EEG into multiple artifact classes

Frølich, Laura; Andersen, Tobias; Mørup, Morten

Published in:
Psychophysiology

Link to article, DOI:
10.1111/psyp.12290

Publication date:
2015

Document Version
Peer reviewed version

Classification of independent components of EEG into multiple artifact classes

Abstract

In this study, we aim to automatically identify multiple artifact types in EEG.

We used multinomial regression to classify independent components of EEG data, selecting from 65 spatial, spectral, and temporal features of independent components using forward selection. The classifier identified neural and five non-neural types of components.

Between subjects within studies, high classification performances were obtained. Between studies, however, classification was more difficult. For neural vs. non-neural classifications, performance was on par with previous results obtained by others.

We found that automatic separation of multiple artifact classes is possible with a small feature set.

Our method can reduce manual workload and allow for the selective removal of artifact classes. Identifying artifacts during EEG recording may be used to instruct subjects to refrain from activity causing them.

Introduction

EEG data is generally contaminated by artifactual, non-neural electrical activity stemming from non-physiological sources such as electrical background noise and loose electrodes, and physiological sources such as subjects' heartbeat, muscle or eye movements. Such non-neural activity can, to some extent, be separated from the data using Independent Component Analysis (ICA), which is a widely used method in data analysis (Hyvärinen and Oja 2000; Comani et al. 2004; Di et al. 2007; C. M. Kim et al. 2003; Kong et al. 2008; Tsai and Lai 2009). Particularly, it is commonly used for pre-processing and analyzing EEG data (Erfanian and Erfani 2004; Acar, Makeig, and Worrell 2008; Ullsperger and Debener 2010). ICA extracts spatial patterns with statistically independent behavior over time from the raw EEG data (Hyvärinen and Oja 2000). These patterns and their corresponding time series are referred to as independent components (ICs).

Non-neural activity in EEG is typically considered a nuisance and the main purpose of separating it from the data using ICA is to exclude it by filtering (Jung et al. 2000). Other approaches to cleaning data include identifying heavily contaminated channels or epochs of EEG data, and then removing such channels or epochs prior to analysis (Nolan, Whelan, and Reilly 2010; Ypparila et al. 2004; Citi, Poli, and Cinel 2010). Unfortunately, this may lead to unnecessary data loss. Simultaneous reordings of e.g. the electrooulogram (EOG) and electrocardiogram along with EEG can also be

used to remove artifats (Nolan, Whelan, and Reilly 2010; Fatourechi et al. 2007; He, Wilson, and Russell 2004) but this approach is not useful for all types of artifacts and requires the additional labor of mounting auxiliary sensors. Therefore, we only consider approaches using ICA based solely on EEG data in the current study.

State of the art

Presently, classification of ICs for artifact detection in EEG is often done manually in a time-consuming and subjective process. While work on fully automated supervised classification methods has increased over the past years, most of this work has focused on the binary problem of distinguishing between neural and non-neural ICs (Winkler, Haufe, and Tangermann 2011; Tangermann et al. 2009; Mognon et al. 2010; LeVan, Urrestarazu, and Gotman 2006; Bartels, Shi, and Lu 2010; Viola et al. 2009; Halder et al. 2007), some using multiple classes as an intermediate step (Mognon et al. 2010; Bartels, Shi, and Lu 2010; Halder et al. 2007) with only few studies evaluating performance for the multi-class problem (Viola et al. 2009; Halder et al. 2007).

Several studies have used simulated artifacts as a ground truth to which they compare their automatic classification e.g. (Nolan, Whelan, and Reilly 2010; Delorme, Sejnowski, and Makeig 2007). This is problematic as real artifacts may affect data in ways different from the simulation. Therefore, we limit our review to those studies that compare performance with human experts' classifications of real artifacts in real data.

The most important factor in performance evaluation is generalization. For a method to be fully automated it should perform well when tested on data from a study that was not used to train the method. Automatic thresholding at e.g. a pre-determined Z-score for a certain feature is one approach that allows this (Nolan, Whelan, and Reilly 2010; Mantini et al. 2008). Another approach is to train a classifier on data from one study and make sure that it performs well on data from another study. This would allow the method to be trained once and then applied to new data without manual intervention. Few studies have tested supervised classifiers for artifact detection at this level of generalization, and only for the binary problem of distinguishing artifactual from neural ICs (Winkler, Haufe, and Tangermann 2011; Mognon et al. 2010).

Winkler, Haufe, and Tangermann (2011) built a classifier based on an initial pool of 38 features from the spatial, spectral, and temporal domains. They compared several classification methods and found that regularized linear discriminant analysis with three spectral, one temporal, and two spatial features obtained the best classification results. We describe most of the features in their initial pool in detail in [app:feats]. They reported error rates of 8.9% and 14.7% within and between studies, respectively. The method ADJUST (Mognon et al. 2010) uses Gaussian densities for classification and incorporates features from the spatial and temporal domains of ICs. ADJUST

employs class-specific classifiers for eye blinks, vertical eye movements, horizontal eye movements and generic discontinuities (non-biological artifacts) to solve the classification problem by classifying an IC as being non-neural if one or more class-specific classifiers labeled it as artifactual. The evaluation measure reported for ADJUST was the percentage of data variance explained by correctly classified ICs. On test data, the ADJUST performance measure was 99.0%, 96.0%, 99.2% and 97.7% for the class-specific classifiers for blinks, vertical eye movements, horizontal eye movements, and generic discontinuities, respectively. In classifying neural vs. non-neural ICs, the ADJUST performance measure was 95.2%. Several classes were considered in ADJUST, and so the method is appropriate to be used for multi-class classification purposes. Since an IC may be assigned to several classes, ADJUST can, strictly speaking, not be tested in the multi-class problem in its current form.

A few studies have addressed multi-class identification of artifacts (Viola et al. 2009; Mantini et al. 2008) at the level of across-subject-generalization within a study. This level of generalization could certainly also be useful as it would allow automatic artifact classification of future subjects once manual classification of some subjects has been achieved.

Viola et al. (2009) introduced the method CORRMAP, which solely uses the scalp map of an IC to classify it as representing a blink, a lateral eye movement, or the heartbeat. CORRMAP classifies an IC using the correlation between the spatial topography of the IC and template topographies from ICs with known classes. If the correlation is higher than a certain threshold, which can either be set manually or determined automatically, then the IC is classified as being of the same class as the template IC.

In Viola et al.'s study, classification rates were calculated for electrode configurations with 30, 68, and 128 channels for three classes: blinks, lateral eye movements, and heartbeats. The mean correlation over electrode arrays between CORRMAP and human experts for these three classes were 0.90, 0.88, and 0.47, respectively. The classification rates for blinks and lateral eye movements were higher for the less dense electrode arrays, while classification of heartbeats improved with denser electrode arrays.

A new fully automated method using the same principle as CORRMAP, of using the correlation between spatial maps as sole feature, has recently been presented (Bigdely-Shamlo et al. 2013) for the identification of eye-related ICs. An area under the receiver operating characteristics (ROC) curve of 0.993 was obtained on ICs from a study that was not used during training. This result shows that the principle behind CORRMAP is a very promising method for automatic artifact identification at the highest level of generalization, namely across studies.

Mantini et al. (2008) used thresholding of a single feature, the approximate entropy of IC time series, to classify ICs of MEG as non-cerebral biological artifacts (low approximate entropy), neural (medium approximate entropy) or environmental noise components (high approximate entropy). They obtained very good results with the area under the ROC curves being above 0.9 with labels by human experts as ground truth. As this method separates artifacts into biological and non-biological ICs it does address the multi-class problem but it is unknown whether it is suitable for a further division of these classes into more specific classes such as lateral eye movements versus eye blinks.

Purpose of study

By distinguishing between multiple types of artifacts such as eye movements and the electrical heartbeat artifact, more diverse uses of an automatic classification method can be imagined since some artifacts may be informative for some purposes, or it may be desirable to remove only some artifact types. The heartbeat, for example, may be an informative signal in some settings, or eye-related ICs could be used to detect drowsiness. Automatic detection and identification of multiple types of artifacts during EEG recording would also allow researchers to instruct subjects to refrain from the activity causing those artifacts.

The purpose of the study is to develop a multi-class artifact detection system covering four diverse artifact classes: eye blinks, horizontal eye movements, heartbeat artifacts and muscle artifacts, as well as ICs consisting of mixed neural and artifactual activity. Importantly, we test the performance of the system at two levels of generalization: between subjects within a study and between studies. A good performance across subjects would allow a classifier to be trained for the first subjects in a study, and then used to automatically classify ICs for the subsequent subjects. A good performance across studies would mean that the classifier can be used on arbitrary studies and subjects without re-calibration. We are also interested in determining the features most relevant to classifying ICs. Hence we aim to answer the following research questions:

1. **Which features are important for a high performance in multi-class classification of ICs?**
2. **Is it possible to distinguish between multiple classes of ICs between subjects within a study?**
3. **Will a classifier generalize between studies?**

Data

Two data sets containing manually labeled ICs were kindly made available by Scott Makeig, Julie Onton and Klaus Gramann (Onton and Makeig 2009; Gramann, Tollner, and Muller 2010).

[tab:data]

One data set was acquired for the purpose of studying the EEG during different emotional states (Onton and Makeig 2009). Subjects were seated in a dimly lit room with eyes closed, imagining emotional states. This study contained recordings from 34 subjects from a Biosemi[1] 250 channel active reference system (Onton and Makeig 2009). Channels that showed highly abnormal activity had been removed manually before performing ICA, leaving 134-235 channels for each subject. The ICA decompositions for this data were obtained by "full-rank decomposition by extended infomax ICA" (Onton and Makeig 2009). The 34 data sets were between eight and eighty-eight minutes long after concatenating the recordings for the various emotions imagined. We will refer to this data set as the *Emotion* data or study. The other data set was recorded to investigate how attention is guided early in visual processing. This was recorded from 64 scalp channels "referenced to Cz and re-referenced off-line to linked mastoids" from 12 subjects during a visual task (Gramann, Tollner, and Muller 2010). ICA was performed with the implementation of the ICA infomax algorithm in the Brain Vision Analyzer software from Brain Products GmbH[2]. The data sets we had access to were between 56 and 66 minutes long for the different subjects. We will refer to data from this study as the *Cue* data or study. The data sets are summarized in Table [tab:data].

The two data sets differed in various ways (see Table [tab:data]). The number of electrodes was much higher in the Emotion study than in the Cue study, implying a higher spatial sampling of the EEG. The Emotion study also contained more subjects, resulting in a total of almost ten times as many ICs in the Emotion study as in the Cue study. Also, different sampling rates and analogue filters were used and the lengths of recordings differed. Additionally, the experimental tasks differed. In the Emotion study, an eyes-closed task was performed while a task requiring responses to visual cues was used in the Cue study. These differences are likely to cause covariate shifts in the data, i.e. differences in distributions of features between training data and future data (Sugiyama and Kawanabe 2012), in the features across studies if features are calculated naively from the raw data. We discuss how we handle this issue in Section [sec:methods].

---

[1] http://www.biosemi.com/

[2] http://www.brainproducts.com/

Both studies contained ICs labeled by experts with the labels “eye blink”, “neural”, “heart”, “lateral eye movement”, and “muscle”. Two experts, one in each study, performed the manual classification of ICs. Figure [fig:scalpmaps] shows examples of scalp maps from the different classes. Neural ICs are the ICs that correspond to activity generated by neural sources within the brain. ICs with the label “heart” represent the electrical heartbeat artifact. The ICs that were not labeled represented, based on visual inspection, mixed ICs containing both artifactual and neural signals. We will refer to the unlabeled ICs as “mixed” ICs. We chose to include mixed ICs in our analysis since mixed ICs will almost always be present in real data. Not including this class would then force mixed ICs to be classified as one of the four artifact, or neural classes. Since mixed ICs have different characteristics from neural ICs, it is likely that many would be classified as artifactual. This is undesirable since mixed ICs also contain traces of neural activity, meaning that the removal of mixed ICs would imply a loss of neural activity in data. The inclusion of mixed ICs can also be seen as a step toward making the classifier mimic human expert classifications as much as possible.

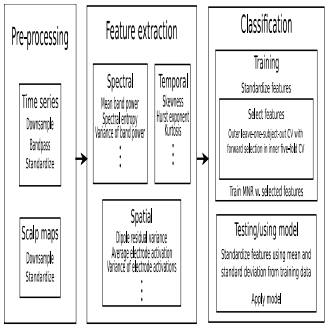
[fig:scalpmaps]

Some types of ICs are much more common than others, which presents a challenge to classification methods as described in Section [sec:methods]. Mixed ICs, for example, make up the majority of available ICs. The numbers and proportions of the different types of ICs in each study are shown in Figure [fig:classprops].

[fig:classprops]

Methods

Figure [fig:classifier_flowchart] shows the pipeline used to train and validate our IC classifier. Each of the steps is described in detail in the remainder of this section.



Processing pipeline for ICs from EEG data. The abbreviations CV and MNR stand for cross-validation and multinomial regression, both explained in section [sssec:classification].

[fig:classifier_flowchart]

We first discuss the steps taken during pre-processing to avoid covariate shifts between studies due to differences in experimental setups. Next, we discuss our feature set. We then describe our classification and feature selection procedures. Finally, we outline how we investigated the research questions posed in the introduction.

Pre-processing

Different EEG studies use different sampling rates, analogue filters, and electrode arrays during recordings, and durations of recordings vary. If features that are influenced by such differences are used, it is improbable that a classifier will generalize across studies.

Higher sampling rates enable spectral features to be determined for higher frequencies. Likewise, different analog filters during recording of EEG cause the spectral content of signals to vary systematically. To avoid such differences, we filter and resample all signals before calculating features. We require that any data given as input was recorded with a sampling rate of at least 200Hz, and that the analogue filter used during recording had a low edge of 3Hz or lower and a high edge of 90Hz or higher. With these requirements in place, it is safe to band-pass filter the signal between 3Hz and 90Hz and downsample all input signals to 200Hz. This ensures that all feature calculations are performed on signals with the same spectral content.

Different durations of recordings entail different uncertainties in the calculation of temporal and spectral features. Invariance to this effect is achieved by using the means and variances of temporal and spectral characterstics of the signal over one-second intervals as temporal and spectral features.

Some features are based on distances between electrodes and are thus clearly influenced by electrode array density. We require that recordings were performed using an array with at least 64 electrodes to ensure a good spatial coverage. We spatially downsample all scalp maps to the 10-20 system electrode array with 64 electrodes. The spatial downsampling is performed with Gaussian kernels using spherical distances between electrodes. We use a standard deviation of 0.5 cm and a head radius of 9 cm.

Before calculating features derived from the spatial distribution of an IC, we standardized the spatial map. Each column of the mixing matrix was standardized to have variance one and mean zero. This ensures that only patterns in the spatial map, and not its scale, determine the features calculated. This is desirable since the magnitude of the mixing matrix cannot be uniquely determined due to an inherent ambiguity in the scaling of the mixing matrix and the matrix of activation time series of ICs. We also standardized time series before calculating temporal and spectral features.

Features

An IC consists of a scalp map containing the contribution of the IC to each EEG channel, and a time series that shows how active the spatial pattern is over time. To quantify the characteristics of an IC, features based on both the spatial and temporal representations have been shown to be relevant (Winkler, Haufe, and Tangermann 2011; Mognon et al. 2010; Viola et al. 2009). Spectral (frequency domain) characteristics of the time series have also been shown to be informative (Winkler, Haufe, and Tangermann 2011). Hence we use features from the spatial, temporal, and spectral domains. We included most of the features described in two recent studies of the binary classification problem (Mognon et al. 2010; Winkler, Haufe, and Tangermann 2011). Descriptions of features are given in [app:feats]. Before training we standardized the features in the training set to have mean zero and variance one. We standardized the test data using the mean and variance from the training data, which is the standard approach (Jayalakshmi and Santhakumaran 2011; Hastie, Tibshirani, and Friedman 2009).

Classification

We used the linear classifier multinomial logistic regression (MNR) since this was found to obtain good results and linear classifiers are desirable both for their interpretability and fast training. Linear classifiers have previously shown good performance in the binary classification of ICs (Winkler, Haufe, and Tangermann 2011).

As is evident from Figure [fig:classprops], the class of mixed ICs makes up the large majority of ICs in both studies. Thus a classifier would achieve a high classification rate by classifying all ICs as mixed. This problem of imbalanced classes is well known, and various approaches to solving it have been proposed (Zadrozny, Langford, and Abe 2003; López et al. 2012). We weighted observations by the reciprocal of their class proportion during training such that the penalty of misclassification was higher for ICs from smaller classes. This weighting scheme can be considered a proxy for optimizing balanced accuracy. Balanced accuracy is a performance measure that weighs all classes equally since it is defined as the mean over classes of the proportion of correct classifications in each class. In the binary case, balanced accuracy is thus the mean of specificity and sensitivity.

Previous studies on the binary classification problem found that only few features are necessary to distinguish between classes (Winkler, Haufe, and Tangermann 2011; Tangermann et al. 2009; Mognon et al. 2010). This motivated us to investigate research question 1 of whether only few features are sufficient in the multi-class problem as well. This was done in a two-level cross-validation (CV). In the outer level, leave-one-subject-out CV was performed over the 34 subjects in the Emotion data. In each outer fold, features were chosen using forward selection in an inner 5-fold stratified CV

by adding features to an MNR model until the test error stopped decreasing. The use of stratified CV ensured that class proportions were as equal as possible across partitions. For each feature, we counted the number of outer CV folds in which it was selected. This number reflects the importance or pertinence of the feature. We then created 35 sets of features consisting of the features that had been selected in at least 0, 1, 2, …, 34 outer CV folds. For each subject, the classifier was trained on each of these feature sets using the 33 other subjects, and tested on the left-out subject. The classes of ICs predicted for each subject in this manner were used to calculate a balanced accuracy for each feature set. As the best feature set we chose the sparsest feature set with acceptable performance.

Investigation of research questions

Research question 1, concerning the features important for multi-class classification, was investigated by comparing classification performances with different feature sets. These feature sets were the ones constructed using the Emotion data as described in Section [sssec:classification]. The Emotion data was also used to choose the best feature set. To evaluate the sensitivity of the classification performance to the choice of features, balanced accuracies were calculated in leave-one-subject out CV on the Cue data and across-study training and testing for each feature set constructed from the Emotion data. If new ICs to be classified have short time series, spectral and temporal features will likely be badly determined. In such cases, the exclusive use of spatial features would be preferable. For this reason, we also tested the classifier using only the spatial features.

Both research questions 2 and 3 were investigated using the feature set determined based on Emotion data. We investigated research question 2, concerning between-subject generalization within studies, through the leave-one-subject-out CV schemes on both the Emotion and Cue data sets. A high classification performance when testing on a subject not used during training would signify that it is possible for a classifier to generalize across subjects within a study, meaning that each class of ICs exhibits certain characteristics independently of the specific subject.

To answer question 3, concerning between-study generalization, we trained a model on each data set using the features selected using the Emotion data. The models were then tested on all subjects from the other study. A good performance on subjects from the other study would indicate that the classifier is able to generalize across studies. We used confusion matrices to inspect the classification performance of the classifiers on a class-by-class basis. We also used the balanced accuracy rate to evaluate performance and compare to classification performances obtained by others.

Results

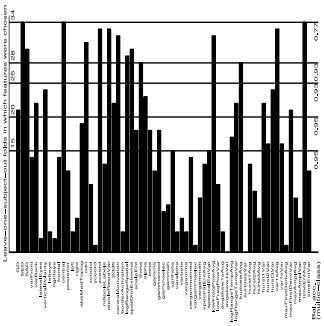
Figure [fig:barplot] shows the number of times each feature was chosen by forward selection in the leave-one-subject-out CV scheme performed on the Emotion data. The balanced accuracies obtained using the features chosen in at least 15, 20, 25, 28, or 34 outer folds are also shown. The feature sets constructed using the thresholds 15, 20, 25, 28, and 34 contain 32, 23, 14, 14, and 3 features, respectively. The two 14-feature sets are identical.

Figure [fig:balaccs$_a$t$_t$hresholds] shows the balanced accuracies obtained with each of the 35 feature sets. The variability of the curves in Figure [fig:balaccs$_a$t$_t$hresholds] gives an idea of how sensitive the classification performance is to the choice of feature set.

Figure [fig:thresh$_c$onfmat] shows the confusion matrices that arose from using the 32-feature set, the 23-feature set, the 14-feature set, and the 3-feature set in a leave-one-subject-out CV on Emotion data. This figure is included to show that the class-wise performances are stable over the different feature sets.

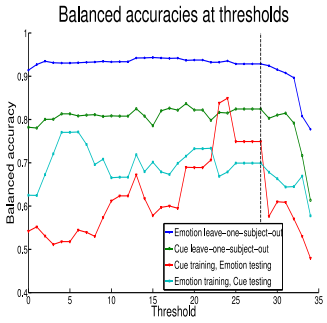
Figure [fig:confusionmatsspatial] shows the confusion matrices obtained in leave-one-subject-out CV on both studies, and with cross-study training and testing using only the spatial features in the initial pool of features. This figure is shown to illustrate the classification performance that can be expected if only short time series of ICs are available, in which case non-spatial features may be unreliable.

Figure [fig:confusionmats] shows the class-wise classification performances when the classifier with the feature set containing 14 features is used. The confusion matrix in the top row shows the performance with leave-one-subject-out evaluation on the Cue data and the two confusion matrices in the bottom row show the cross-study performances. This figure details the class-wise performances, which cannot be derived from the balanced accuracy rates shown in Figure [fig:balaccs$_a$t$_t$hresholds].



Barplot showing the number of folds each feature was chosen by forward selection in leave-one-subject-out CV on the emotion data (34 subjects).

[fig:barplot]



Balanced accuracy obtained with different feature sets constructed by varying the number of CV folds features must have been selected in to be included. The dashed line shows the cut-off chosen based on the blue curve. This choice led to a feature set containing 14 features.

[fig:balaccs$_a$t$_t$hresholds]

[fig:thresh$_c$onfmat]

[fig:confusionmatsspatial]

[fig:confusionmats]

Discussion

Before analyzing the classification performance obtained by our classifier we discuss the classification performance of human experts, which sets the upper bound on the performance we might hope to achieve.

Performance of human experts

As the true underlying content of ICs, i.e. the ground truth, is unknown, we can only rely on classifications made by expert human observers when training and testing classifiers. Several studies have found that the agreement between human experts is generally less than perfect and that it differs for different types of artifacts (Viola et al. 2009; Winkler, Haufe, and Tangermann 2011; Klekowicz et al. 2009). Although the agreement between experts is likely dependent on the particular method of ICA, the information available to the experts, the particular data sets and how experts are instructed to classify ambiguous cases, there seems to be a good agreement between studies on the inter-expert agreement rate (Viola et al. 2009; Winkler, Haufe, and Tangermann 2011; Klekowicz et al. 2009).

Viola et al. (2009) had 11 independent experts classify ICs as eye blinks, lateral eye movements and heartbeat artifacts based solely on the scalp maps of ICs. The data came from three independent studies and observers were under the constraint that a maximum of three ICs could be identified as containing one particular artifact type. In terms of the binary correlation the

inter-expert agreement was very high for eye-blinks (0.82 – 1.00), high but more variable for lateral eye movements (0.55 – 0.93) and low and very variable for heartbeat (0.02 – 0.73).

Winkler, Haufe, and Tangermann (2011) had 2 experts classify ICs from a single study as artifactual or neural based on their spectrum, time series and spatial distribution on the scalp and found that the error rate was 10.6%. They also had one expert re-label the ICs from another study two years after the same expert's first labeling of the same data. The error rate between the two labelings was 13.2%. This is not much higher than the agreement between experts and the disagreement may thus reflect the inherent difficulty of the task rather than differences in technique or approach by different observers.

Klekowicz et al. (2009) made 22 comparisons between expert classifications (artifact vs. neural) based on the EEG time series from 7 polysomnographic recordings and found an agreement of 0.92 in terms of the area under curve of the best fitting ROC curve. Of the 22 comparisons, four were between classifications made by the same expert at different points in time. From their figures (Figure 6 in their article) these agreements were high compared to the agreement between different observers. Hence, their reported overall agreement between human classifications is a high estimate of the agreement between human experts.

The imperfect agreement between human experts should be kept in mind when evaluating automatic artifact detection systems as inter-expert agreement sets the upper limit for what we can hope to achieve through automatic classification. It is very promising that several studies have reported a good agreement between automatic IC classification and human experts, close to the agreement between experts.

Evaluation of classifier

Feature selection

In Figure [fig:balaccs$_a$t$_t$hresholds], the blue curve shows the average leave-one-subject-out CV performance on the Emotion data, the same data used to construct the feature sets. This is also the curve used to determine the feature set to use in the classifier. The feature set resulting from requiring that features must have been included in 28 CV folds or more was chosen as the best feature set since classification performance starts to consistently decrease at this threshold. This feature set includes 14 features, consisting of nine spatial, two spectral, and three temporal features. The red and blue curves are biased upwards since testing for these curves was performed on the Emotion data, which was used to choose the feature sets, implying that the feature sets contain features especially well suited to describing ICs of different classes in the Emotion data. At threshold zero, when all features are included, there is no bias since no features were chosen based on the Emotion data at this point. Figure [fig:balaccs$_a$t$_t$hresholds] shows that, when training

and testing on subjects from the same study (blue and green curves), the performance is stable for most feature sets until the number of features becomes too small. This indicates that, within a study, overfitting to subjects in the training data is not a problem, even for the relatively small amount of data present in the Cue study. The lack of upwards or downwards trends in the performance when training on Emotion data and testing on Cue data (cyan curve) indicates that the Emotion study contains sufficient data that overfitting is avoided. Conversely, when training on Cue data and testing on Emotion data (red curve), the performance peaks with feature sets that are neither too small nor too large. One explanation of this is that there is not enough data in the Cue data set to prevent overfitting when very large feature sets are used. Another explanation is that, since features were chosen based on Emotion data, small feature sets help the model home in on characteristics that best discriminate classes of ICs in Emotion data. All curves indicate that underfitting occurs with feature sets that are too small. In summary, Figure [fig:balaccs$_a$t$_t$hresholds] shows that the classification performance is quite robust to the specific choice of threshold when training and testing on subjects from the same study, whereas the performance is sensitive to the choice of threshold when training on one study and testing on the other study.

The inclusion of both spatial, spectral, and temporal features in nearly all feature sets (Figure [fig:barplot]) shows that all three types of features carry information on the classes of ICs. The features included in the 14-feature set are shown in Table [tab:features], arranged according to the classes they should be good at detecting.

For the spatial feature set, the within-study performances were very similar to those obtained with the 14-feature set (compare confusion matrix (c) in Figure [fig:thresh$_c$onfmat] and confusion matrix (a) in Figure [fig:confusionmats] to confusion matrices (a) and (b) in Figure [fig:confusionmatsspatial]). In the between-study case, the performance improved when testing on Cue data and decreased when testing on Emotion data (compare confusion matrices (b) and (c) in Figure [fig:confusionmats] to confusion matrices (c) and (d) in Figure [fig:confusionmatsspatial]). However, the performance when testing on Emotion data with the 14-feature set is biased upwards since features were chosen using the Emotion data. Thus the decrease seen when testing on Emotion data with the spatial feature set might be artificial, indicating that spatial features may be sufficient if across-study generalization is to be improved.

[tab:features]

Classification performance with the 14-feature set

The following discussion of the classification performance is based on the results given for the classifier with the 14-feature set.

When classifying ICs in the within-study case into only two classes, artifactual or non-artifactual, we obtain balanced accuracy rates of 0.90 and 0.95. This is comparable to performances obtained by others. Balanced accuracy rates of 0.91 and 0.79 were obtained in Winkler, Haufe, and Tangermann (2011) and LeVan, Urrestarazu, and Gotman (2006), respectively, while Halder et al. (2007) and Bartels, Shi, and Lu (2010) report balanced accuracy rates above 0.90 without giving the exact numbers. Likewise, our classifier performs on par with others in the binary across-study case, obtaining balanced accuracy rates of 0.88. In the across-study case, Winkler et al. obtained a balanced accuracy of 0.86 (Winkler, Haufe, and Tangermann 2011). These accuracy rates compare well with the inter-expert agreement seen in previous studies (Viola et al. 2009; Winkler, Haufe, and Tangermann 2011; Klekowicz et al. 2009).

In the following, we discuss the multi-class performance. This is visualized in confusion matrix (c) in Figure [fig:thresh$_c$onfmat] for the leave-one-subject-out CV on the Emotion data, and in Figure [fig:confusionmats] for the leave-one-subject out CV on the Cue data and the cross-study training and testing. The performance on the class of lateral eye movements is high. This could be expected since eye-related ICs have previously been classified well by many others (Mognon et al. 2010; Viola et al. 2009; Bigdely-Shamlo et al. 2013). For the blink class, however, difficulty is experienced when training on Cue data and testing on Emotion data. This could be due to the low number of observations (14) of the blink class in the Cue data, making it difficult for the classifier to learn a good characterization of this class. The high performance on the neural class is also in good agreement with that found by others (Winkler, Haufe, and Tangermann 2011; Mognon et al. 2010). When tested on Cue data, heartbeat ICs tended to be misclassified as neural. Difficulty with the heartbeat class has also been observed in previous work including this class, both for an automatic classifier and for human experts (Viola et al. 2009). The high degree of confusion between the classes of muscle and mixed ICs may partly be explained by the shared characteristic of highly peaked scalp maps in these two classes compared to the other classes. The class most often confused with other classes is that of mixed ICs, which is not surprising since mixed ICs are ICs that do not clearly belong to one class, but may contain characteristics of several classes. The classification of some mixed ICs as neural is arguably difficult to avoid as the contrast between neural and mixed will be based on a threshold, which may be poorly defined.

In general, the classifier performs better when trained on other subjects within the same study than when trained on subjects from another study. High classification performances with balanced accuracies of 93% and 80% for the Emotion and Cue data, respectively, were found in the within-study cases. Evaluation between studies, however, gave balanced accuracies of 74% and 62% when testing on Emotion and Cue data, respectively. Data from more

studies would probably help the between-study performance approach the within-study performance. Another way to improve the across-study performance could be to take into account the distributions of feature values in the test data set compared to the training data set.

Quality of ICA decomposition

Since the quality of an ICA decomposition depends on the pre-processing of data before running ICA, the usefulness of a classifier also depends on the pre-processing steps. If data is subjected to ICA with little pre-processing, many ICs are likely to be either mixed or noisy representations of individual classes. Since such ICs are difficult to classify, the performance of the classifier is likely to decrease. If ICs are truly mixed, classification into separate classes is not possible even for human experts. A future approach to tackling such cases could be to use the class probabilities given by MNR to decide how to handle mixed ICs. If, for example, an IC classified as mixed is also given somewhat high probabilities of representing blinks and lateral eye movements, the IC could be classified as being generally eye-related and discarded. On the other hand, mixed ICs could be retained if the probability of the neural class is above some pre-defined threshold.

Online capability

The reasonable performance of the classifier makes it possible to use it for online monitoring of artifact occurrence while recording EEG. A rule of thumb states that about $20\times n^2$ samples are necessary to perform an ICA of n channels (Ullsperger and Debener 2010). Hence an ICA and classification of resulting ICs can be performed every $20 \times n^2/f$ seconds, where f is the sampling rate. With 64 channels and a sampling rate of 512Hz, three minutes of recorded EEG provides sufficient data for an ICA decomposition. Using the *runica* algorithm in EEGLab (Delorme and Makeig 2004) with at most 50 iterations, an ICA decomposition can be calculated in less than two minutes and calculating the features for an IC takes less than one minute. By distributing the feature calculations for the ICs over several threads, classified ICs can be provided online at a lag of about six minutes.

Conclusion

The presence of artifactual activity in EEG recordings is problematic in the analysis of data. While different approaches to removing such noise exist, these are either subjective and require lengthy manual processing of data or distinguish only between two classes. In this paper, we described an approach to automatic multi-class classification of artifactual ICs of EEG data. We considered neural ICs and five artifact classes: eye blinks, heartbeat, lateral eye movements, muscle, and mixed neural and artifactual activity. Using an initial pool of 65 spatial, spectral, and temporal features invariant to

experimental setup, we investigated which features were important for classification of ICs. We found that features from all three spatial, spectral, and temporal domains carried information important for classification. However, we also saw that classification with a feature set consisting of only the spatial features had very similar performance to the 14-feature set when evaluating the classifier within studies. Across studies, the performance increased with the spatial feature set when testing on Cue data. The performance decreased when testing on Emotion data, but this was compared to the upwards biased performance estimate obtained with the 14-feature set chosen based on Emotion data. The classifier generalizes well across subjects within studies, whereas across-study generalization is more challenging. Collapsing the multi-class classifications into binary classifications (artifact or neural), we obtain classification performances comparable to those found in previous studies both within and between studies (Winkler, Haufe, and Tangermann 2011; Mognon et al. 2010; LeVan, Urrestarazu, and Gotman 2006; Bartels, Shi, and Lu 2010; Viola et al. 2009; Halder et al. 2007). Thus the proposed classifier can be used for binary or multi-class classification interchangeably. The classification performance and speed of obtaining classified ICs allows online use of the classifier to detect artifacts while recording EEG so that subjects can be instructed to refrain from activity producing the detected types of artifacts. Although some artifacts such as the heartbeat are unavoidable, others may be mitigated in some paradigms, e.g. ERP studies, by the experiment being paused to allow subjects to blink or make them aware of muscle tension. Additionally, multi-class classification of artifactual ICs can make researchers aware of overly many artifacts of some class automatically. If possible, the experimental setup could then be redesigned to minimize the risk of such artifacts, e.g. by adjusting seating arrangements for participants to reduce eye and muscle tension. Additionally, the classifier could be used to identify artifacts typical of individual subjects in a short pilot run before performing an experiment.

We provide Matlab code for feature calculation and MNR classifiers trained on different feature sets online at http://www2.imm.dtu.dk/~lffr/publications/IC_MARC.zip. We hope that this will encourage others to further explore automatic classification of artifactual ICs and use this technique to ease data cleaning.

Acknowledgments

We would like to express our gratitude to Julie Onton, Klaus Gramann and Thomas Toellner for the use of their data (Onton and Makeig 2009; Gramann, Tollner, and Muller 2010), without which this study would not have been possible. Laura Frølich would also like to thank Scott Makeig for hosting her at the Swartz Center for Computational Neuroscience and for discussions about IC classification, and Christian Kothe for discussions and aid with

programming. We would also like to thank two anonymous reviewers for their constructive comments which have improved the manuscript.

Features

The 65 features of ICs used for classification are listed here. All these features are sign-invariant since the sign-ambiguity of spatial maps and time series of ICs cannot be resolved through normalization.

Spatial

- (GD) Generic discontinuity measure Mognon et al. (2010). This measure as used in ADJUST Mognon et al. (2010) is defined as

$$\max_n \left| a_n - \frac{1}{c-1} \sum_{m \neq n} \exp\left(-\| y_m - y_n \|\right) a_m \right|,$$

  where y_m is the location of the m^{th} electrode on the scalp, a_m is the activation of the m^{th} electrode by the IC, and c is the number of electrodes. Hence this measure gives a high value if the IC activates any electrode a lot more than the neighboring electrodes, indicative of e.g. a loose electrode.

  We use a slightly modified version of this measure to make the second term a weighted average. Our measure is defined as

$$\max_n \left| a_n - \frac{1}{\sum_{m \neq n} \exp\left(-\| y_m - y_n \|\right)} \sum_{m \neq n} \exp\left(-\| y_m - y_n \|\right) a_m \right|.$$

- (SED) Spatial eye difference Mognon et al. (2010). Absolute value of the difference between activation of electrodes around the left and right eye areas. The left eye area is defined to lie between the angles -61^o and -35^o with a radius larger than 0.3 (where the head radius is assumed to be one, the convention in EEGLab). The right eye area is defined to lie between the angles 34^o and 61^o, also at a radius larger than 0.3. Zero degrees is towards the nose and positive 90^o is at the right ear.
- (SAD) Spatial average difference Mognon et al. (2010). This feature is defined as the absolute value of the mean of frontal electrode activations minus the absolute value of the mean of posterior electrode activations. The frontal area is defined to be the electrodes with absolute angles less than 60^o and radii larger than 0.4. The posterior area consists of the electrodes with absolute angles larger than 110^o.
- (varFront and varBack) Variance of activation of frontal and posterior electrodes Mognon et al. (2010).

- **(lateralEyes) Absolute value of the difference between activation of electrodes around the left and right eye areas. The left eye area is defined as the mean over all electrodes, weighted by a Gaussian bell with center at the location of Fp1 in the 10-20 electrode system. The right eye area is defined as the mean over all electrodes, weighted by a Gaussian bell with center at the location of Fp2 in the 10-20 electrode system. The standard deviation of both Gaussian bells is set to be 1 cm and a head radius of 9 cm is assumed.**
- **(verticalPolarity) Absolute value of the difference between activation of frontal and posterior electrodes. The frontal area is defined as the mean of all electrodes weighted by a Gaussian bell centered at the location of AFz in the 10-20 electrode system. The posterior area is defined as the mean of all electrodes weighted by a Gaussian bell centered at the location of POz in the 10-20 electrode system. The standard deviation of both Gaussian bells is set to be 2 cm and a head radius of 9 cm is assumed.**
- **(lefteye, righteye, frontal, central, posterior, left, right) These features give the absolute values of the mean activations of electrodes in various areas of the scalp. Each area is defined as the mean over all electrodes, where the contribution from each electrode to the mean is weighted by a Gaussian bell. For the areas around the eyes (lefteye and righteye), the standard deviation of the Gaussian bell is 1 cm. For all other areas, it is 2 cm. A 9 cm radius of the scalp is assumed. The Gaussian bells are centered at the locations of Fp1, Fp2, AFz, Cz, POz, C5, and C4, respectively.**
- **(absMedTopog) The absolute value of the median of the values in the scalp map.**
- **(cdn) Current density norm Winkler, Haufe, and Tangermann (2011). The current density norm is a measure of the complexity of the current source distribution of an IC. A high complexity of the current source distribution indicates that the source of the IC is difficult to locate inside the brain, and thus that it is likely to be an artifact. This was one of the six final features included in the classifier described in Winkler, Haufe, and Tangermann (2011), in which a more detailed description can be found.**
- **(xcoord, ycoord, and zcoord) X, Y, and Z coordinates of dipole fit Winkler, Haufe, and Tangermann (2011). The dipole fit used returns a single dipole.**
- **(ndipoleLabels) Number of anatomical areas associated with dipole fit.**
- **dipoleResidualVariance**
- **(2ddft) Average logarithm of band power in high frequencies of spatial pattern Winkler, Haufe, and Tangermann (2011).**

- **(centralActivation) Logarithm of mean of absolute values of activations of central electrodes of IC Winkler, Haufe, and Tangermann (2011).**
- **(borderActivation) Binary feature to detect scalp maps with highest activity at an edge of the pattern. The most active electrode is the electrode for which the IC has the highest absolute value of activation. If the most active electrode in the pattern is in an outer group of electrodes, the feature is defined to be 1. Also, if the local maximum of an outer group is at the edge of the group, and its activation differs by more than two standard deviations from the group mean, then the feature is defined to be 1, too. Otherwise, it is defined to be -1 Winkler, Haufe, and Tangermann (2011).**
- **(logRangeSpatial) Logarithm of range of activation of electrodes. This was one of the six final features included in the classifier described in Winkler, Haufe, and Tangermann (2011).**
- **(spatDistExtrema) Euclidean distance in 3D coordinates between the two electrodes with minimal and maximal activation.**
- **(scalpEntropy) The entropy of the scalp map.**

Spectral

- **(theta, alpha, beta, gamma, gammamed, gammaelec and gammah) Mean over one-second intervals of the logarithm of band power in the θ (4-7Hz), α (8-13Hz), β (13-20Hz), lower γ (21-30Hz), middle γ (30-45Hz), γ around the power grid frequencies (both US and European) (46-65Hz), and higher γ (66-80Hz) bands. The average band power in the α-band was one of the six final features included in the classifier described in Winkler, Haufe, and Tangermann (2011).**
- **(vartheta, varalpha, varbeta, vargamma, vargammamed, vargammaelec and vargammah) The variance over one-second intervals of the logarithm of the bandpower in the same bands as mentioned above.**
- **(spectralEntropyAvg and spectralEntropyVar) The entropy of the power distribution over the bands mentioned above is calculated for one-second intervals of the time series. The feature spectralEntropyAvg is then the average over these one-second intervals, while spectralEntropyVar is the variance of the spectral entropy over the one-second intervals.**
- **(lowFrequentPowerAvg and lowFrequentPowerVar) These features give the band power in the δ band (1-3Hz) relative to the total power in the time series. The spectrogram used for these features is calculated based on the downsampled but un-filtered time series since the filter removes frequencies lower than 3Hz. The spectrogram is calculated over one-second intervals, and the power in the δ band divided by the power over all frequencies is then found. The feature lowFrequentPowerAvg is the**

mean over the one-second intervals of this ratio, and lowFrequentPowerVar is the variance over the one-second intervals.

Temporal

- **(skew1sAvg and skew1sVar) The skewness was calculated for one-second intervals of the time series of ICs. The feature skew1sAvg is the average over these one-second intervals and skew1sVar is the variance over these intervals. The feature skew1sAvg for 15 second intervals was one of the six final features included in the classifier described in Winkler, Haufe, and Tangermann (2011).**
- **(logRangeTemporalAvg and logRangeTemporalVar) The range (maximum value minus minimum value) was calculated for one-second intervals. The feature logRangeTemporalAvg is the average over these one-second intervals and logRangeTemporalVar is the variance.**
- **(kurtosisAvg and kurtosisVar) As for the two above features, the feature kurtosisAvg is the average of the kurtosis in one-second intervals and kurtosisVar is the variance of the kurtosis in one-second intervals. This was also used in Winkler, Haufe, and Tangermann (2011).**
- **(hurst1Avg, hurst2Avg, hurst3Avg, hurst1Var, hurst2Var and hurst3Var) We used the Matlab function *wfbmesti* in the Wavelet toolbox to get three different estimates of the Hurst exponent, which is a measure of the autocorrelation of a time series. These three estimates of the Hurst exponent are found for one-second intervals. The features hurst1Avg, hurst2Avg, and hurst3Avg are the averages over these intervals, and hurst1Var, hurst2Var, and hurst3Var are the variances over the intervals.**
- **(var1sAvg and var1sVar) Again, the variance is found in one-second intervals of the time series. The features var1sAvg and var1sVar are the average and variance over these intervals, respectively. This was also used in Winkler, Haufe, and Tangermann (2011).**
- **(maxFirstDerivAvg and maxFirstDerivVar) In each one-second interval, the maximum difference between consecutive values was found. The average over the intervals is maxFirstDerivAvg and the variance is maxFirstDerivVar. This was also used in Winkler, Haufe, and Tangermann (2011).**
- **(maxAmplAvg adn maxAmplVar) In each one-second interval, the maximum amplitude (maximum absolute value in that interval) was found. The average over these intervals is maxAmplAvg and the variance is maxAmplVar. This was also used in Winkler, Haufe, and Tangermann (2011).**
- **(timeEntropyAvg and timeEntropyVar) In each one-second interval, the entropy was found. The average over these intervals is timeEntropyAvg**

and the variance is timeEntropyVar. This was also used in Winkler, Haufe, and Tangermann (2011).

Acar, Z. A., S. Makeig, and G. Worrell. 2008. "Head Modeling and Cortical Source Localization in Epilepsy." ***Conf Proc IEEE Eng Med Biol Soc*** **2008: 3763–66.**

Bartels, G., Li-Chen Shi, and Bao-Liang Lu. 2010. "Automatic Artifact Removal from EEG - a Mixed Approach Based on Double Blind Source Separation and Support Vector Machine." In ***EMBC*****, 5383–86.**

Bigdely-Shamlo, N., K. Kreutz-Delgado, C. Kothe, and S. Makeig. 2013. "EyeCatch: Data-Mining over Half a Million EEG Independent Components to Construct a Fully-Automated Eye-Component Detector." In ***IEEE Engineering in Biology and Medicine Conference, Osaka, Japan*****.**

Citi, L., R. Poli, and C. Cinel. 2010. "Documenting, Modelling and Exploiting P300 Amplitude Changes Due to Variable Target Delays in Donchin's Speller." ***J Neural Eng*** **7 (5): 056006.**

Comani, S., D. Mantini, P. Pennesi, A. Lagatta, and G. Cancellieri. 2004. "Independent Component Analysis: fetal Signal Reconstruction from Magnetocardiographic Recordings." ***Comput. Methods Programs Biomed.*** **75 (2): 163–77.**

Delorme, A., and S. Makeig. 2004. "EEGLAB: an Open Source Toolbox for Analysis of Single-Trial EEG Dynamics Including Independent Component Analysis." ***J. Neurosci. Meth.*** **134: 9–21.**

Delorme, A., T. Sejnowski, and S. Makeig. 2007. "Enhanced Detection of Artifacts in EEG Data Using Higher-Order Statistics and Independent Component Analysis." ***Neuroimage*** **34 (4): 1443–49. doi:DOI: 10.1016/j.neuroimage.2006.11.004. \url{http://www.sciencedirect.com/science/article/B6WNP-4MNHY2V-4/2/93de9223a58e55c80f19ecdb50e8dcfe}.**

Di, L., N. Rao, C. H. Kuo, S. Bhatt, and V. Dogra. 2007. "Independent Component Analysis Applied to Ultrasound Speckle Texture Analysis and Tissue Characterization." ***Conf. Proc. IEEE Eng. Med. Biol. Soc.*** **2007: 6524–27.**

Erfanian, A., and A. Erfani. 2004. "ICA-Based Classification Scheme for EEG-Based Brain-Computer Interface: the Role of Mental Practice and Concentration Skills." ***Conf Proc IEEE Eng Med Biol Soc*** **1: 235–38.**

Fatourechi, M., A. Bashashati, R. K. Ward, and G. E. Birch. 2007. "EMG and EOG Artifacts in Brain Computer Interface Systems: A Survey." ***Clin. Neurophysiol.*** **118 (3): 480–94. doi:10.1016/j.clinph.2006.10.019. http://www.sciencedirect.com/science/article/pii/S1388245706015124.**

Gramann, K., T. Tollner, and H. J. Muller. 2010. "Dimension-Based Attention Modulates Early Visual Processing." *Psychophysiology* 47 (September): 968–78.

Halder, S., M. Bensch, J. Mellinger, M. Bogdan, A. Kubler, N. Birbaumer, and W. Rosenstiel. 2007. "Online Artifact Removal for Brain-Computer Interfaces Using Support Vector Machines and Blind Source Separation." *Comput. Intell. Neurosci.*, 82069.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York: Springer.

He, P., G. Wilson, and C. Russell. 2004. "Removal of Ocular Artifacts from Electro-Encephalogram by Adaptive Filtering." *Med. Biol. Eng. Comput.* 42 (3): 407–12.

Hyvärinen, A., and E. Oja. 2000. "Independent Component Analysis: algorithms and Applications." *Neural Networks* 13 (4-5): 411–30.

Jayalakshmi, T., and Dr.A. Santhakumaran. 2011. "Statistical Normalization and Back Propagation for Classification." *International Journal of Computer Theory and Engineering* 3 (1): 1793–8201.

Jung, T. P., S. Makeig, C. Humphries, T. W. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski. 2000. "Removing Electroencephalographic Artifacts by Blind Source Separation." *Psychophysiology* 37 (March): 163–78.

Kim, C. M., H. M. Park, T. Kim, Y. K. Choi, and S. Y. Lee. 2003. "FPGA Implementation of ICA Algorithm for Blind Signal Separation and Adaptive Noise Canceling." *IEEE Trans. Neural Netw.* 14 (5): 1038–46.

Klekowicz, H., U. Malinowska, A. J. Piotrowska, D. Wolynczyk-Gmaj, S. Niemcewicz, and P. J. Durka. 2009. "On the Robust Parametric Detection of EEG Artifacts in Polysomnographic Recordings." *Neuroinformatics* 7 (2): 147–60.

Kong, W., C. R. Vanderburg, H. Gunshin, J. T. Rogers, and X. Huang. 2008. "A Review of Independent Component Analysis Application to Microarray Gene Expression Data." *BioTechniques* 45 (5): 501–20.

LeVan, P., E. Urrestarazu, and J. Gotman. 2006. "A System for Automatic Artifact Removal in Ictal Scalp EEG Based on Independent Component Analysis and Bayesian Classification." *Clin. Neurophysiol.* 117 (4): 912–27.

López, Victoria, Alberto Fernández, Jose G. Moreno-Torres, and Francisco Herrera. 2012. "Analysis of Preprocessing Vs. Cost-Sensitive Learning for Imbalanced Classification. Open Problems on Intrinsic Data Characteristics." *Expert Systems with Applications* 39 (7): 6585–6608. doi:10.1016/j.eswa.2011.12.043. http://www.sciencedirect.com/science/article/pii/S0957417411017143.

Mantini, D., R. Franciotti, G. L. Romani, and V. Pizzella. 2008. "Improving MEG Source Localizations: an Automated Method for Complete Artifact Removal Based on Independent Component Analysis." *Neuroimage* 40 (1): 160–73.

Mognon, A., J. Jovicich, L. Bruzzone, and M. Buiatti. 2010. "ADJUST: An Automatic EEG Artifact Detector Based on the Joint Use of Spatial and Temporal Features." *Psychophysiology* 48 (2). Blackwell Publishing Inc: 229–40. doi:10.1111/j.1469-8986.2010.01061.x. http://dx.doi.org/10.1111/j.1469-8986.2010.01061.x.

Nolan, H., R. Whelan, and R. B. Reilly. 2010. "FASTER: Fully Automated Statistical Thresholding for EEG Artifact Rejection." *J. Neurosci. Methods* 192 (September): 152–62.

Onton, J. A., and S. Makeig. 2009. "High-Frequency Broadband Modulation of Electroencephalographic Spectra." *Front. Hum. Neurosci.* 3 (0): 12. doi:10.3389/neuro.09.061.2009. \url{http://www.frontiersin.org/Journal/Abstract.aspx?s=537&name=human neuroscience&ART_DOI=10.3389/neuro.09.061.2009}.

Sugiyama, Masashi, and Motoaki Kawanabe. 2012. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. 1st ed. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts, USA: MIT Press.

Tangermann, M., I. Winkler, S. Haufe, and B. Blankertz. 2009. "Classification of Artifactual ICA Components." *Int. J. Bioelectromagnetism* 11 (2): 110–14.

Tsai, D. M., and S. C. Lai. 2009. "Independent Component Analysis-Based Background Subtraction for Indoor Surveillance." *IEEE Trans. Image Process.* 18 (1): 158–67.

Ullsperger, M., and S. Debener, eds. 2010. "Simultaneous EEG and FMRI: Recording, Analysis, and Application." In 121–35. New York: Oxford University Press.

Viola, F. C., J. Thorne, B. Edmonds, T. Schneider, T. Eichele, and S. Debener. 2009. "Semi-Automatic Identification of Independent Components Representing EEG Artifact." *Clin. Neurophysiol.* 120 (5): 868–77.

Winkler, I., S. Haufe, and M. Tangermann. 2011. "Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals." *Behav Brain Funct* 7 (August): 30.

Ypparila, H., S. Nunes, I. Korhonen, J. Partanen, and E. Ruokonen. 2004. "The Effect of Interruption to Propofol Sedation on Auditory Event-Related Potentials and Electroencephalogram in Intensive Care Patients." *Crit Care* 8 (6): R483–90.

Zadrozny, Bianca, John Langford, and Naoki Abe. 2003. "Cost-Sensitive Learning by Cost-Proportionate Example Weighting." In *Proceedings of the Third IEEE International Conference on Data Mining*, 435. ICDM '03. Washington, DC, USA: IEEE Computer Society. http://dl.acm.org/citation.cfm?id=951949.952181.