Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Autoreject: Automated artifact rejection for MEG and EEG data

Mainak Jas[a,*], Denis A. Engemann[b,c,d,1], Yousra Bekhti[a], Federico Raimondo[d,e,f,g], Alexandre Gramfort[a,1]

[a] LTCI, Télécom ParisTech, Université Paris-Saclay, France
[b] Parietal project-team, INRIA Saclay - Ile de France, France
[c] Cognitive Neuroimaging Unit, Neurospin, CEA DSV/I2BM, INSERM, Université Paris-Sud, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France
[d] Institut du Cerveau et de la Moelle épinière, ICM, PICNIC Lab, F-75013, Paris, France
[e] Laboratorio de Inteligencia Artificial Aplicada, Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
[f] CONICET, Argentina
[g] Sorbonne Universités, UPMC Univ Paris 06, Faculté de Médecine Pitié-Salpêtrière, Paris, France

ARTICLE INFO

Keywords:
Magnetoencephalography (MEG)
Electroencephalogram (EEG)
Preprocessing
Statistical learning
Cross-validation
Automated analysis
Human Connectome Project (HCP)

ABSTRACT

We present an automated algorithm for unified rejection and repair of bad trials in magnetoencephalography (MEG) and electroencephalography (EEG) signals. Our method capitalizes on cross-validation in conjunction with a robust evaluation metric to estimate the optimal peak-to-peak threshold – a quantity commonly used for identifying bad trials in M/EEG. This approach is then extended to a more sophisticated algorithm which estimates this threshold for each sensor yielding trial-wise bad sensors. Depending on the number of bad sensors, the trial is then repaired by interpolation or by excluding it from subsequent analysis. All steps of the algorithm are fully automated thus lending itself to the name *Autoreject*.

In order to assess the practical significance of the algorithm, we conducted extensive validation and comparisons with state-of-the-art methods on four public datasets containing MEG and EEG recordings from more than 200 subjects. The comparisons include purely qualitative efforts as well as quantitatively benchmarking against human supervised and semi-automated preprocessing pipelines. The algorithm allowed us to automate the preprocessing of MEG data from the Human Connectome Project (HCP) going up to the computation of the evoked responses. The automated nature of our method minimizes the burden of human inspection, hence supporting scalability and reliability demanded by data analysis in modern neuroscience.

Introduction

Magneto-/electroencephalography (M/EEG) offer the unique ability to explore and study, non-invasively, the temporal dynamics of the brain and its cognitive processes. The M/EEG community has only recently begun to appreciate the importance of large-scale studies, in an effort to improve replicability and statistical power of experiments. This has given rise to the practice of sharing and publishing data in open archives (Gorgolewski and Poldrack, 2016). Examples of such large electrophysiological datasets include the Human Connectome Project (HCP) (Van Essen et al., 2012; Larson-Prior et al., 2013), the Physiobank (Goldberger et al., 2000), the OMEGA archive (Niso et al., 2016) and Cam-CAN (Taylor et al., 2015). A tendency towards ever-growing massive datasets as well as a shift towards common standards for accessing these databases (Gorgolewski et al., 2016; Bigdely-Shamlo et al.,) is clearly visible. The UK Biobank project (Ollier et al., 2005) which currently hosts data from more than 50,000 subjects is yet another example of this trend.

This has however, given rise to new challenges including automating the analysis pipeline (Gorgolewski and Poldrack, 2016). Automation will not only save time, but also allow scalable analysis and reduce the barriers to reanalysis of data, thus facilitating reproducibility. Engemann and Gramfort (2015) have recently worked towards more automation in M/EEG analysis pipelines by considering the problem of covariance estimation, a step commonly done prior to source localization. Yet, one of the most critical bottlenecks that limits the reanalysis of M/EEG data remains at the preprocessing stage with the annotation and rejection of artifacts. Despite being so fundamental to M/EEG analysis given how easily such data can be corrupted by noise and artifacts, there is currently

[*] Correspondence to: 46 Rue Barrault, Télécom ParisTech, Université Paris-Saclay, France.
E-mail address: mainak.jas@telecom-paristech.fr (M. Jas).

[1] Equal contributions

https://doi.org/10.1016/j.neuroimage.2017.06.030
Received 23 December 2016; Accepted 14 June 2017
Available online 20 June 2017

no consensus in the community on how to address this particular issue.

In the presence of what we will refer to as *bad* data, various data cleaning strategies have been employed. A first intuitive strategy is to exclude bad data from analysis, to *reject* it. While this approach is very often employed, for example, because data cleaning is time consuming, or out of reach for practitioners, it leads to a loss of data that are costly to acquire. This is particularly the case for clinical studies, where patients have difficulties staying still or focusing on the task (Cruse et al., 2012; Goldfine et al., 2013), or even when babies are involved as subjects (Basirat et al., 2014).

When working with M/EEG, the data can be bad due to the presence of bad sensors (also known as channels[2]) and bad trials. A trial refers here to a data segment whose location in time is typically related to an experimental protocol. But here we will also call trial any data segment even if it is acquired during a task-free protocol. Accordingly, a bad trial or bad sensor is one which contains bad data. Ignoring the presence of bad data can adversely affect analysis downstream in the pipeline. For example, when multiple trials time-locked to the stimulation are averaged to estimate an evoked response, ignoring the presence of a single bad trial can corrupt the average. The mean of a random vector is not robust to the presence of strong outliers. Another example quite common in practice, both in the case of EEG and MEG, is the presence of a bad sensor. When kept in the analysis, an artifact present on a single bad sensor can spread to other sensors, for example due to spatial projection. This is why identifying bad sensors is crucial for data cleaning techniques such as the very popular Signal Space Separation (SSS) method (Taulu et al., 2004). Frequency filtering (Widmann et al., 2015) can often suppress many low frequency artifacts, but turns out to be insufficient for broadband artifacts. A common practice to mitigate this issue is to visually inspect the data using an interactive viewer and mark manually, the bad sensors and bad segments in the data. Although trained experts are very likely to agree on the annotation of bad data, their judgement is subject to fluctuations and cannot be repeated. Their judgement can also be biased due to prior training with different experimental setups or equipments, not to mention the difficulty for such experts to allocate some time to review the raw data collected everyday.

Luckily, popular software tools such as Brainstorm (Tadel et al., 2011), EEGLAB (Delorme and Makeig, 2004), FieldTrip (Oostenveld et al., 2011), MNE (Gramfort et al., 2013) or SPM (Litvak et al., 2011) already allow for the rejection of bad data segments based on simple metrics such as peak-to-peak signal amplitude differences that are compared to a manually set threshold value. When the peak-to-peak amplitude in the data exceeds a certain threshold, it is considered as bad. However, while this seems quite easy to understand and simple to use from a practitioner's standpoint, this is not always convenient. In fact, a good peak-to-peak signal amplitude threshold turns out to be data specific, which means that setting it requires some amount of trial and error.

The need for better automated methods for data preprocessing is clearly shared by various research teams, as the literature of the last few years can confirm. On the one hand, are pipeline-based approaches, such as Fully Automated Statistical Thresholding for EEG artifact rejection (FASTER by Nolan et al. (2010)) which detect bad sensors as well as bad trials using fixed thresholds motivated from classical Gaussian statistics. Methods such as PREP (Bigdely-Shamlo et al., 2015), on the other hand, aim to detect and clean the bad sensors only. Unfortunately, they do not offer any solution to reject bad trials. Other methods are available to solve this problem. For example, the Riemannian Potato (Barachant et al., 2013) technique can identify the bad trials as those where the covariance matrix lies outside of the "potato" of covariance matrices for good trials. By doing so, it marks trials as bad but does not identify the sensors causing the problem, hence not offering the ability to repair them. It appears that practitioners are left to choose between different methods to reject trials or repair sensors, whereas they are in fact intricately related problems and must be dealt with together.

Robust regression (Diedrichsen and Shadmehr, 2005) also deals with bad trials using a weighted average which mitigates the effect of outlier trials. Trials with artifacts end up with low contributions in the average. A related approach that is sometimes employed to ignore outlier trials in the average is the trimmed mean as opposed to a regular mean. The trimmed mean is a compromise between the mean which offers a high signal-to-noise ratio (SNR) but can be corrupted by outliers, and the median which is immune to outliers of extreme amplitudes but has a low SNR as it involves no averaging. Of course, neither of these strategies are useful when analyses have to be conducted on single trials. Another approach, which is also data-driven, is Sensor Noise Suppression (SNS) (De Cheveigné and Simon, 2008). It removes the sensor-level noise by spatially projecting the data of each sensor onto the subspace spanned by the principal components of all the other sensors. This projection is repeated in leave-one-sensor-out iterations so as to eventually clean all the sensors. In most of these methods, however, there are parameters which are somewhat dataset dependent and must therefore be manually tuned.

We therefore face the same problem in automated methods as in the case of semi-automated methods such as peak-to-peak rejection thresholds, namely the tuning of model parameters. In fact, setting the model parameters is even more challenging in some of the methods when they do not directly translate into human-interpretable physical units.

This led us to adopt a pragmatic approach in terms of algorithm design, as it focuses on the tuning of the parameters that M/EEG users presently choose manually. The goal is, not only to obtain high quality data but also to develop a method which is transparent and not too disruptive for the majority of M/EEG users. A first question we address below is: can we improve peak-to-peak based rejection methods by automating the process of trial and error? In the following section, we explain how the widely-known statistical method of cross-validation (see Fig. 1 for a preview) in combination with Bayesian optimization (Snoek et al., 2012, Bergstra et al., 2011) can be employed to tackle the problem at hand. We then explain how this strategy can be extended to set thresholds separately for each sensor and mark trials as bad when a large majority of the sensors have high-amplitude artifacts. This process closely mimics how a human expert would mark a trial as bad during visual inspection.

In the rest of the paper, we detail the internals of our algorithm, compare it against various state-of-the-art methods, and position it conceptually with respect to these different approaches. For this purpose, we make use of qualitative visualization techniques as well as quantitative reports. In a major validation effort, we take advantage of cleaned up

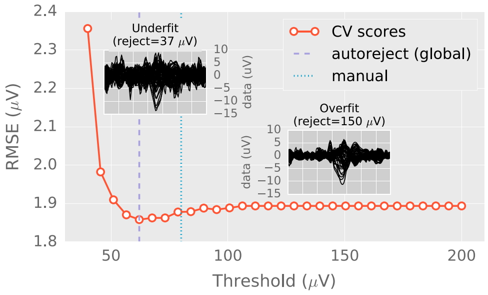

Fig. 1. Cross-validation error as a function of peak-to-peak rejection threshold on one EEG dataset. The root mean squared error (RMSE) between the mean of the training set (after removing the trials marked as bad) and the median of the validation set was used as the cross-validation metric (*Autoreject (global)*). The two insets show the average of the trials as "butterfly plots" (each curve representing one sensor) for very low and high thresholds. For low thresholds, the RMSE is high because most of the trials are rejected (underfit). At high thresholds, the model does not drop any trials (overfit). The optimal data-driven threshold (*autoreject, global*) with minimum RMSE is somewhere in between. It closely matches the human threshold.

[2] They are not necessarily equivalent in the case of a bipolar montage in EEG. However, for the sake of simplicity, we shall use these terms interchangeably in this work.