Bootstrap-based distribution of robust distances for outlier detection in fMRI

fatma parlak

08/18/2021

Abstract

Since fMRI data inevitably artifactual contaminated due to both participant and harware related reasons, fMRI based studies are required to identify artifactual volumes. These volumes contain abnormal signal intensities, which is known as outliers in statitical terminology. There exist many outlier-detection methods for these types of datasets. However, these methods either are non-robust or do not yield principled threshold value. Although there exist robust approaches that are originated Mahalanobis distance known as Robust Distance (RD), we observe clear violations of their assumtions of gaussianity and independence in the fMRI context. In addition to that the distribution of these RD are unknown that preventing us to obtain quantile based threshold values. We develop a robust outlier detection method, which determines the distribution of RD. This procedure starts with utilizes univarite outlier imputation followed by a non-parametric bootstrap procedure to estimate an upper quantile of the distribution of RD, which serves as a threshold for outliers. The performance comparison of our RD-based approach is done with recent approaches employing 5 highly noisy rsfMRI sessions from HCP.

Keywords: outliers, fMRI, robust estimates

1 Introduction

Functional magnetic resonance imaging (fMRI) is a non-invasive technique that can be used to localize task-specific active brain regions and assist in predicting psychological or disease states (Lindquist 2008). During neuronal activity the brain hemodynamics alter due to an increase in the blood oxygenation level in the related brain cells. These changes in the brain are captured in a magnetic resonance imaging (MRI) scanner. These MR images are collected over the time (2~60 min.) of the experiment and consist of ~$100,000$ evenly sized cubes known as voxels, which collectively represent an individual's whole brain. Therefore fMRI data are a form of high dimensional (HD) data containing the received signals from each voxel at each time point. An acquired image from all voxels at one time point is called a *volume*.

To employ artifact-contaminated fMRI data reduces the quality of the results and influences the statistical result by reducing the signal-to-noise ratio (SNR) and violating common statistical assumptions. A low SNR is one of the shortcomings of fMRI that makes it more difficult to identify active brain regions associated with an activation task because these artifactual signals might mask the real brain signals. Artifacts may be either participant-related or equipment-related. Head movements, eye movements, breathing, and heartbeats are examples of participant-related artifacts (Friston et al. 1996; Beauchamp et al. 2003; Frank, Buxton, and Wong 2001; Krüger and Glover 2001). An example of equipment-related artifacts could be a scanner drift. Before data analysis, the identification of artifact-contaminated volumes is crucial.

In this work, we propose a robust outlier-detection approach for identifying artifact-contaminated fMRI volumes. Our approach considers the HD, auto-correlated, and non-Gaussian aspects of fMRI data.

Standard outlier detection methods fail in HD data. But, in many cases dimension reduction techniques can be used, after which existing outlier detection methods can be applied. The literature provides various distance-based outlier detection methods for low-dimensional data. One of the oldest is Mahalanobis Distance (MD) (Mahalanobis 1936) which is based on the sample mean and sample covariance. However, both the center and scaling factors are obtained by using all observations and may be influenced by outliers, which can cause "masking" and "swamping" effects (Rousseeuw and Van Zomeren 1990). The masking effect is failure to identify true outliers, while the swamping effect is to flag non-outliers as outliers. So MD is a non-robust measure.

Rousseeuw and Van Zomeren (1990) proposed the use of Minimum Covariance Determinant (MCD) estimates of mean and covariance rather than the conventional ones to produce a robust distance (RD) measure. MCD estimates are calculated by splitting the data into two subsets, one of which contains the observations located very close to the center of the data and therefore unlikely to represent outliers. This subset is used to estimate the location and scale parameters to avoid the influence of outliers.

Hardin and Rocke (2005) derived the theoretical distribution of MCD-based RD for Gaussian distributed data. An upper quantile of that distribution can be used to identify outliers. For example, using the 99 th quantile of the theoretical distribution, on average 1% of non-outlying observations will be flagged as outliers. Unfortunately, the empirical distribution of RD has often been observed to deviate from the theoretical distribution. Previous work employing MCD-based RDs attempted ad-hoc approaches to improve the distributional fit, for example median matching (Maronna and Zamar 2002; Filzmoser, Maronna, and Werner 2008; Mejia et al. 2017). Some reasons for this deviation might be due to violation of assumptions of independence, identical, and normally distributed.

fMRI data typically violate these key assumptions, in particular those of Gaussianity and independence. Here, we develop a nonparametric bootstrap-based approach to determine the distribution of RDs and estimate an upper quantile of that distribution used to threshold outliers. This method can be applied to any type of low-dimensional data, so we assume the fMRI data has undergone dimension reduction using previously proposed techniques (Pham et al. 2021). The main steps of our method are as follows. First, we start with applying a univariate outlier imputation to mitigate the influence of outliers on the procedure. Second, we define the two MCD subsets. Third, we employ a nonparametric bootstrap within each subset to estimate the distribution of RD and the quantile for thresholding of outliers. Finally, this threshold is applied to the RDs of the original data to identify outliers.

The remainder of this paper is organized as follows. In section 2, we introduce our proposed method and compare it with Hardin and Rocke's theoretical approach by employing simulated data and two "toy" fMRI datasets. In section 3, we apply our method to real-world fMRI data and compare it with an existing statistical method (Mejia et al. 2017) and a hardware-based approache (Power et al. 2012, 2014). In section 4, we briefly summarize the results and discuss possible future work.

2 Method

The main purpose of this study is to develop an outlier-detection method to identify artifact-contaminated fMRI volumes. Since neuronal activities account for a small amount of variance in fMRI data, temporal spikes or outliers in signals are assumed to be of artifactual origin. Our approach consists of four steps, described in the following subsections. The first step is to reduce the dimensions and to select high-kurtosis components, which are likely to represent artifactual volumes. The second step is to compute a robust distance measure. The third step, our novel contribution, is to estimate the null distribution of RD and estimate an upper quantile of this distribution to serve as an outlier detection threshold using a bootstrap procedure. The fourth step is to apply this threshold to identify artifactual volumes.

2.1 Dimension Reduction & Selection

Here, we adopt a dimension-reduction and selection approach proposed by Pham et al. (2021). Briefly, their technique reduces the dimensions of data by applying independent component analysis (ICA) and then selecting the high-kurtosis components. For fMRI, ICA is a popular tool for dimension reduction and for identifying spatially independent components, which can be of neuronal or artifactual origin. ICA for artifact detection and removal in fMRI datasets (Griffanti et al. 2014) and ICA-AROMA (Pruim et al. 2015). Consider an fMRI dataset, $\mathbf{Y}_{T\times V}$, where T is the duration of an fRMI experiment, V is the total number of voxels of the brain, where Q is the number of statistically independent components ($Q < T \ll V$). ICA decomposes $\mathbf{Y}$ into a spatial source signal matrix ($\mathbf{S}$) containing the independent components (ICs) and a temporal mixing matrix ($\mathbf{A}$) containing the temporal activation profile of each IC. That is, $\mathbf{Y} = \mathbf{A}\,\mathbf{S} + \mathbf{E}$. $\mathbf{A}$ can be considered a dimension-reduced version of $\mathbf{Y}$ along the space dimension. While the ICs of $\mathbf{S}$ may represent signal or noise, artifacts tend to appear in burst noise causing high spikes in the corresponding time courses. So the columns of $\mathbf{A}$ related to artifactual ICs will more likely contain extreme values. Kurtosis is an indicator of potential outliers. Hence, we select the K high-kurtosis components within the Q independent components. The resultant matrix, $\mathbf{X}_{T\times K}$, consists of the columns of $\mathbf{A}$ corresponding to the rows of $\mathbf{S}$ that are likely to represent artifacts. The goal of our method, described below, is to detect K-dimensional outliers among the T volumes via $\mathbf{A}$. These extreme observations can be excluded from analysis to avoid undue influence.

2.2 Robust Distance

Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)^T$ be data with n observations of p dimensions. The MD of the i^{th} observation $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ is defined as $MD(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \hat{\mu})^T\, \hat{\Sigma}^{-1}\, (\mathbf{x}_i - \hat{\mu})}$ where $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_p)^T$ and $\hat{\Sigma}_{p\times p}$ are the sample mean and sample covariance, respectively, across the n observations. Rousseeuw and Van Zomeren (1990) proposed minimum covariance determinant (MCD) estimators of the mean, $\hat{\mu}_{MCD}$, and covariance, $\hat{\Sigma}_{MCD}$, in order to prevent outliers from distorting the MD. These MCD estimates are obtained by using a subset of $h < n$ observations, which the MCD estimator achieves its maximum breakdown point by choosing $h = \lfloor (n + p + 1)/2 \rfloor$ (Lopuhaa and Rousseeuw 1991). For n observations $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ with p dimensions, the subset of h observations, $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, ..., \mathbf{x}_{i_h}$, is chosen to provide the minimum possible determinant of the covariance among any subset. That is, $det(\hat{\Sigma}_{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, ..., \mathbf{x}_{i_h}}) \leq det(\hat{\Sigma}_{\mathbf{x}_{k_1}, \mathbf{x}_{k_2}, ..., \mathbf{x}_{k_h}})$ for any $k_1, k_2, ..., k_h$. We call the observations $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, ..., \mathbf{x}_{i_h}$ *included* observations and the rest *excluded* observations. Let $S_1 = \{i_1, i_2, ..., i_h\}$ index the *included* observations, and let $S_2 = \{1, 2, ..., n\} \backslash S_1$ index the *excluded* observations. $\hat{\Sigma}_{MCD}$ and $\hat{\mu}_{MCD}$ are obtained by only using *included* observations. MCD estimators

have a breakdown point of $1 - \frac{h}{n}$ so that as long as the outlier fractions do not exceed $1 - \frac{h}{n}$ of the observations, the MCD estimator of the mean and covariance is robust. The use of MCD estimates of the mean and covariance in the MD formula yields a robust distance (RD) metric.

$$RD(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \hat{\mu}_{MCD})^T \ \hat{\Sigma}_{MCD}^{-1} \ (\mathbf{x}_i - \hat{\mu}_{MCD})}$$

To illustrate their approach, we employ publicly available data based on a single slice of one fMRI session from the Autism Brain Imaging Data Exchange (ABIDE) resource (Di Martino et al. 2014). This fMRI session was determined to be relatively free of artifacts based on visual inspection of the original fMRI data. For the illustration purpose, we choose 5 components from the dimension-reduced and high-kurtosis components selected fMRI data. Each component is a time series, shown in Figure 1. The horizontal orange lines indicate *included* time points, while the horizontal turquoise ones indicate *excluded* time points. We obtain RDs of each observation in the timeseries based on the mean and covariance calculated from *included* observations. As we can seen below, *excluded* time points tend to have higher signal occurrences and higher RD.

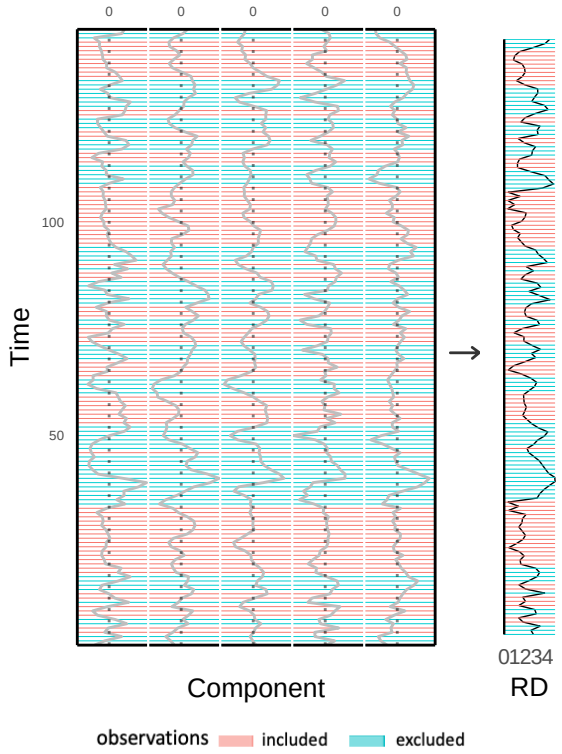


Figure 1: RD of 5 components of a ABIDE dataset

2.3 Theoretical Distribution of MCD-based RD

To identify outliers based on RDs, it is necessary to know the distribution of RD among non-outlying observations. An upper quantile of that distribution could be used as a threshold to identify outliers. Hardin and Rocke (2005) derived the distribution of MCD-based RDs of i.i.d. Gaussian data. They prove that RDs of *excluded* observations approximately follow a scaled F distribution. However, this theoretical result becomes invalid if any of the assumptions of Gaussianity, independence, or identical distribution are violated. As we show below, these assumptions may not hold for fMRI

data, making this theoretical result inapplicable to this type of data.

We visualize Hardin & Rocke's theoretical results on three outlier-free and Gaussian simulated data in Figure 2 showing empirical distribution of RD. Because the *excluded* observations are a subset of the dataset with a distribution truncated at the $(1-\frac{h}{n})$-th quantile, both *included* and *excluded* observations are displayed on the histograms for ease of visualizion. The scaled F distribution is displayed to show empirical versus theoretical distribution fit of *excluded* observations' RDs. Observations are identified as outliers when they lie beyond the turquoise vertical lines that are the 99^{th} quantile of the theoretical F distribution. We replicate these three datasets 100 times with exactly the same settings. We compute the average false positive rate (FPR), which indicates the rates of observations being wrongly labelled as outliers. Since the 99^{th} quantile is used, a FPR near 1% is expected.

Figure 2a shows the RDs of the first dataset, which is generated from independent and identical Normal distribution. Recall that under these assumptions, the *excluded* observations RD follow a scaled F-distribution. The second and the third datasets, shown in Figure 2b and Figure 2c, are generated from a Gaussian first-order auto-regressive ($AR(1)$) model to reflect one feature of fMRI data, autocorrelation. To mimic typical fMRI data we choose $\phi = 0.3$, a value is commonly assumed for the temporal correlation of fMRI data. For the third dataset, we choose $\phi = 0.6$, which imitates a more modern fMRI acquisition with fast temporal resolution. The theoretical F distribution performs reasonably well on the three histograms displayed in Figure 1. As expected the 99^{th} quantile lies at the tail of the distribution of *excluded observations'* RDs for each dataset, near the nominal level of 1%. The average FPR in the three scenarios are 1.3%, 1.4%, and 1.5% respectively. Even though the assumption of independence is violated in the second and the third datasets, FPRs suggest that the violation of the independence assumption has a relatively minor effect on the validity of the theoretical F distribution.

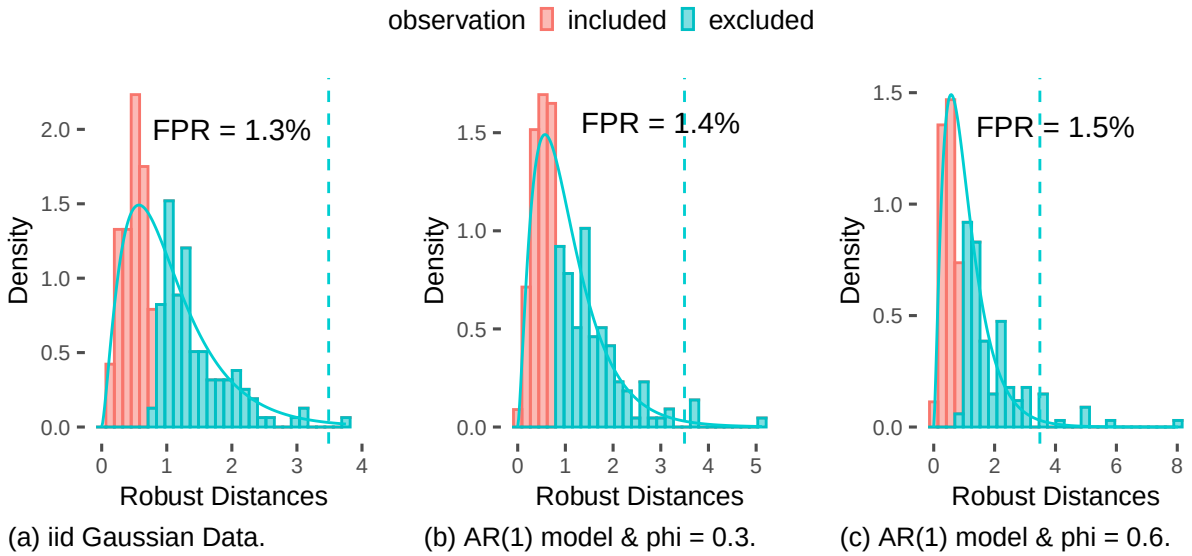
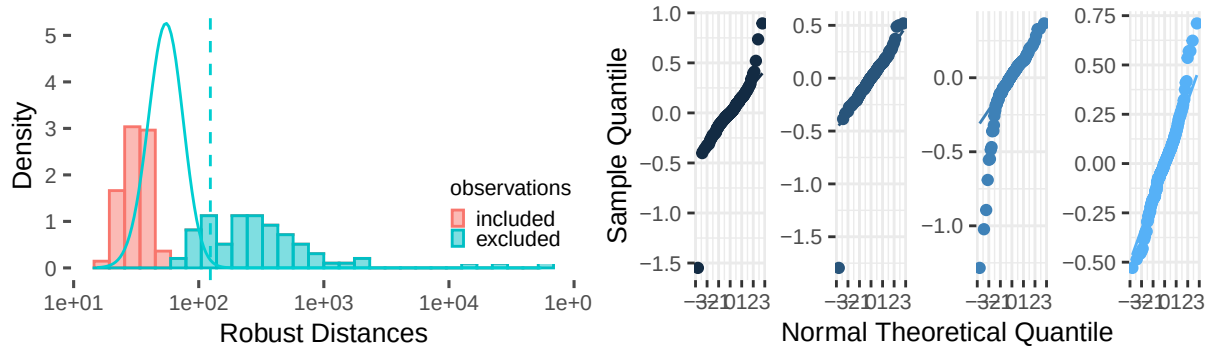


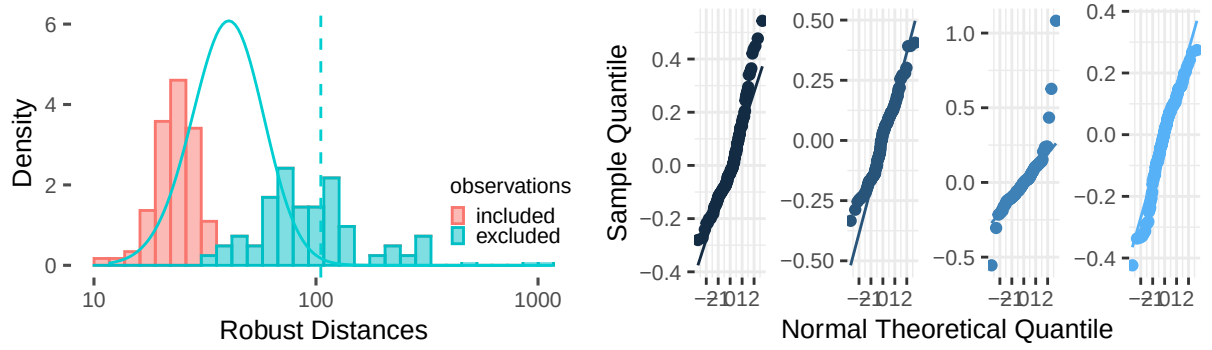
Figure 2: Distributions of RDs from simulated datasets (size = 250) along with a scaled F distribution. The vertical-dashed line indicates the 99th quantile of that F distribution. Reported FPR are obtained by replicating data with the same setting 100 times.

Apart from autocorrelation, fMRI data could violate the Normality assumption as well. To demon-

strate that violation and the subsequent failure of their approach, we employ two "toy" fMRI sessions based on a single slice of the brain from an fMRI session from the publicly available ABIDE database (Di Martino et al. 2014). Based on visual inspection of the original fMRI data, the first fMRI session, *ABIDE 1*, is known to be artifact contaminated, while the second fMRI session, *ABIDE 2*, is relatively free of artifacts. Figure 3 displays the empirical distribution of RDs of each ABIDE dataset, which are obtained after applying a dimension reduction by using ICA as described in section 2.1. Quantile-Quantile (Q-Q) plots are shown to check the Normality assumption for 4 of the components. Both datasets show violations of the assumptions in Hardin & Rocke's approach. First, the scaled F distributions clearly fail to fit the distribution of the RDs' *excluded* observations from both datasets. More specifically, the empirical distribution of RDs are much larger than what would be expected based on the theoretical distribution. The 99^{th} quantile of the theoretical F distribution falls within the middle of the distribution of *excluded* observations for both dataset. This would lead to many likely non-outlying observations being classified as outliers. The Q-Q plots of the selected components highlight that the selected components deviate from Normality which cannot be explained solely by the presence of outliers. Their distributions are non-Gaussian in different ways, which prevents application of a common transformation to achieve Normality. The conclusion is that, while Hardin & Rocke's theoretical results succeed when applied to i.i.d. Gaussian data and even correlated Gaussian data, their approach fails when assumptions of Normality are violated- a common occurrence in fMRI data.



(a) ABIDE 1



(b) ABIDE 2

Figure 3: Histograms of the Robust Distances of ABIDE datasets and along with a scaled F-distribution. The vertical lines indicate the 99th quantile of that F-distribution. QQ-plots of four components obtained from the dimension and selection approach described in Section 2.1 are displayed to check the Normality assumption.

2.4 Bootstrap-based Distribution of Robust Distance

Since fMRI data is autocorrelated and non-Normal, two key assumptions of Hardin & Rocke's approach to determining a distribution for RDs of *excluded* observations are violated. Although typical fMRI data could also violate their third assumption, which is to be from identically distributed data, we overcome this violation by applying mean and variance detrending to each independent component. Therefore, a method of determining the distribution of RD and identifying outliers is required to overcome independence and Gaussianity violation. Here, we introduce a nonparametric bootstrap procedure to estimate the $(1-\alpha)$ quantile of the distribution of RD, the desired threshold for outliers.

The nonparametric bootstrap is a statistical technique for obtaining an empirical distribution of a desired statistic by randomly resampling with replacement from the main sample with equal probability. Each bootstrap sample is used to calculate the desired statistics yielding a distribution for statistics of interest which could be used in inference. Without making assumptions about the population distribution. The following three subsections are to describe the procedure of our method. Section 2.4.1 describes a univariate outlier imputation step that we apply before bootstrapping the data. Section 2.4.2 describes the details of estimating the $(1-\alpha)$ quantile of the distribution of RD by using nonparametric bootstrap. Section 2.4.3 describes the choice of a threshold measure summarizing over bootstrap samples to identify outliers.

2.4.1 Univariate Outlier Imputation

The existence of outliers in the data alters the distribution of RD, particularly the upper quantiles. For instance, for a dataset of size n, assume that there is only one outlier. Bootstrap samples from that dataset would contain that outlier with probability $\left(1-\frac{1}{n}\right)^n \to \frac{1}{e} \approx 0.37$ as $n \to \infty$. In fMRI data, there are typically numerous outlying volumes. Prior to application of bootstrapping, we therefore apply univariate outlier imputation. To each component of the dimension-reduced and dimension-selected dataset. `Algorithm 1` describes the steps to obtain imputed data. First, we identify outliers by using the median absolute deviation (MAD), a robust measure of variability for univariate data. For a given observation $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, the MAD is defined as the median of the absolute deviations from the median of data $MAD = med(|\mathbf{x} - M|)$ where $M = med(\mathbf{x})$. We use a scaling factor of 1.4826 (Rousseeuw and Croux 1993), which makes the MAD a consistent estimator of the standard deviation for Gauissian data. Observations lying beyond 4 MADs of the median are considered outliers. To obtain an imputed dataset for each outlying observation, we use the mean of the nearest preceding and subsequent non-outlier observations. We will use this imputed data set, $\mathbf{X}^0_{T\times K}$, instead of the raw data in the bootstrap procedure to determine the distribution of RDs in the absence of outliers, as described next.

Algorithm 1: Univariate Outlier Imputation

Input: $\mathbf{X}_{T\times K}$ (dimension reduced, detrended & selected fMRI data)
Initialize $\mathbf{X}^0_{T\times K} \leftarrow \mathbf{X}_{T\times K}$
for *k = 1,2,.., K* **do**
  $\mathbf{x}_k$ k-th column vector and $M_k = med(\mathbf{x}_k)$
  Calculate $\text{MAD}_k = med(|\mathbf{x}_k - M_k)|)$
  Define $\mathcal{T}_k = \left\{t : x_{t,k} \notin M_k \pm [4 \cdot (1.4826 \cdot \text{MAD}_k)]\right\}$ the set of univariate outlier indices
  for $t \in \mathcal{T}_k$ **do**
    $x^0_{t,k} \leftarrow \text{mean}\left\{x_{t-a,k}, x_{t+b,k}\right\}$ where $a = \min\{1, 2, ..., t-1 : x_{t-a,k} \notin \mathcal{T}_k\}$ and $b = \max\{1, 2, ..., T-t : x_{t+b,k} \notin \mathcal{T}_k\}$. The corresponding $x_{t-a,k}$ and $x_{t+b,k}$ will be dropped if either a or b is null.
  end
end
return $\mathbf{X}^0_{T\times K}$ the imputed data matrix

2.4.2 Bootstrap Procedure

The goal of our bootstrap procedure is to estimate the $(1-\alpha)$ quantile of the distribution of RDs. This quantile will be used to threshold the observed RDs to identify outliers, where α represents the proportion of non-outliers that will be expected to be labeled as outliers, i.e. false positives. `Algorithm 2` describes the bootstrap procedure. First, we divide observations into *included* and *excluded* MCD subsets as described in Section 2.2. Second, we estimate the MCD covariance using the original sample (the "main" MCD covariance) because we discovered that the determinant of the covariance obtained from bootstrapped samples is smaller than that of the original dataset. This would reduce the RD to be overestimated due to the inverse factor of the covariance matrix. Consequently, the bootstrap-based $(1-\alpha)$ quantile will be overestimated, resulting in fewer outliers identified. To avoid this source of bias, we calculate estimates of the MCD-based covariance matrix using the original sample.

Third, we bootstrap *included* and *excluded* observations separately to preserve the MCD structure. For each bootstrap sample, we obtain the MCD estimate of the mean. We then compute the RD of each observation in the bootstrap sample using the main MCD covariance and the bootstrap MCD mean. We record the $(1-\alpha)$ quantile of the RD in each bootstrap sample. This process is repeated across B bootstrap samples to obtain a bootstrap distribution of the $(1-\alpha)$ quantile of the distribution of RD, which will be used to determine a cutoff point for outlier detection. Details of how we choose a threshold measure among several possible summary statistics from this bootstrap distribution are given next.

Algorithm 2: Bootstrap-based Estimation of the $(1-\alpha)$ Quantile of the distribution of RD

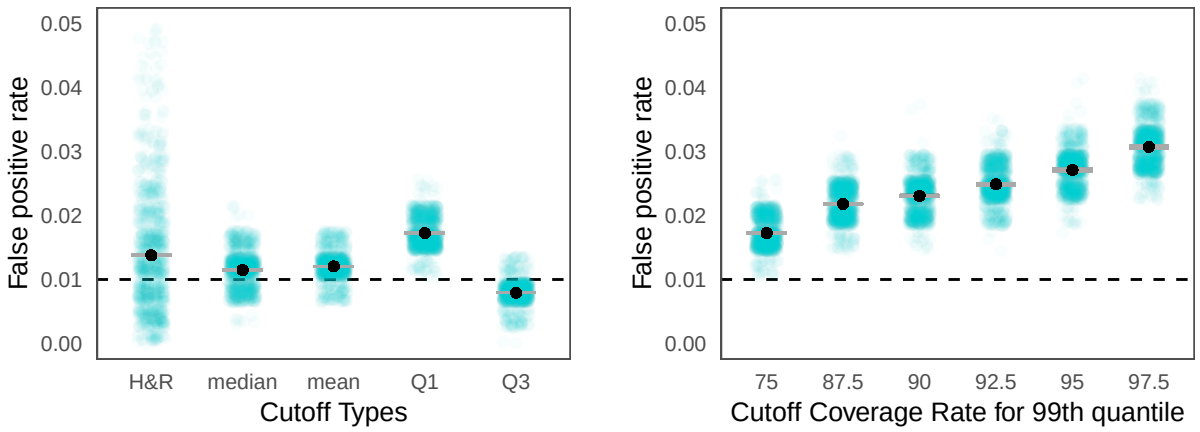
Input: univariate outlier imputed data $\mathbf{X}^0_{T\times K}$ and $\alpha \in (0,1)$
Define the index sets S_1 and S_2 for *included* and *excluded* observations of $\mathbf{X}^0$, respectively
Compute main MCD covariance $\hat{\Sigma}_{MCD}$ from $\{\mathbf{x}^0_t : t \in S_1\}$ where $\mathbf{X}^0_t$ is the t-th row of $\mathbf{X}^0$
for *b = 1,2,..,* B **do**
    use $F_1 = \frac{1}{n}$ probability mass to sample each of n included observations to create $S_1^{(b)}$
    use $F_2 = \frac{1}{n-h}$ probability mass to sample each of $n-h$ excluded observations to create $S_2^{(b)}$
    compute MCD estimate of sample mean $\hat{\mu}^{(b)}_{MCD}$ from $\{\mathbf{x}^0_t : t \in S_1^{(b)}\}$
    compute bootstrap-based RDs of $\{\mathbf{x}^0_t : t \in S_1^{(b)} \bigcup S_2^{(b)}\}$ by using $\hat{\Sigma}_{MCD}$ and $\hat{\mu}^{(b)}_{MCD}$
    calculate $(1-\alpha)$ quantile estimates, $\hat{Q}^{(b)}_{(1-\alpha)}$, from bootstrap-based RDs
end
return the estimated $\hat{Q}^{(b)}_{(1-\alpha)}$ for $b = 1, 2, ..., B$

2.4.3 Determining a Threshold Measure

To determine a threshold value for outlier detection, we consider several possible summary statistics across the bootstrap samples. Depending on the aim of a study, there are different objectives to control the FPR across samples: on average across sample FPR of α, FPR of *at least* α , and FPR of *at most* α with some statistical certainty (e.g. in 95% of samples). For example, a study may want to achieve *at least* α FPR in consideration of not sacrificing the power. We consider the first two possible objectives.

In order to examine the average FPR and its variability, we generate 1000 outlier-free replicates of i.i.d and Gaussian data of size $n = 250$. Consider a value for α of 0.01. We start by applying *Algorithm 2*, which results in a distribution of the bootstrap-based 99^{th} quantile. Then, cutoff candidates are obtained by calculating the median, mean, first quartile, and third quartile of these bootstrap-based quantiles. For comparison, we also calculate the 99^{th} quantile of the theoretical F distribution proposed by Hardin and Rocke (2005). The FPR for each replicate is illustrated in Figure 4.a along with the means across replicates. According to this plot, the theoretical F distribution-FPR is more variable than other cutoff types and is on average above the nominal rate of 0.01. The median and mean have FPR closer to the nominal rate and are less variable than the theoretical distribution.

For the objective of achieving at least FPR of α to avoid a drop in power, we consider one-sided bootstrap confidence intervals (CIs) of the $(1-\alpha)$ quantile as potential cutoffs. Specifically, we consider 75%, 87.5%, 90%, 92.5%, 95%, and 97.5% lower one-sided CIs. Note that the 75% coverage rate cutoff is the same as $Q1$ considered previously. As seen in Figure 4.b., all cutoffs achieve FPR above 0.01 in nearly all replicates. We prefer the 75% coverage cutoff, since it achieves the strictest type-1 error control, maintaining FPR below 0.02 in most replicates in this simulated example.



(a) FPR distribution by using different summary statistics and Hardin & Rocke's approach.

(b) FPR distribution by using the 75%, 87.5%, 90%, 92.5%, 95%, and 97.5% coverages.

Figure 4: FPR obtained by using different bootstrap-based cutoff types. The black dots represent the mean over replicates, and the gray bars are the standard error of the mean across 1000 replicates.

Based in this simulated example, we adopt two cutoff types on for further exploration both ABIDE datasets: the median and the 75% coverage bound (Q1). We apply the same approach to both toy ABIDE datasets. As illustrated in Figure 5, it is clear that H&R's quantile performs poorly with more than half of the *excluded* observations labelled as outliers particularly bad for ABIDE1, the more with artifacts. By contrast, the median and 75% coverage cutoffs perform well, successfully identifying the long upper tail likely to consists of outliers in both datasets.

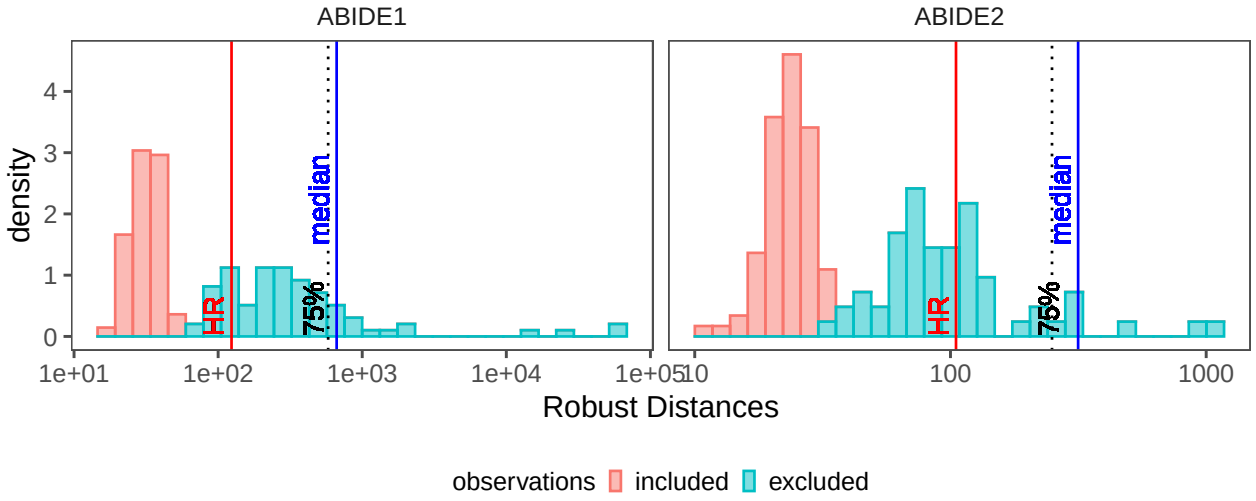


Figure 5: Distribution of RD in ABIDE datasets with different cutoffs.

3 Experimental Data Results

We apply our bootstrap-based outlier identification method to five highly noisy fMRI datasets and compare its results with existing methods. This section is organized as follows. In Section 3.1, we introduce the fMRI data we employ. In Section 3.2, we apply our bootstrap-based outlier identification procedure and present the results. In Section 3.3.1, we compare the bootstrap-based approach

with an existing non-robust distance approach In Section 3.3.2, we make another comparison between our bootstrap-based method and a popular hardware approach.

3.1 fMRI dataset

To evaluate performance of the bootstrap-based artifact-identification method in fMRI data, we employ freely available data from the Human Connectome Project (HCP) (Van Essen et al. 2013). This data resource includes resting-state fMRI (rs-fMRI) from over 1000 healthy young adults. Details of acquisition and processing can be found in Van Essen et al. (2013) and Glasser et al. (2013). Each subject was scanned over two sessions (REST1 and REST2), and at each session there were two runs acquired with different methods (LR and RL). We choose two rs-fMRI datasets from the same subject "A" and three rs-fMRI datasets from different subjects ("B," "C," and "D") exhibiting high levels of noise and artifacts. Each dataset includes 1200 volumes acquired every 0.72 seconds over approximately 15 minutes. The total number of voxels inside the brain is approximately $200,000$ but varies across subjects. We refer to these five sessions as *A1*, *A2*, *B*, *C*, and *D*. Each fMRI session is 4 dimensional ($4D$) array of size $91 \times 109 \times 91 \times 1200$. The first 3 dimensions represent $[x, y, x]$ information, while the last dimension represents time. In order to convert this data into a matrix format, we start with creating a subject-specific brain masks indicating which voxels are located inside the brain. That masks enable us to convert this $4D$ array into a matrix $\mathbf{Y}_{T\times V}$ where T is the total acquired volumes for an experiment, while V is the total number of voxels in the brain.

3.2 Bootstrap-based artifact detection in rs-fMRI datasets

For each session, we first apply dimension reduction and selection method as described in Section 2.1. Second, we compute the RDs of each volume and apply the bootstrap-based outlier-detection method described in Section 2.4.2. We obtain outlier thereshold values based on the median and 75% coverage rates as described in Section 2.4.3. Figure 6 shows the empirical distribution of RDs of each fMRI session. Since both cutoff values look very close to each other, we adopt the median as a threshold for further analysis.

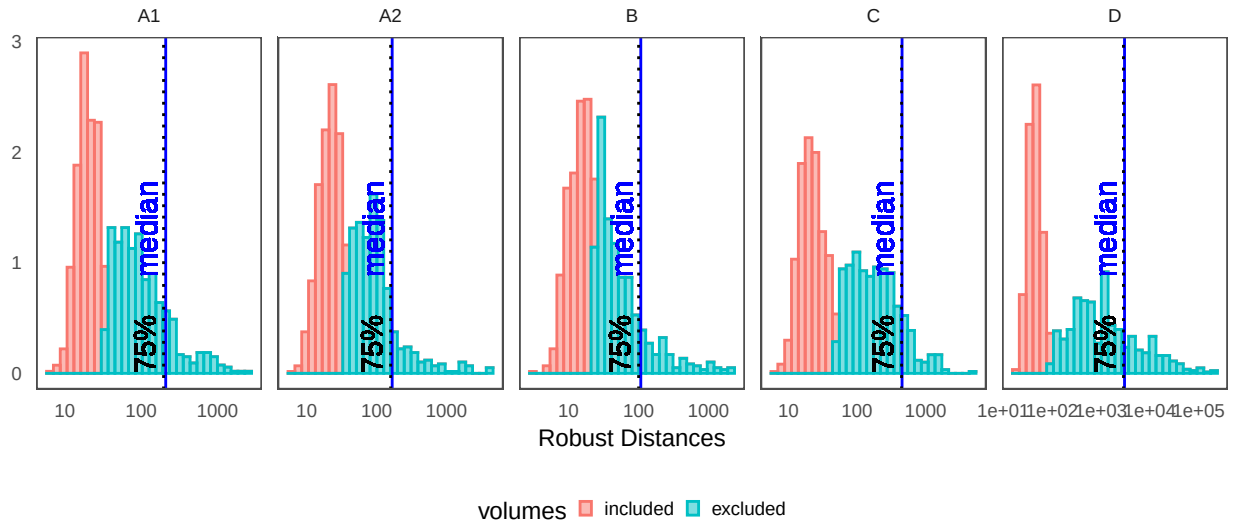


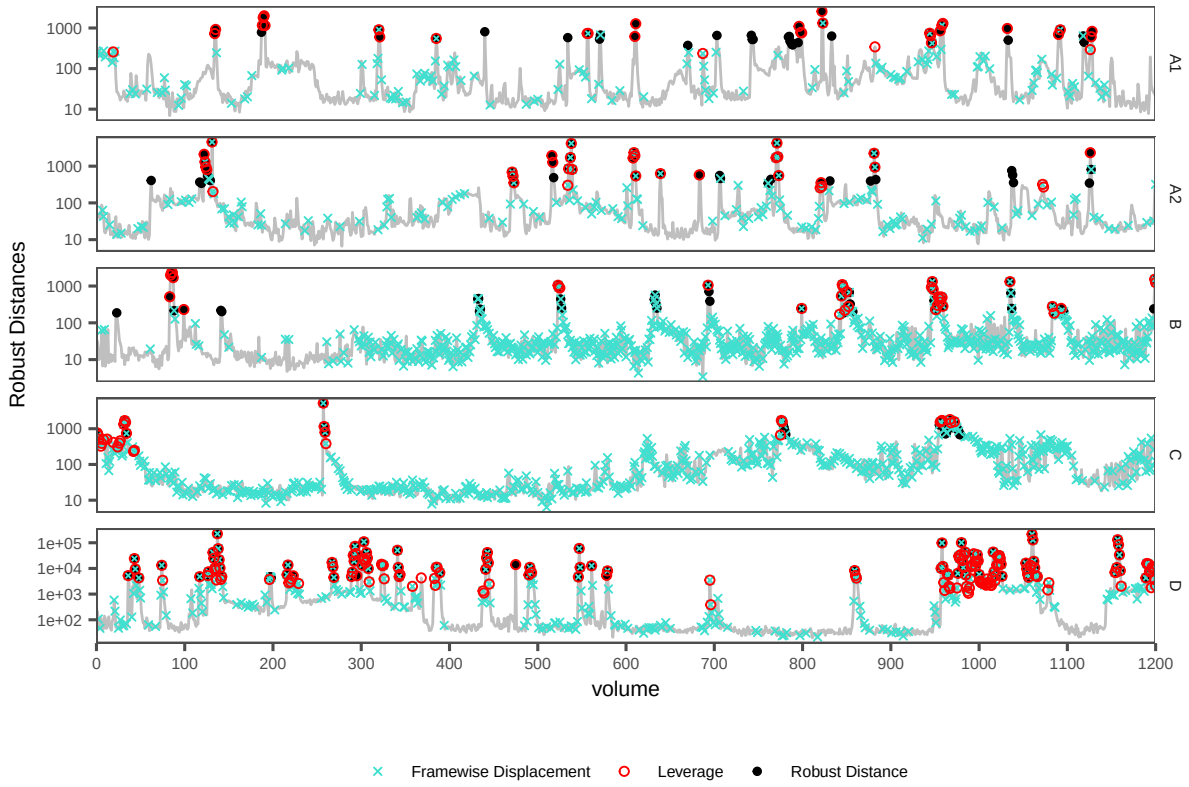
Figure 6: For each session, distributions of the Robust Distances for each volume. The vertical black lines are two cutoff types: median(solid) and 75 percent coverage (dotted)

3.3 Evaluation of the method

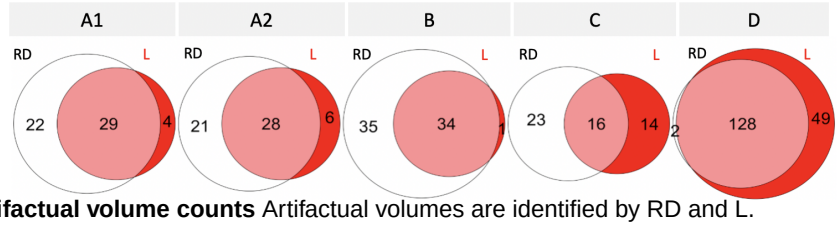
We evaluate our bootstrap-based artifact-identification by comparing with "ICA Leverage," which is based on the same dimension reduction and selection procedure but employs a non-robust distance measure, leverage, and uses an ad-hoc threshold for outliers (Pham et al. 2021). Specifically, artifactual volumes are identified based on having a leverage value greater than 4 times of the median across all volumes. While this approach has been shown to outperform existing data-driven and hardware-based approaches (Pham et al. 2021), it may suffer from masking and/or swamping effect due to the relationship of leverage with Mahalanobis Distance. Another common practice to detect artifactual volumes in fMRI data is known as motion scrubbing using frame-wise displacement (FD), a measure of subject head motion. This approach aims to detect artifacts caused by head movement (Power et al. 2012, 2014). FD is a summary measure based on six rigid body realignment parameters used to align a subject's brain across volumes within an fMRI session. A threshold is applied to these time-specific FD measures to flag high motion volumes. Although this approach is commonly used in fMRI analysis, there is no principled or universally accepted threshold value. By following the most common practice across studies, we use 0.3 as a FD threshold.

We compare the occurrence of temporal artifacts detected by the RD, leverage, and FD approaches in Figure 7.a. Although the RD and leverage approaches mostly identify the same volumes, it can be seen that leverage fails to detect some volumes with high RD in *A1*, *A2*, *B*, and *C*. However, the RD approach seems to be inefficient for detecting artifactual volumes with the high RD in *D*. To clarify this comparison, Venn diagrams of both methods for each subjects are also displayed in Figure 7.b, which shows the total counts of identified outliers. It concludes that the RD approach almost doubled the number of identified artifactual volumes that are identified by leverage in *A1*, *A2*, *B*, and *C* datasets. Unlike the performance on these datasets, the RD approach identifies a subset of the volumes flagged with leverage in *D*. This may indicate that RD is missing many outliers. The comparison of RD's magnitude across datasets indicate that unlike the dataset *D* has more extreme RD magnitude than the others. This could indicate that the underlying distribution of the components are more right-skewed. This could lead to failure of our univariate outlier imputation approach, which is based on a measure appropriate for normally distributed data.

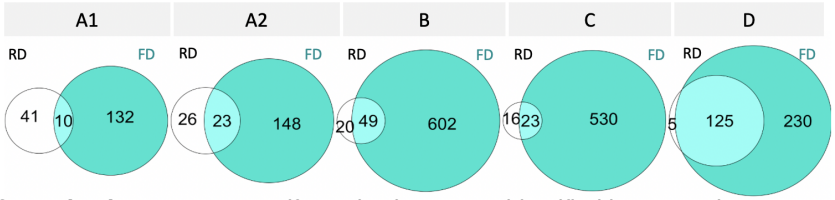
To compare the RD approach with FD, we first visualize the temporal occurrence of artifacts as shown in Figure 7.a. In Figure 7.c., it can be seen that by using the FD artifactual volumes substantially outnumbered RD-based identified volumes in terms of temporal occurrence. That is, motion scrubbing removes a lot of volumes, and many of those do not have high RD. Therefore, motion scrubbing may be identifying volumes that are not significantly abnormal. FD may also fail to identify volumes affected by hardware-induced artifacts. In Figure 7.c. , we also present Venn diagrams, which show the total number of artifactual volumes that are detected by both approaches. This figure illustrates the RD based approach tends to detect fewer artifactual volumes than FD. Since RD based approach is not designed to identify only motion related artifacts it also identifies additional volumes in each session. However, it is interesting to see that, for subject *D*, this approach only identify 5 artifactual volumes. This is further evidence that RD based approach may fail to identify some outliers in this session.



(a) Temporal occurence of identified artifacts Artifactual volumes are identified by RD and L appraoches displayed in each subject's rs−fMRI data. The gray lines are robust distance of each volume for each session.



(b) Temporal artifactual volume counts Artifactual volumes are identified by RD and L.



(c) Temporal artifactual volume counts Artifactual volumes are identified by RD and FD.

Figure 7: Comparison between the RD-based (RD), Leverage (L), and Framewise Displacement (FD) approaches in terms of artifactual-volumes' temporal occurrence locations and counts

Finally, we visualize the artifacts associated with spatial patterns of RD-based artifactual volumes for each high RD volume. Recall that $\mathbf{X}$ is obtained by selecting a subset of the detrended columns of $\mathbf{A}$. Let $\mathbf{A}^*$ and $\mathbf{S}^*$ contain only the selected rows and columns of $\mathbf{A}$ and $\mathbf{S}$. For each outlying volume t, an image of artifact intensity can be obtained by multiplying the t-th row of $\mathbf{A}^*$ by $\mathbf{S}^*$. Since each of these volumes cannot be displayed due to space constraints, we visualize the average

across all outlying volumes, giving an overall measure of artifact intensity at every voxel of the brain.

We display 2 orthogonal views of brain images (slices) acquired from each subject, which contain artifact intensities, in Figure 8. The images show the artifact intensities of artifactual images. The intensities are lower when the color is more red and are higher when the color is more yellow. The axial views of each subject show a ring of intensity on the outer edge of the brain, which is often caused by head/ brain movements during the data acquisition. It is possible that these head movements cause mislocalizing signals, which cause artifacts in the brain volume, that can also been seen on the center of each brain images on the axial view of subjects. The sagittal view of images that are obtained from *A1*, *A2*, *B*, and *C* data demonstrates high artifact intensity around the spinal cord (white rectangle on the sagittal views), is another indication of head movement related artifacts. It is interesting to see a similar pattern on the same source of artifacts on the same subject (*A1* & *A2*) from different visits. In contrast to the source of artifacts identified from *A1*, *A2*, *B*, and *C* data, the sagittal view of the subject *D* illustrates tissue activation around edge of the cerebellum (white circle on the sagittal views), which could be an indication of pulsatile artifact from blood being pumped in brain blood vessels with each heart beat. In addition to these conclusions, the subject-based color scales highlight the signal changes across subjects. For instance, subject *D* exhibit more intense artifacts compared with for subject *A*.

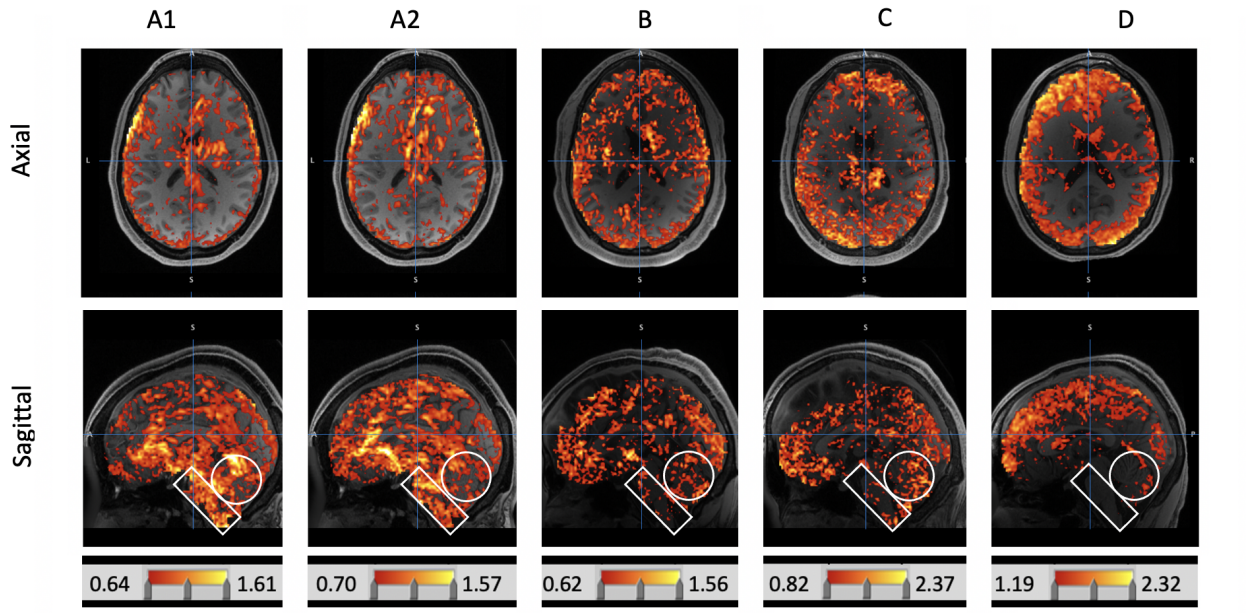


Figure 8: Absolute value of artifact intensity, averaged across outlying volumes. The bottom color scales indicate the subject-specific signal magnitudes values below the mean are excluded for better visualization. The white rectengular on each sagittal view shows the spinal cord of the brain. The white circle on each sagittal view highlights the cerebellum of the brain.

4 Discussion and Future Work

We have proposed a data-driven outlier detection method that is applicable to non-Normal and dependent data. Because the existing statistical outlier detection methods, which are based on theoretical distribution of RD, are more suitable for identical, independent, and normally distributed

data, these methods cannot be applied to fMRI data (Hardin and Rocke 2005; Filzmoser, Maronna, and Werner 2008; Cerioli 2010; Green and Martin 2014 ; Mejia et al. 2017). Violation of the assumptions causes the distribution of RD to deviate from the theoretical distribution, which can lead to an increase in FPR or a decrease in sensitivity to outliers, depending on the nature of the violations. These existing approaches, attempt to reduce the distributional deviation between the theoretical and empirical fit by using ad hoc methods. The proposed method instead uses a bootstrap procedure to estimate relevant summary statistics of the true distribution of RD and does not require any quantile matching.

The propose method offers a statistically principled alternative to common hardware-based methods to identify artifactual volumes in fMRI data. Specifically, "motion scrubbing" is a artifact detection approach in fMRI using frame-wise displacement, which is common all subjects. Motion scrubbing often employs an ad hoc but strict threshold value which could cause more than half of the volumes to be flagged as outliers and excluding potentially valuable data. Moreover, "motion scrubbing" can only detect artifacts coinciding with head motion and may therefore miss other types of artifacts. Because the fMRI data are acquired in different subjects, the signal intensities can vary in each individual. Therefore, a single threshold to identify artifactual volume could be inappropriate. Our proposed method, not only is applicable to non-normal and dependent data but also provides subject-specific threshold value. Although there are more general "data-driven scrubbing" techniques (Pham et al. 2021; Power et al. 2012), our method improves upon these, as a formal robust outlier detection framework that is appropriate for fMRI data.

Although the proposed method outperforms the existing methods in many aspects, it has several limitations. First, it utilizes a univariate outlier imputation by using the MAD factor for normally distributed data prior to the bootstrap even though the data show significant violation from normality. Another limitation of that imputation technique is to use neighboring observations of each column of data. To resolve this, we could use a more generalize outlier imputation approaches, which impute better by using information from other columns. Incorporating multiple imputation into the bootstrap samples could also provide better results. Second, during the development of the method, it has been observed that the determinant of MCD-based covariance estimation is underestimated in bootstrapped samples. This underestimation negatively affects the robust distance estimation for bootstrapped data. Therefore, we use the common MCD-based covariance estimate obtained from the original data. Third, we have not done a comparative analysis with any quantile matching approaches adopted by Mejia et al. (2017).

There are several exciting future opportunities to better understand the performance of our method. First, while the proposed method's excellent performance is demonstrated in simulated data, some additional tests are still needed. For example, the method could be tested on non-normal and outlier-contaminated simulated data. Second, the method has been applied to resting state fMRI data which contains task-free signals. But, the method could be applied and tested on task-related fMRI data by regressing out the task-related signals. Third, we employed data from one subject with two separate imaging sessions that showed very similar patterns for artifacts. This attribute could be studied for more subjects to explore whether there is a recognizable inter-individual variability in the spatial patterns of artifacts in fMRI datasets. Analyzing more subject would also allow us to better quantify the performance of our method compared with existing data-driven and software-based scrubbing techniques. Forth, since identification of eye movement related artifacts is still a hot topic, it is inspired to apply our method to a dataset where eye tracking is available to see if it is capable of detecting blinks and eye movement better than existing scrubbing techniques (Beauchamp 2003; Frey, Nau, and Doeller 2020).

References

Beauchamp, Michael S. 2003. "Detection of Eye Movements from fMRI Data." *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 49 (2): 376–80.

Beauchamp, Michael S, Kathryn E Lee, James V Haxby, and Alex Martin. 2003. "FMRI Responses to Video and Point-Light Displays of Moving Humans and Manipulable Objects." *Journal of Cognitive Neuroscience* 15 (7): 991–1001.

Cerioli, Andrea. 2010. "Multivariate Outlier Detection with High-Breakdown Estimators." *Journal of the American Statistical Association* 105 (489): 147–56.

Di Martino, Adriana, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, et al. 2014. "The Autism Brain Imaging Data Exchange: Towards a Large-Scale Evaluation of the Intrinsic Brain Architecture in Autism." *Molecular Psychiatry* 19 (6): 659–67.

Filzmoser, Peter, Ricardo Maronna, and Mark Werner. 2008. "Outlier Identification in High Dimensions." *Computational Statistics & Data Analysis* 52 (3): 1694–1711.

Frank, Lawrence R, Richard B Buxton, and Eric C Wong. 2001. "Estimation of Respiration-Induced Noise Fluctuations from Undersampled Multislice fMRI Data." *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 45 (4): 635–44.

Frey, Markus, Matthias Nau, and Christian F Doeller. 2020. "MR-Based Camera-Less Eye Tracking Using Deep Neural Networks." *BioRxiv*.

Friston, Karl J., Steven Williams, Robert Howard, Richard S. J. Frackowiak, and Robert Turner. 1996. "Movement-Related Effects in fMRI Time-Series." *Magnetic Resonance in Medicine* 35 (3): 346–55. https://doi.org/https://doi.org/10.1002/mrm.1910350312.

Glasser, Matthew F, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, et al. 2013. "The Minimal Preprocessing Pipelines for the Human Connectome Project." *Neuroimage* 80: 105–24.

Green, Christopher G, and R Douglas Martin. 2014. "An Extension of a Method of Hardin and Rocke, with an Application to Multivariate Outlier Detection via the IRMCD Method of Cerioli." Working Paper, 2017. Available from http://christopherggreen. github. io ….

Griffanti, Ludovica, Gholamreza Salimi-Khorshidi, Christian F Beckmann, Edward J Auerbach, Gwenaëlle Douaud, Claire E Sexton, Enikő Zsoldos, et al. 2014. "ICA-Based Artefact Removal and Accelerated fMRI Acquisition for Improved Resting State Network Imaging." *Neuroimage* 95: 232–47.

Hardin, Johanna, and David M Rocke. 2005. "The Distribution of Robust Distances." *Journal of Computational and Graphical Statistics* 14 (4): 928–46.

Krüger, Gunnar, and Gary H Glover. 2001. "Physiological Noise in Oxygenation-Sensitive Magnetic Resonance Imaging." *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 46 (4): 631–37.

Lindquist, Martin A. 2008. "The Statistical Analysis of fMRI Data." *Statistical Science* 23 (4): 439–64. https://doi.org/10.1214/09-STS282.

Lopuhaa, Hendrik P, and Peter J Rousseeuw. 1991. "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices." *The Annals of Statistics*, 229–48.

Mahalanobis, Prasanta Chandra. 1936. "On the Generalized Distance in Statistics." In. National Institute of Science of India.

Maronna, Ricardo A, and Ruben H Zamar. 2002. "Robust Estimates of Location and Dispersion for High-Dimensional Datasets." *Technometrics* 44 (4): 307–17.

Mejia, Amanda F, Mary Beth Nebel, Ani Eloyan, Brian Caffo, and Martin A Lindquist. 2017. "PCA Leverage: Outlier Detection for High-Dimensional Functional Magnetic Resonance Imaging Data." *Biostatistics* 18 (3): 521–36.

Pham, Damon, Daniel McDonald, Lei Ding, Mary Beth Nebel, and Amanda Mejia. 2021. "Projection Scrubbing: A More Effective, Data-Driven fMRI Denoising Method." http://arxiv.org/abs/2108.00319.

Power, Jonathan D, Kelly A Barnes, Abraham Z Snyder, Bradley L Schlaggar, and Steven E Petersen. 2012. "Spurious but Systematic Correlations in Functional Connectivity MRI Networks Arise from Subject Motion." *Neuroimage* 59 (3): 2142–54.

Power, Jonathan D, Anish Mitra, Timothy O Laumann, Abraham Z Snyder, Bradley L Schlaggar, and Steven E Petersen. 2014. "Methods to Detect, Characterize, and Remove Motion Artifact in Resting State fMRI." *Neuroimage* 84: 320–41.

Pruim, Raimon HR, Maarten Mennes, Daan van Rooij, Alberto Llera, Jan K Buitelaar, and Christian F Beckmann. 2015. "ICA-AROMA: A Robust ICA-Based Strategy for Removing Motion Artifacts from fMRI Data." *Neuroimage* 112: 267–77.

Rousseeuw, Peter J, and Christophe Croux. 1993. "Alternatives to the Median Absolute Deviation." *Journal of the American Statistical Association* 88 (424): 1273–83.

Rousseeuw, Peter J, and Bert C Van Zomeren. 1990. "Unmasking Multivariate Outliers and Leverage Points." *Journal of the American Statistical Association* 85 (411): 633–39.

Van Essen, David C, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, and others. 2013. "The WU-Minn Human Connectome Project: An Overview." *Neuroimage* 80: 62–79.