Lossless EEG pre-processing pipeline for objective signal quality assessment using data annotation and ICA

James A. Desjardins, MA[1,2*]
Stefon J.R. van Noordt, PhD[1,3]
Sidney J. Segalowitz, PhD[4]
Mayada Elsabbagh, PhD[1,3]

[1] *Montreal Neurological Institute, McGill University, Montréal, Canada*

[2] *SHARCNET, Compute Ontario, Compute Canada*

[3] *Douglas Mental Health University Institute, Verdun, Canada*

[4] *Cognitive and Affective Neuroscience Lab, Brock University, St. Catharines, ON Canada*

*Corresponding author

james.desjardins@computeontario.ca
stefonv0@gmail.com
sid.segalowitz@brocku.ca
mayada.elsabbagh@mcgill.ca

Keywords

EEG, pre-processing, high performance computing, ICA,

Highlights

- Standardized approach to pre-processing electroencephalography data that produces signal quality annotations and ICA decompositions, while minimizing data rejection and manipulation
- Using pipeline annotations to remove ICA isolated artifacts results in a significant increase in data retention as well as ERP effect precision
- Pipeline includes extensive data characterization and interactive quality control
- Procedures are compatible with high performance computing clusters

Abstract

The methods available for pre-processing EEG data are rapidly evolving as researchers gain access to vast computational resources; however, the field currently lacks a set of standardized approaches for data characterization, efficient interactive quality control review procedures, and large-scale automated processing that is compatible with High Performance Computing (HPC) resources. In this paper we describe an infrastructure for the development of standardized procedures for automated pre-processing of EEG data. Our pipeline incorporates several methods to isolate cortical signal from noise, maintain maximal information from raw recordings and provide comprehensive quality control and data visualization. In addition, batch processing procedures are integrated to scale up analyses for processing hundreds or thousands of data sets using HPC clusters. We demonstrate here that by using the lossless pipeline's signal quality annotations, significant increase in data retention is achieved when applying subsequent post-processing channel and segment rejection procedures. Further, we demonstrate that the increase in data retention is not achieved at the expense of an ERP effect, but rather a significant increase in precision around the ERP effect is measured.

1 Introduction

1.1 Overview of EEG data standardization

Electroencephalography (EEG) provides a wealth of data about neural dynamics and circuits, as well as their relationships to the underpinnings of information processing and behavior. In an attempt to isolate cortical signal from noise in EEG recordings, many analysis strategies include a number of pre-processing procedures that typically result in considerable data removal (e.g. channel and/or time period removal) as well as data manipulation (e.g. spectral filtering, eye artifact regression, Independent Component Analysis (ICA) artifact correction). In addition, these processing steps vary widely across laboratories and studies. As a consequence, it is common that EEG analyses not only discard a substantial amount of potentially meaningful information about cortical dynamics, but they also introduce challenges for cross-lab and cross-pipeline replications, processing archived data, or combining data across studies or experimental conditions. The methods available for signal processing, modeling, and data analysis are rapidly evolving as researchers gain access to vast computational resources; however, the field currently lacks a set of standardized approaches for data characterization, efficient interactive quality control review procedures, and large-scale automated processing that is compatible with High Performance Computing (HPC) resources.

Several groups have contributed to automated and standardized EEG pre-processing initiatives. For example, the work of Bigdely-Shamlo and colleagues (Bigdely-Shamlo, Cockfield, Makeig, Rognon, La Valle, Miyakoshi, & Robbons, 2015) has led to the development of a comprehensive infrastructure to transform data into successive standard level formats, including the integration of all the information that would be required to annotate raw data and prepare it for subsequent processing. This group has also introduced the PREP pipeline, which focuses on issues related to noisy channels and their impact on referencing (Bigdely-Shamlo, Mullen, Kothe, Su, & Robbins, 2015). A unique feature of PREP is the use of a robust referencing scheme and detailed reports on signal quality and transformations for individual EEG datasets. Once raw EEG data sets have been harmonized to a common standard, there are several pre-processing pipelines available, which vary in terms of artifact detection methods and the specific combination of procedures and data transformations. A common goal of pre-processing is to identify signals and artifacts in channels, components, or time varying activations, in order to isolate reliable cortical signal in the EEG. It is beyond the scope of the current paper to review or compare the many pre-processing options available for EEG; however, we briefly describe some existing methods to give context for our pre-processing pipeline, which moves raw data into a state that is ready for hypothesis testing and study-level analyses.

Comprehensive pipelines, which complement our approach, have recently been made available, including The Harvard Automated Processing Pipeline for EEG (HAPPE; Gabard-Durnam, Leal, Wilkinson, & Levin, 2018) and the Batch Electroencephalography Automated Processing Platform (BEAPP; Levin, Leal, Gabard-Durnam, & O'Leary, 2018). HAPPE offers a standardized automated approach to deal with EEG recordings that vary in signal quality, and may be especially useful for EEG recordings that are commonly contaminated with high levels of artifact (e.g., data from young children, neurodevelopmental populations, psychiatric populations), and contain other constraints such as a short recording duration. The HAPPE pipeline takes raw data through a sequence of cleaning procedures to enhance data quality and ensure suitability for subsequent post-processing. Compared to several other pipelines, HAPPE offers optimal trade-offs in terms of the signal-to-noise ratio of the pre-processed EEG. Whereas HAPPE describes a specific sequence of EEG processing steps, other software, such as BEAPP (Levin, Leal, Gabard-Durnam, & O'Leary, 2018), focuses on the flexibility to choose from various procedures. Together, HAPPE and BEAPP are examples of standardized, yet modular, pipelines that can be used for various samples and multi-site studies that combine data from different acquisition machines or across labs. A fundamental difference in our approach is the minimization of data removal and signal manipulation during pre-processing; instead, our procedures involve building extensive annotations about signal quality that can easily be visualized with the raw data and modified during quality control review.

1.2 Brain Imaging Data Structure for EEG (BIDS-EEG)

Given the multitude of data collection procedures, equipment, data streams, event marking, and software, it is critical for researchers to utilize some form of standardization when taking information from raw and metadata files in order to implement a universal pre-processing trajectory that is transferable across studies. The BIDS standard provides a set of guidelines that allow researchers (and processing pipelines) to expect where to find all relevant information that is common to EEG data collection and analysis. The BIDS standard offers researchers a way to store transformed data in successive states or “derivatives” (Gorgolewski et al., 2016).

Recent efforts to establish BIDS-EEG provide the reference point for sharing and integrating multi-site data while also simplifying the development of processing pipeline tools that aim to establish standardized data states (e.g., robust ICA). The lossless pipeline expects a BIDS compliant data set as input and outputs a BIDS compliant “derivative” state of the data containing extensive annotations about signal quality properties. Because of the compliance in input and output to the BIDS standard, although the current tool implementation is developed in EEGLAB (Delorme & Makeig, 2004), the Lossless state of data outputs are agnostic to software tools.

1.3 Overview and goals of the lossless pipeline

The primary goals of the lossless pipeline are to (1) maximally isolate cortical signal from the various forms of noise contained in EEG while also (2) maintaining maximal information from the raw recording throughout the process, (3) remaining generalizable to various EEG recording parameters and constraints, (4) being replicable across sites and projects, and (5) being scalable to large sample sizes. The output of the lossless pipeline is the minimally altered original continuous data (all scalp channels and all time points, except for the final moments of the recording that may be pruned to the nearest window duration [e.g., one second]) with the only data manipulations being an interpolated average re-reference, and optional spectral filters (e.g., typically a 1 Hz high pass and a line noise notch filter if required). To the original EEG file the lossless pipeline adds multiple Independent Component Analysis (ICA) decompositions and annotations regarding various signal properties associated with time points, scalp channels, and Independent Components (ICs). These outputs of the automated pre-processing pipeline are then used in an optimized interactive quality control review procedure where each data file is inspected and annotations that describe the isolation of signal and noise in the recording are potentially modified by an expert reviewer. This lossless approach to establishing a standardized data state for pre-processed EEG introduces minimal constraints on post-processing procedures.

The lossless pipeline takes EEG data from a BIDS-compliant raw state, using EEGLAB, and passes it through a set of maximally automated procedures for producing a replicable ICA decomposition, and implementing an optimized interactive quality control procedure. Although standardization and objective automation are important issues in EEG research, another critical component is some form of quality control review (particularly in the context of ICA decompositions). This quality control takes the form of comprehensive annotation of data properties accumulated during the processing pipeline so that expert reviewers can visualize and modify the decisions regarding data quality. The lossless data state solves this issue by including algorithms that flag various properties of the EEG data (as described below) and generate all of the annotation information that is required for appropriate interactive quality control . Following the quality control check, the data in the lossless state have channels flagged (based on several criteria), time periods flagged (based on several criteria), and classified independent components (ICs). The following sections describe sequentially the steps in the lossless pipeline that are applied to raw EEG data files. Finally, we provide an example to demonstrate the impact that the lossless pipeline has on data retention, as well as the precision of an ERP effect measurement.

2 Methods

2.1 EEG data set

In order to assess the performance of the lossless pipeline in retaining signal in EEG recording, we examine the EEG data from 105 7-month old infants taken from the sample described in Elsabbagh et al. (2015). Of the original 105 EEG sessions 5 of the recordings were rejected from the analysis due to a combination of short recording time and large noise contamination. The data were acquired using a 129-channel saline solution EGI sensor net while infants were presented images of faces and controlled phase scrambled noise stimuli on a computer monitor. The recording durations varied from 93 to 893 seconds with a mean of 450 and standard deviation of 151 seconds. This data set was selected as an example of an ERP protocol having heightened constraints on signal quality at acquisition time. EEG data acquisition on infants, particularly those with neurodevelopmental vulnerabilities, is typically limited to short recording durations and is likely to be contaminated with artifacts from movement behavior.

2.2 Software tool implementation

The lossless pipeline performs a cascade of signal assessment procedures described in details below. The lossless pipeline is made available as a git repository hosted at https://github.com/BUCANL/BIDS-Lossless-EEG. The git repository includes all of the scripts and toolboxes required to run the pipeline. The current implementation is developed within the EEGLAB software framework. Being developed in EEGLAB the lossless pipeline has some dependencies on Matlab for graphical user interface (GUI) usage. Other than the GUI usage, the scripted procedures have been made compatible with GNU Octave (https://www.gnu.org/software/octave/) which is freely available and can be easily deployed across multiple compute nodes of compute clusters. The two GUI requirements of running the lossless pipeline are for job submission to compute cluster and the interactive quality control procedure. The job submission is achieved with the Batch-Context extension for EEGLAB (https://github.com/BUCANL/Batch-Context). The interactive quality control and data property annotation tools are implemented in the Vised-Marks extension for EEGLAB (https://github.com/BUCANL/Vised-Marks). Besides containing all of the software tools for running the lossless pipeline with EEGLAB in Matlab and Octave the github wiki contains documentation for the resources as well as links to extensive tutorials for replicating the analysis on a demo data set of the recordings used in Desjardins and Segalowitz (2013).

2.3 Common criterion function

A common criterion function for classification is used throughout the pipeline when signals or time points are being flagged for a specific property. We describe this flagging method first here before describing the sequence of measures that it is applied to during the pipeline in the following sections. This classification method accepts an input measure in the form of a two-dimensional numeric array that has signals (scalp channels or independent component activations) as rows (Y axis) and time points as columns (X axis). The time points typically consist of the measure calculated on a specific interval of time (e.g., the continuous data are epoched into one-second non-overlapping time windows). Measures are assessed against the criterion function to determine whether signals (rows) or time points (columns) are consistent outliers on the measure. The criterion function examines the distributions of values along one dimension, looking for outliers, and then classifies the signal or time point based on the consistency of it being identified as an outlier. This is a process of identifying outlier values along one dimension of the input measure array and then examining the consistency of the outlier locations along the second dimension.

In the case of high voltage variance time periods, the criterion function calculates the distribution of values across time windows for each channel and then identifies outlying time points for each channel. For example, outlying time windows for each channel can be identified by calculating the median as well as an upper and lower quantile of the input measure for the channel, then setting a threshold of a given inter-quantile range (e.g., 6 times the distance between the median and the inner edge of the upper or lower quantile) to flag the outliers for the current channel. After identifying the outlying time windows for each channel, the flags are tallied across all channels and a criterion is set to identify the time windows (e.g., any time window that was flagged as an outlier in 20% of the channels). This process is illustrated in Figure 1 where panel A (top left) depicts the input measure of voltage standard deviation for each channel (Y axis) and time window (X axis). Panel B (top right) illustrates the median (dark grey dots), 30% and 70% quantiles (light grey dots), critical values as a distance from the median that is 6 times the median to quantile range (blue dots) and outliers (red “+”). In panel C (middle left) the outliers identified in panel B are placed in their channel by time locations (white points). Panel D (bottom left) illustrates the percentage of channels that where flagged as outliers in each time window (blue line), and time windows that surpass the critical value of 20% (red line) are identified as being an artifact time windows based on the input measure of voltage variance. A flagged time window in this case is a time window that is unlike other time windows in regards to the input measure for a critical percentage of channels. Put another way, a flagged time window is when too many channels are unlike themselves on the input measure in relation to other time points.

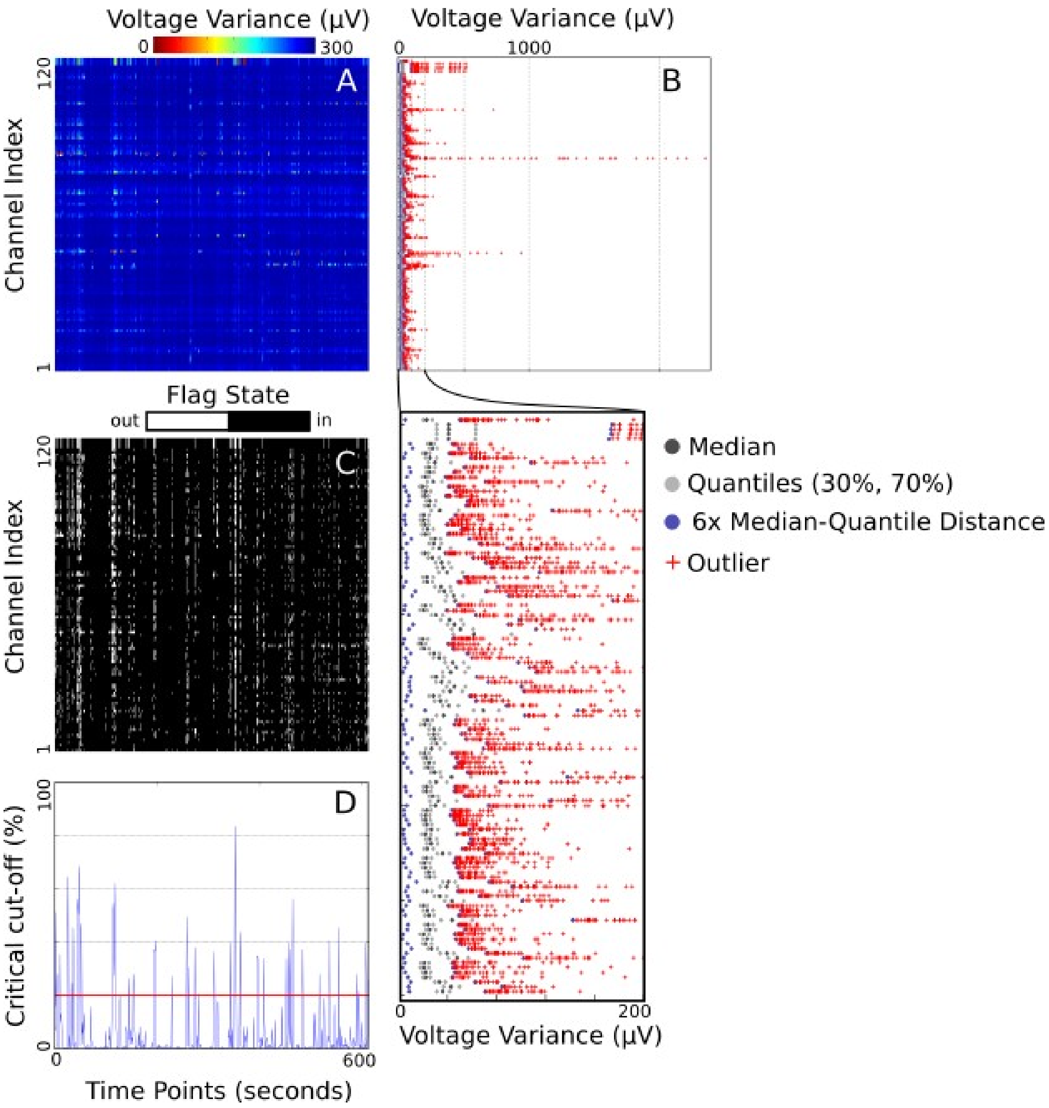

Figure 1. *Criterion function time outlier flagging based on voltage variance across time.*

Similar to the identification of time windows, the identification of channels can be accomplished by a rotation of the dimension along which the criterion function operates. Like the time window flagging criteria, a flagged channel is a channel that is unlike other channels in regards to the input measure for a critical percentage of time windows. For example, in Figure 2 the standard deviation of scalp channel voltage input array is used to identify channels where a threshold time window of unusual voltages across channels is reached. In panel A (top left) there is another example similar to Figure 1 panel A, illustrating the input measure of scalp channel voltage variance for each channel (Y axis) and calculated on each one second time interval (X axis). Panel B (bottom left) illustrates the outliers for each time window, identified by obtaining the median and quantiles for each time window across channels. In panel C (top middle) the outliers are plotted in their channel by time array locations. Panel D (top right) illustrates the percentage of time windows where each channel was flagged as an outlier (blue points), and channels that surpass the critical value of 20% (red line) are identified as being artifact-laden channels based on the input measure of voltage variance.

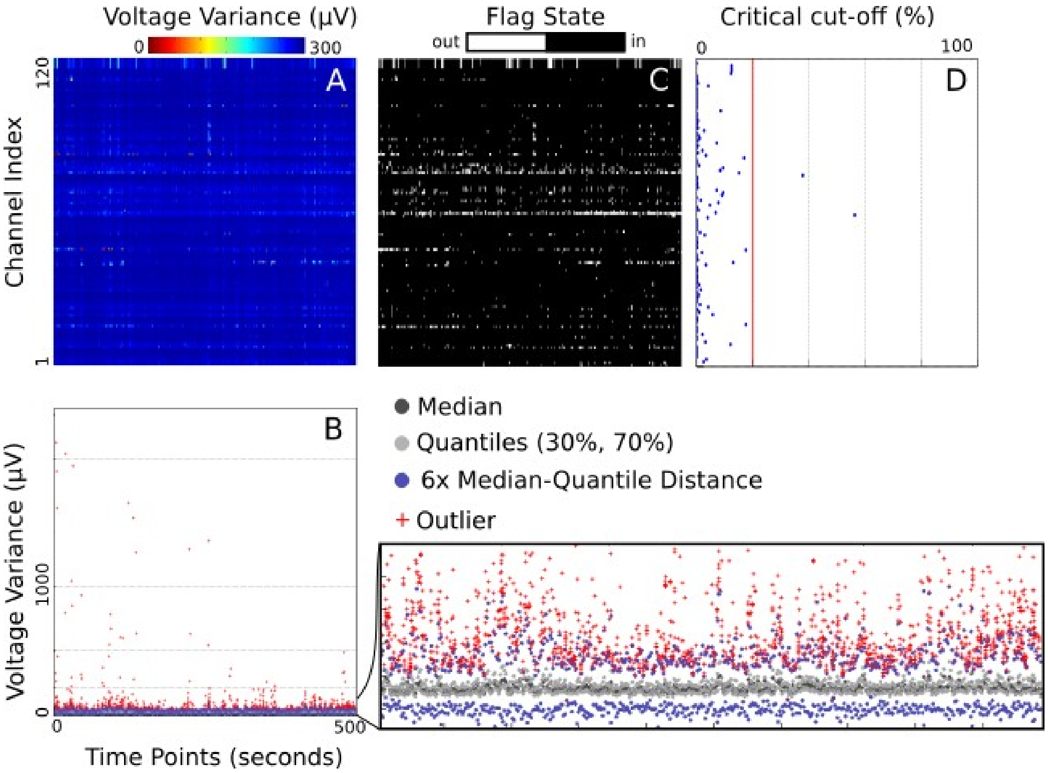

Figure 2. *Criterion function channel outlier flagging based on voltage variance across channels.*

2.4 Sequence of lossless pipeline procedures

In the following sections is a detailed description of the processing procedures illustrated in the Figure 3 lossless pipeline schematic. The pipeline is described in chronological order where section headers are labeled in correspondence with the processing blocks from the pipeline schematic. The lossless pipeline schematic lays out the signal quality assessment units chronologically from top to bottom, starting with 1- “scalp signal”, 2- “Initial ICA signal”, then 3- “final ICA signal”. The specific procedures included in each of the signal quality assessment units is described in detail below. At each stage the scripts use a common criterion function (described above) to assess signals or time periods on various data quality measures. The scripts are divided into units that execute sequentially on each data file. The procedures are divided into units based on the various compute resource demands that each stage requires (e.g., scalp signal assessment requires more memory, while Adaptive Mixture Independent Component Analysis (AMICA; Palmer, Kruetz-Delgado, & Makeig, 2011) requires more cores, etc.). The junction between the units of the scripts also represents logical restart or inspection points for executing the pipeline. For example, if a user chooses to run the entire pipeline in a semi-automated way, they could run the first stage of scalp signal assessment, then perform a visual inspection and annotation modification prior to submitting the data to the initial AMICA model. Further, considering that there are substantial computational resources consumed during the execution of the pipeline (including four multi-core AMICA MPI executions) this allows for efficient checkpoint restarting of the process if it is prematurely terminated.

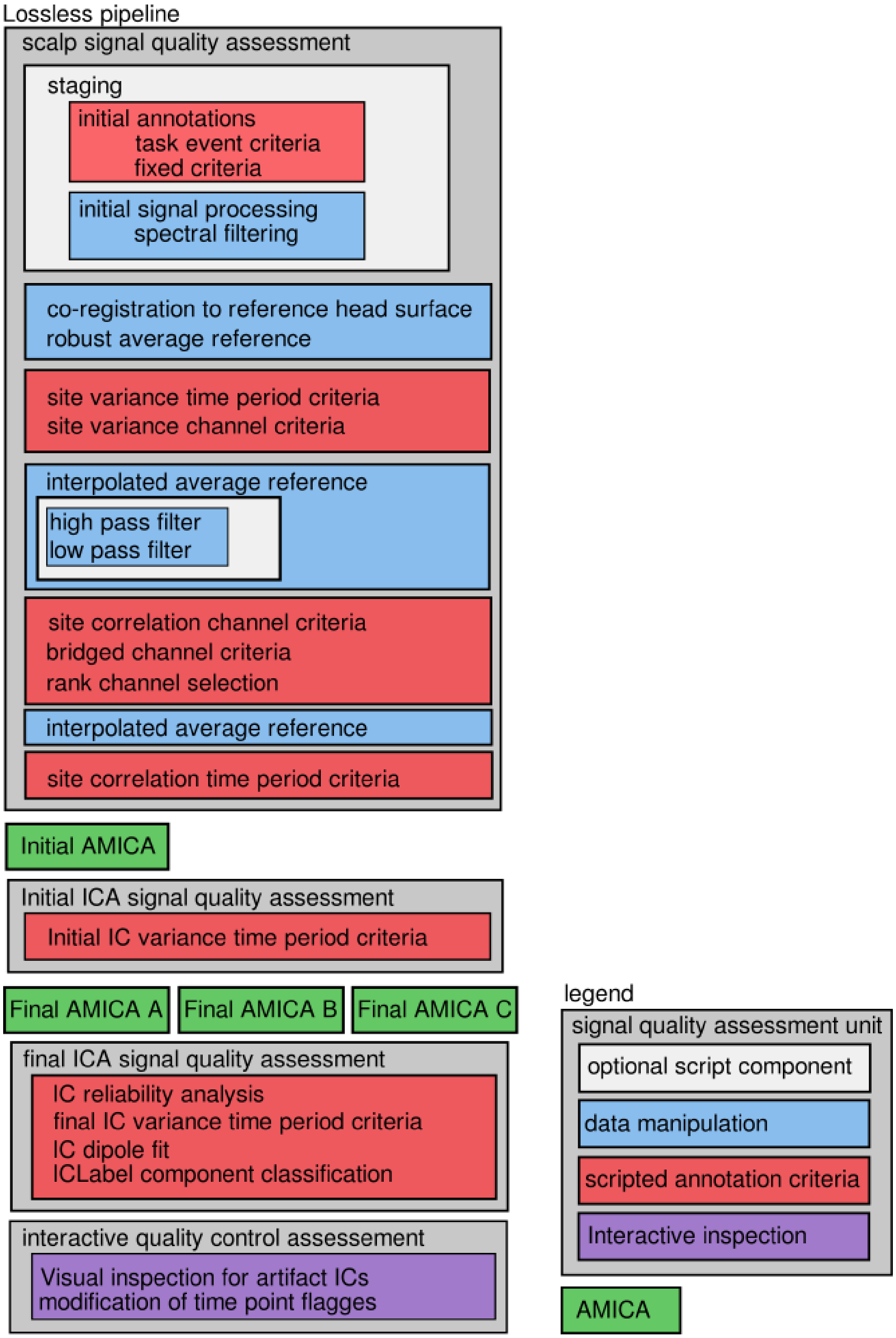
Lossless pipeline
scalp signal quality assessment
staging
initial annotations
task event criteria
fixed criteria
initial signal processing
spectral filtering
co-registration to reference head surface
robust average reference
site variance time period criteria
site variance channel criteria
interpolated average reference
high pass filter
low pass filter
site correlation channel criteria
bridged channel criteria
rank channel selection
interpolated average reference
site correlation time period criteria
Initial AMICA
Initial ICA signal quality assessment
Initial IC variance time period criteria
Final AMICA A
Final AMICA B
Final AMICA C
final ICA signal quality assessment
IC reliability analysis
final IC variance time period criteria
IC dipole fit
ICLabel component classification
interactive quality control assessment
Visual inspection for artifact ICs
modification of time point flagges
legend
signal quality assessment unit
optional script component
data manipulation
scripted annotation criteria
Interactive inspection
AMICA

Figure 3. *Lossless pipeline process sequence schematic.*

2.4.1 Scalp signal quality assessment

In order to achieve the goal of producing a reliable ICA decomposition, the lossless pipeline begins by assessing scalp channel properties in order to flag channels and time points that do not appear to be consistent with a set of spatially fixed projections. Specifically, in order to achieve a reliable ICA decomposition the voltage fluctuations at the scalp should be representative of a set of field projections that are picked up by many channels and the relative influence of each field projection across channels is stationary (e.g., the field projection from eye blinks affects channels in a way that has a topographical projection that does not change over time). This initial scalp signal quality assessment unit attempts to use the recording signals to identify channels and time points that do not have voltage fluctuations that are consistent with spatial stationarity.

The lossless pipeline begins by loading BIDS-compliant raw data files and adds a time annotation field *init_ind* that stores the data point index at this initial stage of the processing. This *init_ind* annotation can be used in subsequent processes to identify the initial positioning in the file of time points even if time intervals are removed from the data array. Subsequently, automated robust cleaning of the scalp data is performed by flagging channels and time that should not be given as input to the AMICA modeling.

Staging

Once the raw file is imported into the EEGLAB EEG structure, optional initializing steps can be performed prior to standardized quality assessment. Typical staging steps include project specific script that identify time periods in the recording that should not be included in standard data quality assessment measures.

Staging task event criteria

The Vised-Marks extension for EEGLAB provides tools for marking periods of time based on task event markers in the files. This process of marking time based on task event properties allows the signal quality measures to ignore periods of time when the participants may have been outside of the experimental tasks (e.g., during task instructions, during a break, when they may have moved around). This initial marking of channels and time points can also be performed interactively in a scroll plot if desired.

In the case of the current test data set no task event criteria annotations were used in the staging process. This is because the recordings of this test data set were already of short duration and it was important to allow as much of these short recordings as possible to be included in the ICA modeling. Further, with infants the task markers have relatively little bearing on the movement behavior of the participant making it equally likely that movement artifacts will occur during task intervals and break intervals.

Staging fixed criteria based on study parameter estimates

During the staging of the data it may also be desirable to apply data quality criteria annotation based on group level parameter estimates. The standard automated scripts of the lossless pipeline use distributions of values within each single data file in order to classify portions of data that are outliers. In the case of large portions of a data file containing non-representative signal, the pipeline may not be able to properly classify artifacts. For example, if the pipeline receives a file in which there are large non-stationary movement artifacts in the majority of time points of the recording, it would not be able to make fine grained distinctions about artifacts as the measure distributions (although robust to some extreme noise) would be contaminated by extreme values making up too much of the distribution. In data sets where large portions of the files contain substantial artifacts it is recommended to perform parameter estimate diagnostics on the full set of files to identify parameters that can be used initially to isolate large artifacts from the data, even if these artifacts are present in the majority of time points in a data file. One example of this is to estimate an acceptable standard deviation of voltage for each channel over all one-second epochs in all the data files in a study. Calculating a robust outlier estimate on each data file, and then taking the median outlier cutoff across all data files is a good strategy for initially isolating extreme artifacts that may be present in the majority of time points in a data file. While this standard deviation of voltage outlier estimate may change from data set to data set, measures such as standard deviation of voltage are intended to be as generalizable across studies as possible for classifying extreme periods of artifact contamination.

In the case of this test data set, a fixed criteria of 50µV was determined to be a reliable outlier cutoff for the criterion function based on an analysis of voltage variances during one second intervals in all the recording sessions from the test set. This 50µV outlier criteria was used in a criterion function calculation for both channel flagging (channel annotation labeled “ch_s_sd”) as well as time interval flagging (time annotation labeled “ch_s_sd”).

Staging spectral filtering

Performing the study level parameter estimate diagnostics provides the opportunity to examine aspects of the data files that may affect the performance of the pipeline. For example, if there is a very large variation across files, it is valuable at this initial stage to examine outlier files to determine the source of the anomalies. In many cases there can be a source of variance across files that may be handled within the staging script to make the files of a study more comparable to each other. For example, large variance from file to file in terms of line noise amplitude may lead to including a notch filter in the staging script to initially make all files more similar to one another in relation to a known artifact.

In the case of this test data set two spectral filters were applied to help account for large artifact voltage variability across recordings. The first filter applied to this data set was a 1Hz high pass filter in order to help account for the large variance in movement artifacts across recordings. The second spectral filter applied to the data was a 49Hz-51Hz notch filter to attenuate the large variance in line noise contamination across recordings.

Co-registration to reference head surface

The first data manipulation of the standard signal quality assessment procedures is the co-registration of the scalp channel coordinates to a standard head surface. This is an important step used to equate the re-referencing procedure and make the re-referencing site independent of the project’s channel montage. Throughout the lossless pipeline, any time that channels are identified as problematic (e.g., extreme voltage variance, low correlation between neighboring channels, bridged channels), the data are re-referenced to an interpolated average site, excluding the flagged channels. The interpolated average reference procedure uses all remaining channels from the original montage that have not been flagged as problematic to interpolate a montage of 19 spatially distributed 10-20 sites on a common head surface. The average of the 19 interpolated channels is calculated and then subtracted from all of the original scalp channels. The interpolated average reference procedure is the same as a typical average reference procedure (subtracting the average channel from each channel in the original montage) except it is not affected by different channels being excluded across data files (each data file uses the non-flagged channels to interpolate to the common reference montage) nor is it affected by varying recording montages used between recording sessions or studies (as each data file's recording montage is co-registered to the common head surface containing the reference locations).

Robust average reference

Prior to any post-staging signal quality assessment criteria the data are re-referenced to a selected channel average reference. Note that the initial selected channel average reference is different from the interpolated average reference that is used subsequently in the pipeline following bad channel classification procedures. At this initial stage it is assumed that large variance contaminated channels can be in the data array and if included in the interpolated average reference (or full average reference) they could contaminate all of the signals and negatively impact initial channel assessment criteria. Channels are excluded from the selected average channel reference based on their outlying voltage variance across one second time intervals over the recording period. Specifically the data are segmented into one second non-overlapping consecutive segments and the distributions of voltage standard deviation for each segment and each channel is used to identify outlying channels based on quantile distances from the median. It is important to note that when applying assessment criteria that considers distribution values across channels, it is necessary to use an average reference. Any reference that does not attempt to be equally distant from all channels introduces unnecessary voltage variance between channels which affect the distributions used to identify outliers.

Site variance time period criteria

Following the staging procedure and initial montage modifications, the criterion function is used with various input measures, in succession, to identify progressively more fine grained artifacts in the data. The lossless pipeline begins this criterion succession by examining the scalp channel signals for voltage variance as described in the criterion function examples above. The one second consecutive segmented data returned from the selected average channel operation are used to calculate the standard deviation of the voltage values for each channel-by-time window. These standard deviation values are stored and used as the input measure to the criterion function. It is important to note at this point that only channels and time points that have not already been identified for exclusion are included in the input measure array to the criterion function. The first criterion function identifies time periods in which too many channels have outlying voltage variances (as described above) using the 30% and 70%

quantiles, a 6 inter-quantile range outlier classification and a 20% channel criterion (as described above). Similar to the time classification implemented in the staging script, this classification measure is intended to identify very large voltage artifacts so a one-second padding is added to each side of identified time intervals to extend the duration of the annotation. The periods of time that are marked as having unusual scalp voltage variance are stored in the *time_info* annotation structure with the label “*ch_sd*”.

Site variance channel criteria

Following the criteria for classifying periods of time with artifacts, all identified periods for exclusion to this point are ignored and the channel voltage standard deviations are recalculated for each of the remaining channels and time windows. These standard deviation values are then passed as input to the criterion function in order to classify artifactual channels with voltage variance that is consistently identified as outliers compared to other remaining channels. The channel criterion function identifies channels that have outlying voltage variances in more than 20% of the time segments using the 30% and 70% quantiles and the 6 inter-quantile range outlier classification. The scalp channels that are marked as having unusual voltage variance are stored in the *chan_info* annotation structure with the label “*ch_sd*”. After each channel exclusion process, the remaining channels are used to create the interpolated average re-reference described previously.

Optional high pass and low pass filters

Following the scalp voltage variance criteria for gross artifacts and interpolated re-referencing, there is an opportunity to filter the data. Typically, if the data sets are filtered during the staging process, no filters are applied at this point. Common filters applied at this point are intended help establish a reliable ICA decomposition. Due to the tendency of nonstationary artifacts being made up of large, low frequency oscillations (e.g. movement artifact and sweat artifacts), ICA decompositions are more reliable when a high pass filter is applied to the data (e.g. 1Hz; Debener et al., 2010; Winkler et al., 2015). Similarly, a low pass filter may also provide a substantial improvement in ICA decompositions in the case where recordings are heavily contaminated with EMG signals. Many small muscle groups lie directly below the skin adjacent to EEG recording sites. These muscular sources of voltage do not have wide field projections like the cortical sources coming from within the skull. Because of this, it is common to have dozens of muscle sources uniquely contributing to the EEG recordings. Each of these muscle sources can take up an independent component when their voltage account for variance at the scalp. Because much of the muscle activity recorded at the scalp sites are at a frequency higher than many of the EEG frequencies of interest, it is sometimes desirable to apply a low pass filter (e.g., 30Hz) prior to running an ICA decomposition. It is important to note at this point that the full spectrum of the original recording is recoverable at any point during the pipeline, and is commonly reintroduced at the end of the pipeline for subsequent analysis. This recovery of the full recording spectrum is facilitated by the lossless property that the size of the data array does not change throughout the process. At the end of the pipeline the annotations could be applied back to the original signals and subsequent analysis could be carried out on the full-spectrum signals taking advantage of the annotations obtained from the reduced spectrum analysis.

In the case of the current test data set no spectral filters were applied to the data at this stage as the appropriate high pass and notch filters were applied during the staging process at the beginning of the pipeline.

Site correlation channel criteria

The standard deviation of voltage values criteria is a coarse measure to identify gross artifacts in the data and extremely noisy channels. The remaining input measures given to the criterion function are more fine grained and are intended to identify issues of spatial non-stationarity. If the voltages at the scalp are the result of sources in the brain, it is expected that neighboring channels on the scalp in a dense array EEG recording will have highly similar signal properties. If neighboring channels are consistently uncorrelated, or a period of time has an unusual number of uncorrelated channels, this is taken as an indication that there may be a spatial non-stationary artifact. Further, high and invariable correlations between neighboring channels indicates that the channels may be bridged due to an electrolyte link.

The remaining channels and time windows not marked for exclusion are then used to generate a correlation input measure to the criterion function. The correlation input measure is created by calculating, for each time window, the correlation coefficient between each channel with its three spatially nearest neighbors. The maximum correlation coefficient of the three neighbors is stored for each channel and each time window. This maximum correlation coefficient array is then passed to the criterion function in order to identify channels that are consistently outliers, over time, with respect to low maximum neighbor correlations relative to other channels. The channel criterion function identifies channels that, too often over time, have outlying neighbor correlation values using the 30% and 70% quantiles, the 6 inter-quantile range outlier classification and a 20% time criteria. The scalp channels that are marked as having unusually low maximum correlations to their nearest neighbors are stored in the *chan_info* annotation structure with the label “*low_r*”.

Bridged channel criteria

Following the identification of channels that are poorly correlated with neighboring channels, bridged channels are identified by taking the maximum neighbor correlation array and calculating the median and inter-quartile-range for each channel across time. A composite value is then created by dividing the median channel *r* values by their corresponding inter-quartile-ranges producing a value that accentuates high and invariable correlation distributions. Channels are then identified as bridged if their composite value falls 6 standard deviations (40% trimmed) from the mean (40% trimmed) across channels. The scalp channels that are marked as bridged to their nearest neighbors are stored in the *chan_info* annotation structure with the label “*bridge*”.

Rank channel selection

The marking of bridged channels is the final direct channel criterion in the lossless pipeline sequence. The channels that are not marked for exclusion at this stage will be included in the subsequent ICA decompositions. Because the pipeline uses a form of average reference for the data included in the ICA decomposition, it is necessary to account for the N-1 rank deficit (each channel in an average referenced montage is perfectly inversely correlated to the average of all other channels included in the reference). To account for the rank deficit, the lossless pipeline identifies one more channel to be excluded from the ICA model (but potentially included in other subsequent procedures). The “rank” channel (non-artifact channel that is not included in the ICA decomposition) is identified by the *r* array as the channel with the least unique information (or the remaining channel with the highest maximum *r* values across time points). The scalp channel that is marked as having the least amount of unique information based on correlations to nearest neighbors is stored in the *chan_info* annotation structure with the label “*rank*”.

Site correlation time period criteria

Following the neighboring channel correlation criteria for excluding high and low correlated channels, the data are re-referenced to the interpolated average and then a new correlation array is generated (as described above) with the remaining data. At this point the new correlation array is used as the input measure to the criterion function to classify periods of time when 20% of the channels are unlike themselves based on the 30% and 70% quantiles, and a 6 inter-quantile range outlier classification. Identifying periods of time when a substantial percentage of channels have unusual correlations to their neighbors is the first measurement for periods of spatial non-stationarity. Subsequent measures of spatial non-stationarity are performed on the IC activation signals. The time periods that are marked as having unusually low maximum correlations with nearest neighbor channels are stored in the *time_info* annotation structure with the label “*low_r*”.

At this point it should be noted that the parameters passed to the criterion function may be adjusted based on the properties of the recordings. Using the 30% and 70% quantiles and the 6 inter-quantile range outlier classification for scalp channel classification is a relatively strict cutoff that in some cases may remove time periods or channels that could be better accounted for by ICA decomposition. For example, this strict criteria may flag channels or time points due to eye blinks. In such cases we would prefer that the channel criteria is adjusted to allow eye blinks to remain in the data so that they can be isolated with ICA and the channels and time can remain in the recording. Typically this can be optimized by increasing the inter-quartile range outlier classification value. Although a value of 6 is used in the example above it is common to use values as high as 16 when the recordings are well suited for ICA decompositions.

2.4.2 Initial AMICA decomposition

After assessing the correlations between neighboring channel, the data are ready for the initial AMICA decomposition. The one-second windowed data are concatenated into continuous signals then periods of time and scalp channels identified for exclusion are not included in the input to the initial AMICA decomposition. The next step in the automated pre-processing pipeline is to perform an ICA decomposition using AMICA on the channels and time segments that were not flagged as problematic by the scalp measures described above. AMICA performs

ICA on the input data using an adaptive mixture model. Although AMICA is capable of generating multiple models during a decomposition, the lossless pipeline uses AMICA to learn a single model and store the log likelihood of the model at each time point. Another benefit of the AMICA implementation is that it can take advantage of multi-core systems and be performed quickly on HPC resources. At this stage a single AMICA decomposition is performed to establish a model of latent factors representing unique field projections in the EEG. Further, the log likelihood returned by the AMICA decomposition is added to the *time_info* structure in an annotation named "*logl_init*".

2.4.3 Initial ICA signal quality assessment

Following the initial AMICA decomposition, the scalp data and ICA information are used to perform further flagging of time windows based on the standard deviation of the ICA activations. This process of flagging time periods when too many ICA activations are unlike themselves is the same as the initial time flagging of scalp channel variance except it is performed on the ICA activation signals. Specifically, the continuous data are windowed into one-second non-overlapping epochs and the standard deviation of the voltage values for each IC activation signal by time window are stored and used as the input measure to the criterion function. The first IC criterion function identifies time periods in which too many ICs have outlying voltage variances using the 30% and 70% quantiles, the 6 inter-quantile range outlier classification and a 20% component count criteria. The periods of time that are marked in this initial run as having unusual IC activation variance are stored in the *time_info* annotation structure with the label "*ic_sd1*". This method for flagging periods of time where too many ICs have unusual activation variances is useful for identifying the time windows that did not fit the ICA model. With AMICA the log likelihood of the model is often low during these periods, indicating that it attenuates the influence of these non-stationary time intervals, but a subsequent run of the AMICA decomposition ignoring these time periods is a robust method for achieving a reliable ICA decomposition. This flagging of time periods based on the IC activation variances is the final property classification prior to performing the final AMICA decomposition. At this point, the AMICA model is only used to identify problematic periods of time, ICs are not classified until the final decomposition.

2.4.4 Final AMICA replication model

Once time periods have been flagged based on the variance of IC activations, 3 simultaneous single model AMICA decompositions are executed on the remaining data channels and time points that have not been flagged by any of the artifact classification procedures. Although the computational time of executing ICA decompositions has traditionally been a constraint for this type of ICA replication procedure, the multiple levels of parallelization on HPC clusters make this process feasible. Specifically, each of the three AMICA decompositions is performed as an 8-core MPI process, each of the 3 decompositions is performed simultaneously, and each subject recording is executed simultaneously. With this strategy, hundreds of files can be processed through the automated multi-AMICA pipeline within hours (depending on the size of the cluster, workload in the system's job queue and properties of the data file).

2.4.5 Final ICA signal quality assessment

IC reliability analysis

After the three simultaneous AMICA decompositions are performed on the identical remaining data points, and the new log likelihood arrays are added to the *time_info* annotation structure labeled “*logl_A*”, *logl_B*” and “*logl_C*”, the three models are tested for component replication across decompositions using *isctest* (Hyvärinen, 2011) with the default 0.05 false-positive and false-discovery parameters. AMICA will always return a model for a data file. If the data are not sufficient to generate a reliable decomposition, the models will vary significantly from run to run even though the models are performed on the exact same data array. Using *isctest*, the pipeline annotates the replicability of the AMICA model by identifying independent components that occur in each model. ICs from the first of the three decompositions that do not cluster with components from the subsequent decompositions are flagged in the *comp_info* annotation structure with the label “*ic_rt*”.

At this point it is worth clarifying the specific reason for doing each of the four AMICA decomposition during the lossless pipeline procedure. The Initial AMICA following the scalp signal quality assessment unit is not intended to be used for analyzing ICs themselves, but rather the IC time courses of activation are used to identify periods of time that should be flagged before performing a subsequent AMICA decomposition. Flagging ICs prior to performing subsequent AMICA decompositions does not have an effect on the quality of the decomposition (this simply reduces the rank of the data); however, flagging times in which numerous components have outlying activation variances can have a substantial effect on the reliability of subsequent AMICA decompositions. This is because having many ICs with outlying activation variances is an indication of spatial nonstationarity as the variance of scalp signals could not be modeled with a small set of components. The first of the three final AMICA decompositions “final AMICA A” is the decomposition that is intended for use in subsequent artifact correction, feature extraction and hypothesis testing. “Final AMICA A” is performed on the identical input data as “final AMICA B” and “final AMICA C”. Like the “initial AMICA” “final AMICA B” and “final AMICA C” are not intended to be used in subsequent analysis but rather are only intended to be used for data quality assessment. Specifically, they are used to determine if the input data array provided to AMICA is sufficient to produce a replicable ICA decomposition. If the input data array is not sufficient AMICA may still return a model; however, the models may vary substantially between A, B and C. By testing the replication of ICs across the final AMICA A, B and C, we are able to assess the sufficiency of the input data to the final AMICA modeling at the end of the lossless pipeline.

Final IC variance time period criteria

Once the final AMICA decompositions are established, the IC activation variance criteria previously performed on the initial AMICA decomposition is repeated on the first of the three final AMICA models and periods of time that are marked in this final AMICA run as having unusual IC activation variance are stored in the *time_info* annotation structure with the label "*ic_sd2*". There are no subsequent time classification procedures or ICA decompositions. This classification is intended to identify periods of time that should not be included in subsequent signal extraction or hypothesis testing.

Dipole fit and ICLabel component classification

The remainder of the pipeline prepares the data for the interactive QC process and subsequent post-processing. A single dipole is fit to each IC weights topography and the *ICLabel* (Pion-Tonachini, Kruetz-Delgado, & Makieg, 2019) extension for EEGLAB is used to classify ICs into seven categories each added to the *comp_info* annotation structure as "*brain*", "*eye*", "*muscle*", "*heart*", "*chan_noise*", "*line_noise*", and "*other*". Classification of independent components into the phenomena that they capture is typically achieved by expert review considering multiple properties of the ICs (e.g., topographies, spectral analysis, dipole fit). ICLabel is an EEGLAB extension that classifies components by examining the spatio-temporal measure in the ICLabel data set containing over 200,000 ICs from more than 6000 EEG recordings. The scale of this labeled classification set used in the ICLabel learning implementation is achieved via the ICLabel website that employs crowd sourcing strategies to collect expert IC labeling data.

2.5 Interactive Quality Control

At the end of the fully automated quality assessment and annotation stage of the lossless pipeline, the output files have all the information that is required to isolate the cortical signal from the various noise sources in the EEG traces. This information contains annotations indicating time periods and channels that should not be included in post-processing feature extraction and hypothesis testing procedures (based on various measures), and IC classification by the *ICLabel* toolbox. At this point, if the researcher is fully confident in the decisions made by the various criteria in the automated pipeline, this can be treated as a fully automated process that can purge channels, time points, and artifact ICs without inspection. A primary focus of the lossless pipeline, however, is rather to make the decisions made by the various criteria in the automated pipeline visible and easily adjusted by expert inspection.

The interaction with the data during the QC review stage of processing occurs inside a modified EEGLAB IC activation scroll plot. The IC activation scroll plot interactive figure is combined with other figures to create the QC dashboard illustrated in Figure 4. Axis background colors of the scroll plots (Figure 4, Panels A and B) indicate periods of time that have been flagged for various reasons by the pipeline, plus a "manual" time flag. The "manual" time flag indicates a period of time where the expert has decided to remove the data for subsequent analysis. At the onset of the QC process, the "manual" annotation includes all time points that have been marked as problematic in any of the automated measures. The expert reviewer is then free to add more time points to the "manual" flag, or remove any instances of "manual" flagged time points created at the initiation of the QC process. Using this "manual" flag method, the annotations from the automated calculations are never modified and the history of algorithm flags and manual interaction flags are permanently logged and isolated.

The modified EEGLAB scroll plot, provided with the *Vised-Marks* extension, allows users to associate keyboard presses with various actions on the data annotations. At the onset of the QC procedure the ICs are all flagged with an annotation corresponding to one of the seven categories classified by the *ICLabel* extension. The ICLabel probabilities for component classification at depicted in Figure 4 Panel C as a stacked histogram). If the highest probability classification of a component is “*other*”, the second highest probability classification is used in the annotation if its probability is measured to be greater than 30%. For example if a component is 60% “*other*”, 35% “*eye*” and 5% “*muscle*” the component is classified as “*eye*” in the comp_info annotation structure. However, if a component is 60% “*other*”, 25% “*eye*” and 15% “*muscle*” the component is classified as “*other”* in the *comp_info* annotation structure. This strategy of preferring specific classifications over “*other*” achieved closer agreement with expert review than simply selecting the highest probability classification. In preparation for the interactive QC review of each data file, “manual” flags are added to all components that have been labeled “*eye*”, “*muscle*”, “*heart*” and “*chan_noise*”. Following the QC review process, it is the intention that all channels, time points, and components that are flagged with the “manual” label are purged before post-processing.

The storage of the modified annotations following the QC review process is modular within the BIDS standard data storage strategy. The process of performing a QC review process of the lossless data state introduces no modifications to the data signals, but rather only modifies data property annotation. As such, the entire data file is not re-saved, instead, only the data annotation fields are stored to their respective BIDS compliant annotation files. Further, several versions of the annotations can be saved uniquely for the same data file each with their own identifier. With this flexibility in storing multiple versions of the annotations for each data file, research groups can easily implement strategies for inter-rater-reliablility measures.

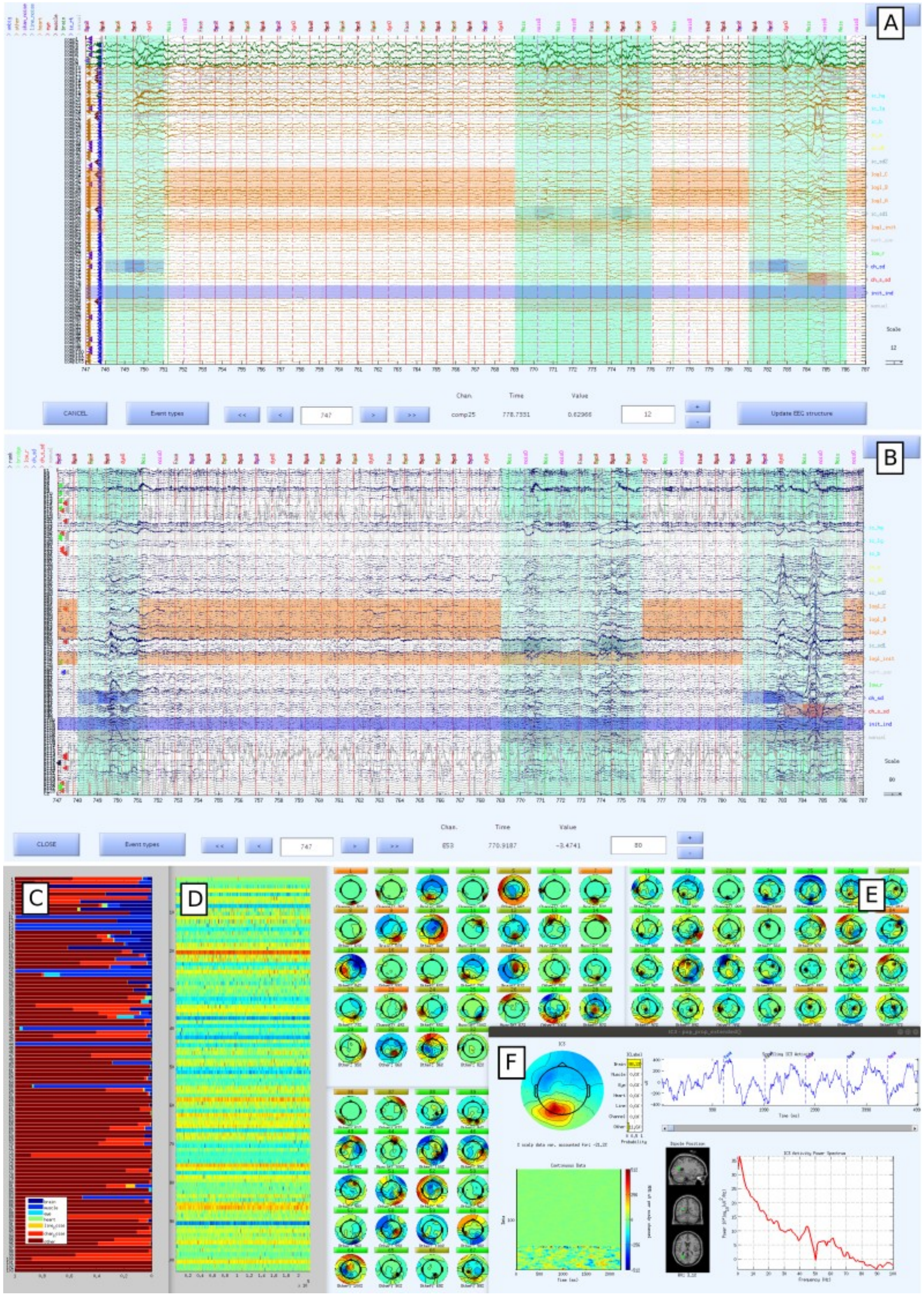

Figure 4.*Interactive quality control dashboard.*

A fundamental property of the lossless pipeline is that the entire duration of a channel can be rejected from post-processing (a portion of a channel cannot be rejected), all channels for a given time period can be rejected (a subset of channels cannot be rejected for a period of time), or an IC can be removed from the duration of the recording. Because of these properties, it is important that the automated pipeline leave in aspects of the data that can be removed by ICs. For example, the automated pipeline should not reject time periods because of eye movements or blinks if those artifacts can be isolated in ICs and removed to reconstitute the artifact free EEG. Further, if a scalp channel experiences noise for a brief period of time, neither the channel nor the time period of the artifact should be rejected if the artifact on that channel can be isolated by a single or multiple ICs. Given these priorities, the role of the QC reviewer is to scroll through the data and mark with a “manual” label (by right clicking on the component waveform) any component that accounts for identifiable artifacts in the scalp data.

Although the annotation interaction performed by the reviewer during the quality control procedure is handled in the IC scroll plot, the complete QC dashboard contains several figures designed to most efficiently provide them with the information required to make accurate and replicable decisions about annotation modifications. The primary panel in the QC dashboard is the IC activation scroll plot (Figure 4, Panel A) in which a right-click of the mouse will toggle the “manual” annotation of the selected IC. Left-clicking selects time periods and pressing “m” on a selected time period adds a “manual” mark to that period. Pressing “M” over a selected time-period removes the “manual” annotations in that time period. The scalp data scroll plot (Figure 4, Panel B) is linked to the IC scroll plot such that paging forward on the IC scroll plot updates the time displayed in the scalp channel scroll plot. Understanding the scalp data as it relates to the IC scroll plot is important for making decisions about IC inclusion or removal. To help with this interpretation, several interactions are enabled on the scalp data scroll plot. The first is a topographical plot that is generated for the time point under the mouse pointer when the “t” key is pressed. The “o” key when pressed while viewing the scalp scroll plot toggles the IC projection overlay. The IC projection overlay collects the remaining (not marked for rejection) ICs and projects the data back to scalp. This projection back to the remaining scalp channels is overlaid onto the scalp scroll plot as gray waveforms over the original (blue) scalp signals. If more ICs are flagged for removal in the IC scroll plot, the “u” key can be pressed (while the scalp scroll plot is active) to update the projection overlay to be re-generated from the currently selected set of components. Being able to visualize the effects of IC selection in this objective way is a key aspect of replicable IC selection QC.

Beyond the scroll plots, other IC properties are available for display via the *ViewProp*s plugin to EEGLAB (https://sccn.ucsd.edu/wiki/Viewprops). The *ViewProps* displays (Figure 4, Panel E and F) provide an interactive IC topography array in which topographical projections of each component are displayed along with their numeric labels as well as *ICLabel* classification tag and likelihood. By clicking on the IC index button further properties of the ICs are displayed in a new panel (Figure 4, Panel F). This subsequent panel includes time-course displayed of the component activation, the spectra of the activation, a topographical map, dipole location displays and *ICLabel* probabilities associated with each of the possible *ICLabel* classification categories. In order to summarize the *ICLabel* classification of the IC set, a stacked histogram is plotted vertically in line with the IC scroll plot (Figure 4, Panel C) displaying the relative likelihood that each IC belongs to each of the seven *ICLabel* classifications. Finally, a surface plot illustrating the normalized voltage for all non-manual mark time points for each IC is also displayed in line with the IC scroll plot (Figure 4, Panel D). This plot is used to see at a glance whether an IC has evenly distributed activation over the duration of the recording, or if the component has isolated bursts of activation at specific times in the recording.

2.6 Post-processing generalization

An overarching goal of the lossless pipeline is that the automated processes used to establish the best possible chance of a reliable ICA decomposition place no constraints on any potential post-processing analyses. Specifically, any analysis that could have been performed on the original raw data can still be performed on the output of the lossless pipeline, except now isolated cortical signal can be extracted (based on annotations) prior to performing the desired post-processing procedure. Although the post processing procedures are unlimited, the scripts for each post-processing project on lossless state data can begin with a standard procedure that extracts the signal from the data based on the annotations. The process is simply to purge time periods and channels that have the “manual” flag and to retain only the ICs of interest.

The *Vised-Marks* plugin for EEGLAB provides functions for purging data based on annotations in the *EEG.marks* structure. The following three lines of code purges the channels, time points, and then components that have been flagged with the “*manual*” label.

```
EEG = pop_marks_select_data(EEG,'channel marks',[],'labels',{'manual'
'rank'},'remove','on');
EEG = pop_marks_select_data(EEG,'time marks',[],'labels',{'manual'},'remove','on');
EEG = pop_marks_select_data(EEG,'component marks',[],'labels',
{'manual'},'remove','on');
```

Example post-processing scripts are available at the BIDS-lossless-EEG pipeline gitlab repository.

2.7 Post-processing for signal retention and ERP effect robustness

In order to assess the impact of the signal quality characterization obtained in the lossless pipeline on ERP measures, two sets of ERPs were created from the sample data. Each set contained the segmented data required to perform a robust ERP contrast between “face” and “noise” stimulus conditions. The first no-lossless (noLL) ERP set had a conventional artifact rejection procedure performed on the segmented data without removing the ICs flagged for rejection during the lossless pipeline. In the lossless (LL) ERP set, the data were segmented with the same artifact rejection procedure, but after removing the ICs flagged during the lossless pipeline.

Although there are several signal quality assessment annotations stored in the files following the lossless pipeline, this analysis is focused on assessing the impact of IC annotations on the subsequent ERP analyses. Specifically, the channels flagged in the pipeline were interpolated, a 30Hz low pass filter was applied, the data were re-referenced to the average site and then segmented around “face” and “noise” stimuli with an interval of -200 ms to 800 ms relative to onset. At this point one data set did not remove any ICs (noLL) and the second data set removed the flagged-as-artifact ICs (LL). Time periods marked for rejection during the pipeline were included during segmentation and the two data sets were then passed through a common rejection procedure in which channels were identified for rejection using EEGLAB’s channel spectra outlier criteria (3 standard deviations), as well as trial rejection using voltage cutoff of +/-100μV. The resulting ICs-retained (noLL) and ICs-removed (LL) segmented data sets were analyzed for data retention quantities as well ERP effect precision.

2.8 Statistical comparison of data retention and ERP effect precision

Channel count, trial count, and subject count following the rejection criteria of the two segmented data sets were used to assess the data retention effect of using the IC annotations generated by the lossless pipeline. To ensure a fully within subjects design, only files that contained all four segmented sets (ICs-retained vs. ICs-removed and “face” vs. “noise”) were included.

We also examined an ERP effect across the ICs-retained and ICs-removed data sets. The STATSLAB software package was used to implement a robust 2 (condition: face, noise) by 2 (artifact handling: IC-retained, IC-removed) ANOVA, using percentile bootstrap, to examine the impact of lossless IC classification handling on the ERP effect (Campopiano, van Noordt, & Segalowitz, 2018). In addition, we carried out the bootstrap tests in each individual subject to generate single subject-level distributions of the ERP condition difference magnitudes independently from the precision (confidence interval width) of the ERP differences. Specifically, in each condition the single-trial data were randomly re-sampled with replacement, averaged, and the difference wave was calculated. Repeating this process 1000 times generates a distribution of the average difference between conditions and allows for the examination of the magnitude of the ERP difference, as well as the precision of the difference as the width of the confidence interval around the bootstrapped difference wave within individuals.

3 Results

3.1 Signal retention

Table 1 contains the data retention results for subject count, trial count, and channel count. By removing the flagged artifact ICs a significant increase of data retention was measured. First, the total number of subjects retained from the origin 100 sessions that completed the pipeline increased when artifact ICs were removed (recording sessions were removed if they contained less than 4 trials after artitfact rejection). Specifically, the subject retention of the original 100 participants for IC-removed (LL) segmented face and noise sets were 96 and 91 respectively, while the ICs-retained (noLL) had a retention of 78 and 65, respectively. Further, within the recording sessions that were retained for all conditions there was a significant effect of the IC removal resulting in an increased data retention of both channels and trials (*p*<.001). Most notably the average trial retention percentage for IC-removed (LL) segmented face and noise sets were 61.86% and 60.86% respectively, while the ICs-retained (noLL) sets had a retention percentage of 28.0% and 27.95%, respectively.

Table 1. Summary of descriptive statistics resulting in lossless (LL) and no-lossless (noLL) procedures for both face and noise stimulus conditions.

Condition	N	Total trials	Remaining trials (sd)	Remaining channels (sd)
Face LL	96	62.69	38.78 (19.75)	126.57 (1.73)
Face noLL	78	65.95	18.47 (12.38)	124.74 (2.18)
Noise LL	91	44.81	27.27 (12.18)	126.53 (1.81)
Noise noLL	65	47.15	13.18 (7.27)	124.68 (2.19)

3.2 ERP effect precision

Data retention is a primary goal of EEG pre-processing procedures. The data retention results of the lossless pipeline indicate that this strategy provides significant improvements for data retention if the ICs marked for rejection are removed from the data before performing subsequent post-processing procedures. It is important, however, to also confirm that the data retention achieved by the method is not at the expense of deteriorating cortical signal effects. In order to test that the data retention improvements reported here were not at the expense of an ERP effect, we ran a set of robust bootstrap tests on the resulting ERP segmented files.

We first examined the interaction between condition (face vs. noise) and artifact handling procedure (ICs-retained vs. ICs-removed) in the robust bootstrapped grand average ERPs. Panel A (top left) in Figure 5 shows that both post-processing strategies show a larger positive deflection after 100ms to noise stimuli compared to face stimuli. The confidence intervals around the interaction in Panel B of Figure 5 shows that there is not a significant difference in ERP effects across choice of post-processing procedure. Specifically, the 95% confidence interval of the interaction contains zero over the duration of the ERP.

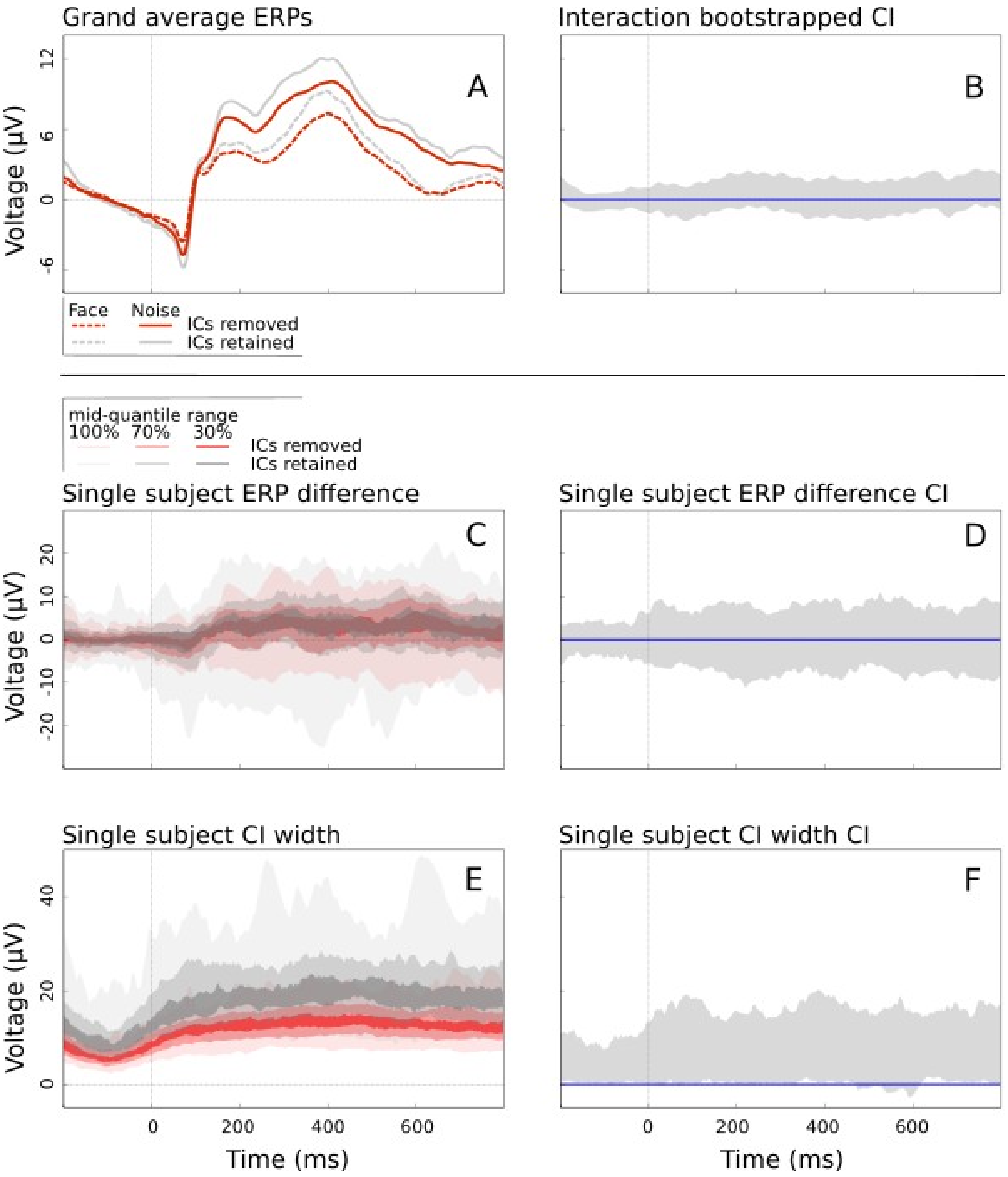

Figure 5. *ERP condition difference distribution overlay*

Potential differences in the ERP effect magnitude and precision were examined further by assessing the single subjects' data. Panel C (middle left) of Figure 5 illustrates that the distribution of single subjects ERP difference waves (i.e., condition difference waves) do not differ across post-processing strategies. The confidence intervals of the individual subject robust ERPs stimulus condition differences for ICs-retained (gray) and ICs-removed (red) are largely overlapping. Panel D of Figure 5 (middle right) shows that the 95% confidence interval around the single subject interaction contains zero over the duration of the ERP. Although post-processing strategies does not reliably impact measured ERP differences, we demonstrate in the Panel E (bottom left) of Figure 5 that there is a significant difference in the precision of the single subject differences. Specifically, by measuring the width of the confidence interval for single subject ERP differences, then plotting the distributions of these widths across subjects, we find that the IC-removed post-processing strategy contains less variance around the stimulus condition difference waves than ICs-retained. The result that the middle 70% of the ICs-removed single subject robust ERP confidence interval (red) does not overlap with the middle 30% of the ICs-retained confidence interval distribution (gray) is a clear demonstration of the benefit in IC-removal for single subject ERP measure precision. The systematic difference in confidence interval width of single subject ERP stimulus condition differences is illustrated in Panel F (bottom right) of Figure 5. The 95% confidence interval for difference in single subject ERP effect precision does not contain zero during a majority of the ERP duration.

Taken together, these results indicate that the significant increase in data retention achieved by removing the ICs flagged during the lossless pipeline is not at the expense of the ERP effect from the stimulus set. In fact the magnitude of the ERP differences do not change and the precision of the difference increases with IC-removal as a post-processing strategy.

4 Conclusions

The lossless pipeline produces and ideal state for long term, concatenated and integrated data sets by focusing on (1) maximally isolating cortical signal from the various forms of noise contained in EEG while also (2) maintaining maximal information from the raw recording throughout the process, (3) remaining generalizable to various EEG recording parameters and constraints, (4) being replicable across sites and projects, and (5) being scalable to large samples size. EEG data provides substantial potential for insights within large scale integrated neuroinformatics projects and the lossless state of EEG data provides the infrastructure for storing the data in a way which allows for maximal signal retention, flexibility in post processing strategies, and replication across projects.

By implementing tools for both batch dispatch automated procedures to HPC compute clusters and providing tools for interactive extendable data quality annotation, this pipeline is able to achieve an optimal balance between the automation of computationally demanding signal quality assessment and insightful interactive quality control review of the data. By collecting signal quality assessments from the computational pipeline into annotations that can be viewed and modified through visual inspection of the data, we are able to increase both the efficiency and integrity of the quality control procedure.

By minimizing the data loss and manipulation during the processing in this pipeline (hence the “lossless” name for the pipeline) the outcome of these procedures is ideal for long term and standardized integration across projects. By focusing on adding descriptive data to EEG session files (in the form of data quality annotations and ICA models, etc.) without removing portions of data or applying data manipulations (other than minimal filters and re-referencing), the outcome data state of the lossless pipeline minimizes constraints on subsequent post-processing decisions.

The outcome of the lossless pipeline is well suited as a standard preprocessed state of EEG data, due to the maximization of signal and noise isolation, the minimization of the data removal and/or manipulation, and the comprehensiveness of the interactive quality control review.

Finally, beyond the benefits in the standardization of the data state, this strategy results in significant increases in data retention following artifact rejection that are not attributable to ERP effect reduction. This increased data retention is a critical accomplishment for EEG preprocessing. This is particularly the case in the context of large scale integrated neuroinformatics platforms that have the potential for including brain data in multi-modal machine learning efforts to establish high impact biomarkers for neurological conditions. The impact of a standard approach to storing quality assessed EEG data goes beyond the results of single studies but rather changes the landscape of how EEG data can be integrated with open data platforms and included in multi-modal neuroinformatic outcomes.

Supplements

EEGLAB extensions

Vised-Marks

In order for a pre-processing pipeline to accumulate information about data properties and perform operations on specific portions of the data without rejecting data, a method for expandable and exhaustive data annotation is required. The EEGLAB extension *Vised-Marks* (https://github.com/BUCANL/Vised-Marks) allows users to visualize a variety of annotation information that is incorporated into a marks structure that is added to the EEGLAB EEG structure (i.e., *EEG.marks*). The annotation within the marks structure spans the time domain, channels, and independent components, and is fully extendable so that new measure annotations can easily be added. This plugin works directly with an adapted version of the EEGLAB scroll plot (*eegplot*) so that users can visualize and interact with the data along with the annotations. This interface not only supports the visualization and manipulation of key elements of the data file (e.g., *events*), but also plots any annotations that are generated during pre-processing. The Vised-Marks plugin communicates with a configuration file that allows users to set a range of parameters for how the scroll window will interact with the marks structure. This configuration file allows users to specify the type of annotation information that is plotted, as well as various key commands that can be executed while interacting with the data (e.g., plotting topographical maps, etc.). From the *Vised-Marks* menu, users also have the option to epoch and concatenate data, edit the mark structure information, and select data based on the information in the marks structure. Although the *Vised-Marks* is primarily designed to make it easier to interact with the data, it also makes it easier for testing new methods of flagging and visualizing artifacts.

Batch-Context

The EEGLAB extension Batch-Context (https://github.com/BUCANL/Batch-Context) offers users an efficient way to generate text files and edit *replace strings* so that scripts can be executed in a batch procedure across multiple files. The *replace strings* can be altered in configuration files across studies without the need to modify the sequence of commands or the actual scripts that are being executed. A *replace strings* is a *key* and *string* pair that allows the user to define *key* names in a configuration text file and associate a *string* with each of the *keys*. At run time each instance of the *keys* in the scripts are replaced with the associated strings before the procedure is executed. At the most basic level, the *replace strings* are used for fields holding information about the location of the data files, swapping out the file name for the loading and saving of each data set. Each script that carries a set of procedures is paired with a configuration file, which holds the *replace string* information. With this approach, users can add *replace strings* to a configuration file so that this information is incorporated into the batching script that will be executed across data sets. For example, in addition to loading and saving functions, users could set a *replace string* that specifies the high and low pass cutoffs for filtering. Any change to these parameters simply requires modifying the corresponding *replace string* comments in the configuration file, leaving the batch script untouched. This extension facilitates the scaling up of pipelines by dispatching batch procedures and running procedures in parallel on HPC compute clusters.

Figure Legends

Figure 1.Criterion function time outlier flagging based on voltage variance across time.

*Top left (**A**): Channel voltage variance across time. Top right (**B**): Voltage variance distribution of time windows for each channel, including the median (dark grey), 30% and 70% quantiles (grey), values that are six times the median-quantile distance (purple), and outliers (red). left middle (**C**): Flagging criteria array for outliers based on voltage variance. Left bottom (**D**): Summed flagging criteria to identify time windows that exceed the critical cut-off (time windows that are outliers in more than 20% of the channels).*

Figure 2.Criterion function channel outlier flagging based on voltage variance across channels.

*Left top (**A**): Channel voltage variance across time. Bottom left (**B**): Voltage variance distribution of channels for each time window, including the median (dark grey), 30% and 70% quantiles (grey), values that are six times the median-quantile distance (purple), and outliers (red). top middle (**C**): Flagging criteria array for outliers based on voltage variance. Top right (**D**): Summed flagging criteria to identify channels that exceed the critical cut-off (channels that are outliers in more than 20% of time windows).*

Figure 3.Lossless pipeline process sequence schematic.

The sequence of the procedures in the lossless pipeline are depicted as a set of nested colored rectangles in this schematic. Overall the pipeline consists of three automated signal quality assessment units (gray rectangles) which flank two stages of AMICA decompositions. The overall process ends with an "interactive quality control assessment" that isolates all of the data quality visualization into a single interactive inspection (purple box). The "scalp signal quality assessment" is the first of the signal quality assessment units which examines measures of channel signal quality to determine which channels and time points should be hidden from the AMICA decomposition. This assessment unit begins with an optional (white box) staging script for study specific annotation (red box) and signal manipulation (blue box). The remainder of this assessment unit progresses from measures of voltage magnitude assessment to more fine grained measures of neighboring channel correlation (in order to start identifying channels and periods of time which are spatially non-stationary). Following each of the subsequent AMICA decomposition stages the assessment units examine the ICA signal quality to identify periods of time in which too many components have unusual voltage activations.

Figure 4.Interactive quality control dashboard.

Panel A is the IC scroll plot showing the continuous time course of component activations and data annotations for component classification as well as time period flagging. Panel B is the scalp data scroll plot showing the continuous time course of scalp signal activations and annotation for flagging channels and time periods. This scalp data scroll can generate topographical plots for any time point and plot the IC projection overlay. IC projection overlay collects the remaining (not marked for rejection) ICs and projects them back to scalp. This projection back to the remaining scalp channels is overlaid onto the scalp scroll plot as gray waveforms over the original (blue) signals. Top: Panel C is a stacked histogram displaying the *relative likelihood that each IC belongs to each of the seven ICLabel classifications. Panel D is a surface plot illustrating the normalized voltage for all non-manual marked time points for each IC. Panel E is the ViewProps display showing an interactive IC topography array along with numeric identifiers, ICLabel classification and likelihood. Properties of the ICs can be displayed in a new figure illustrated in Panel F, which includes component time-course activation, frequency spectra curve, topographical map, dipole location, and ICLabel probabilities associated with each of the possible ICLabel classification categories.*

Figure 5. ERP condition difference distribution overlay

Panel A contains the grand average ERP overlays for the interaction between stimulus condition (Face = dashed line, Noise = solid line) and post-processing method (IC retained = gray and IC removed = red). The ERP overlay shows a clear morphology similarity and effect between the two post-processing methods. Panel B illustrates the confidence interval for the grand average bootstrap interaction test showing that the 95% confidence interval of the interaction encompasses zero during the period of the ERP. Panel C depicts the single subject ERP difference distributions overlaid across post-processing method IC retained (gray) and ICs removed (red). The areas of the distributions are represented by color intensity where the middle 30% of the distribution is darkest, the middle 70% is lighter, and the full distribution range is the lightest. Panel D is the single subject ERP difference effect confidence interval. The gray area represents the 95% confidence interval of the distribution of single subject interactions between stimulus condition ERP difference and post-processing method. Panel E illustrates the distributions around the width of the single subject bootstrapped confidence intervals of the ERP effect using the same color assignment as Panel C, Panel F is the single subject ERP difference effect width confidence interval for the interaction between stimulus condition and post-processing method. The gray area represents the 95% confidence interval of the distribution of single subject bootstrapped ERP difference variance between post-processing method showing that there is a significant difference in the precision of the ERP effect between the two post-processing methods.

References

Palmer, Jason & Kreutz-Delgado, Ken & Makeig, Scott. (2011). AMICA: An Adaptive Mixture of Independent Component Analyzers with Shared Components.

Acknowledgments

Substantial technical contributions were made by members of the Brock University Cognitive and Affective Neuroscience Lab: Allan Campopiano, Andrew Lofts, Mae Kennedy, Mike Cichonski, Tyler Kennedy Collins, and Sara Stephenson.

Computational resources provided by SHARCNET, Compute Ontario, Compute Canada.

This work was supported by NSERC, funding to Sid Segalowitz, funding to Mayada Elsabbagh, and Dedicated Programming support from SHARCNET.