Robust outlier detection for high dimensional neuroimaging data with principal components analysis and split-half resampling.

Derek Beaton[1,2], Kelly M. Sunderland[1,2], Abiramy Uthirakumaran[1,2], Stephen R. Arnott[1,2], Robert Bartha[3], Sandra E. Black[4], Leanne Casaubon[5], Morris Freedman[2], Richard H. Swartz[4], Sean Symons[4], ONDRI Investigators[6], Malcolm A. Binns[1,2,7], and Stephen C. Strother[1,2,7]

[1]Rotman Research Institute
[2]Baycrest Health Sciences
[3]Robarts Research Institute
[4]Sunnybrook Health Sciences Centre
[5]Krembil Research Institute
[6]ONDRI
[7]University of Toronto

Abstract

The Ontario Neurodegenerative Disease Research Initiative (ONDRI) has collected hundreds of thousands of variables per participant across many data modalities. Such large and complex data across heterogeneous cohorts could be highly susceptible to outliers. To ensure high quality data and inference we need methods that can identify outliers and make robust estimates. Few such methods exist. We propose a novel framework based on principal components analysis (PCA) and split-half resampling (SHR) to identify outliers via Mahalanobis distance (MD) and robust subspaces. PCA+SHR allows us to: (1) make predictive estimates of MD in rank-deficient or highly collinear data in order to find outliers, and (2) identify a robust subspace through reproducibility estimates. We applied PCA+SHR to resting state functional neuroimaging data in one cohort of ONDRI participants (I=109 << J=32,768). Compared to existing techniques, PCA+SHR has many advantages: it provides multiple distance metrics with distributions to assess outlierness (e.g., average estimates or stability estimates), through SHR it provides a robust subspace, and it is a general purpose tool for outlier detection and robust subspaces because it can be used on data of any dimensionality. Code with examples available: https://www.github.com/derekbeaton/ours.

Key Words: principal components analysis, split-half, resampling, outliers, Mahalanobis distance, neurodegenerative diseases

1. Introduction

Outliers can have undue influence on results and thus skew interpretation. Undue influence is especially problematic in complex data such as functional neuroimaging, where the data are multivariate, have low signal-to-noise, and have high variability. The complexity of such multivariate data makes it difficult to perform quality control and outlier analyses; procedures that are easier and often more intuitive for univariate data. To make more reliable conclusions with complex, multivariate data we need tools that can both produce robust estimates and identify outliers. Most current tools generally provide either robust estimates or ways to identify outliers but not both. Herein we propose a framework based on principal components analysis (PCA) and split-half resampling (SHR; PCA+SHR for short) that provides robust estimates and multiple metrics for outlier identification.

The Ontario Neurodegenerative Disease Research Initiative (ONDRI) is a longitudinal, multi-site, "deep-phenotyping" project (Farhan et al., 2016a). The ONDRI project has unprecedented breadth of disorders and depth of data. ONDRI's breadth spans five neurodegenerative cohorts: Alzheimer's disease, vascular cognitive impairment, Parkinson's disease, amyotrophic lateral sclerosis, and frontotemporal dementia. ONDRI's depth spans genetics & genomics (Farhan et al., 2016b), brain structure & function, and neuropsychological, cognitive, behavioral, & clinical batteries, as well as data "between" these levels: ocular imaging, eye-tracking, gait & motor phenotypes (Montero-Odasso et al., 2017), and eventually neuropathology (see http://ondri.ca/assessments). Amongst the many assessment platforms and teams within ONDRI there also exists a neuroinformatics and biostatistics team whose roles and areas of responsibilities include but are not limited to data management and infrastructures as well as statistical tools and expertise. One of the most critical roles the neuroinformatics and biostatistics team plays is that we provide a quality control (QC) assessment and outlier identification as part of the release cycle, independently from the other platforms.

The neuroinformatics and biostatistics teams employ a variety of multivariate tools for the QC and outlier analyses. These include techniques such as the minimum covariance determinant (MCD; Hubert and Debruyne (2010)) and the Corr-Max transformation (Garthwaite and Koch, 2016). However, the ONDRI data are complex with a wide-variety of issues that require novel techniques, such as the generalized MCD (Beaton et al., 2018) for data of mixed types (e.g., categorical, continuous, ordinal). All the aforementioned techniques work only for low-dimensional (more observations than variables), full-rank data with little-to-no collinearity because they depend on the ability to compute a Mahalanobis distance (MD). MD only exists for data that are full rank on the columns. When data are high dimension-low sample size (HDLSS) then MD is no longer defined. Thus many individual data sets (e.g., almost any of the neuroimaging measures) and any practical mixtures of data (e.g., neuroimaging + neuropsychology assessments) require different techniques. For HDLSS data an analog to the MCD is robust PCA (ROBPCA; Hubert et al. (2005)), and more recently a novel approach based on leverage (Wold et al., 1987) was proposed for functional neuroimaging (Mejia et al., 2017). However both ROBPCA and the leverage approach rely on alternate distance measures (e.g., score distances, orthogonal distance, partial MD a.k.a. leverage) because, as previously noted, MD is undefined for high-dimensional or rank-deficient data and techniques such as MCD and Corr-Max breakdown for highly collinear data. Here we show that we can in fact compute a "predicted" MD, as well as a variety of other predicted distances, through the PCA+SHR framework. PCA+SHR is a more generalized approach than, and overcomes many of the limitations of, other techniques.

Our paper is outlined as follows. First we introduce the required notation and background, where we define MD, explain PCA, and show how the two are connected, and then explain PCA+SHR with its various distances. Next, we illustrate the PCA+SHR framework on resting state functional MRI (rsfMRI) data from the Alzheimer's disease/mild cognitive impairment cohort in ONDRI. We compare distances obtained PCA+SHR to ROBPCA and Mejia et al.'s leverage approach. Finally, we discuss advantages, extensions, and limitations of PCA+SHR.

2. PCA+SHR for outlier detection and robust estimates

2.1 Notation

Bold uppercase letters denote matrices (e.g., $\mathbf{X}$), bold lowercase letters denote vectors (e.g., $\mathbf{x}$), and italic lowercase letters denote specific elements (e.g., x). Upper case italic letters denote cardinality, size, or length (e.g., I) where a lower case italic denotes a specific index (e.g., i). A generic element of $\mathbf{X}$ would be denoted as $x_{i,j}$. Common letters of varying type faces for example $\mathbf{X}$, $\mathbf{x}$, $x_{i,j}$ come from the same data structure. Vectors are assumed to be column vectors unless otherwise specified. Two matrices side-by-side denote standard matrix multiplication (e.g., $\mathbf{XY}$), while $\odot$ denotes element-wise multiplication. The matrix $\mathbf{I}$ denotes the identity matrix. The diagonal operation, diag$\{\}$, when given a vector will transform it into a diagonal matrix, or when given a matrix, will extract the diagonal elements as a vector. Superscript ^{T} denotes the transpose operation, superscript $^{-1}$ denotes standard matrix inversion, and superscript $^+$ denotes the Moore-Penrose pseudo-inverse. Throughout the following sections we typically refer to the data matrix of interest as $\mathbf{X}$ with I rows and J columns and $\mathbf{X}$ could be in one of two possible conditions: (1) full rank (on the columns) where $I > J$ or (2) rank deficient where $I < J$.

2.2 Mahalanobis distance

Let $\mathbf{X}$ be a column-wise centered data matrix that is full rank where we define its covariance matrix as $\mathbf{C} = (\mathbf{X}^T\mathbf{X}) \times (I-1)^{-1}$. Squared Mahalanobis distances (MDs) are usually defined as $\mathbf{m} = (\mathbf{X}\mathbf{C}^{-1} \odot \mathbf{X})\mathbf{1}$ where $\mathbf{1}$ is a conformable $J \times 1$ vector of ones and $\mathbf{m}$ is a $I \times 1$ column vector (where $\mathbf{m}^T$ is a row vector) of squared MDs. If we relax the definition of squared MD to exclude the degrees of freedom scaling factor in the covariance as $\mathbf{C'} = \mathbf{X}^T\mathbf{X}$, then:

$$\mathbf{M'} = \mathbf{X}\mathbf{C'}^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \tag{1}$$

where $\mathbf{m'} = \text{diag}\{\mathbf{M'}\}$ and $\text{diag}\{\mathbf{M'}\} = \text{diag}\{\mathbf{M}\} \times (I-1)$. Henceforth when we refer to squared MD it is the definition we provide in Eq. 1. Generally, MD is only defined for data that are full rank because when data are (column-wise) rank deficient, $\mathbf{X}$ is not invertible.

2.3 Principal components analysis and distances

The principal components analysis (PCA) of $\mathbf{X}$ that is column-wise centered and/or normalized where $\mathbf{X}$ is of rank L is performed through the singular value decomposition (SVD) as:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{2}$$

where the following properties hold regardless of rank:

1. $\mathbf{U}$ and $\mathbf{V}$ are orthonormal left and right singular vectors of sizes $I \times L$ and $J \times L$, respectively with $\mathbf{U}^T\mathbf{U} = \mathbf{I} = \mathbf{V}^T\mathbf{V}$.

2. $\mathbf{\Sigma}$ is the $L \times L$ diagonal matrix of singular values and $\mathbf{\Lambda} = \mathbf{\Sigma}\mathbf{\Sigma}$ is the diagonal matrix of eigenvalues (squared singular values).

In PCA there are two sets of component scores: $\mathbf{F}_I = \mathbf{U}\mathbf{\Sigma}$ for rows and $\mathbf{F}_J = \mathbf{V}\mathbf{\Sigma}$ for columns. The component scores can also be defined through projections or rotations as:

$$\begin{aligned}\mathbf{F}_I &= \mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V} = \mathbf{U}\mathbf{\Sigma} \\ \mathbf{F}_J &= \mathbf{X}^T\mathbf{U} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U} = \mathbf{V}\mathbf{\Sigma}.\end{aligned} \tag{3}$$

The first set of distances we can obtain from PCA are score distances (SD) computed as $\mathbf{s} = \text{diag}\{\mathbf{S}\}$ where $\mathbf{S} = \mathbf{F}_I\mathbf{F}_I^T$. We can also compute $\mathbf{U}$ and $\mathbf{V}$ through projections akin to Eq. 3:

$$\begin{aligned}\mathbf{U} &= \mathbf{X}\mathbf{V}\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}^{-1} \\ \mathbf{V} &= \mathbf{X}^T\mathbf{U}\mathbf{\Sigma}^{-1} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}^{-1},\end{aligned} \tag{4}$$

because $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{\Sigma}\mathbf{\Sigma}^{-1} = \mathbf{I}$. When $\mathbf{X}$ is full rank (i.e., $\mathbf{X}$ is $I \times J$ where $I > J$), there are some specific additional properties of PCA:

1. $\mathbf{V}\mathbf{V}^T = \mathbf{I}$ and
2. $\mathbf{U}\mathbf{U}^T = \mathbf{M}'$.

The second property shows that PCA provides squared MD defined in Eq. 1 through the left singular vectors (i.e., the I rows or observations):

$$\begin{aligned}\mathbf{M}' = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \\ \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T = \\ \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T = \\ \mathbf{U}\mathbf{\Sigma}\Lambda^{-1}\mathbf{\Sigma}\mathbf{U}^T = \mathbf{U}\mathbf{U}^T.\end{aligned} \tag{5}$$

In the case of $I > L$: $\mathbf{V}^T = \mathbf{V}^{-1}$, $(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T)^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T$, and $\mathbf{C}' = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. This alternative form of MD is also sometimes referred to as leverage (Wold et al., 1987; Mejia et al., 2017; Beaton et al., 2018). When $\mathbf{X}$ is full rank the use of all components provides MD as defined above. However some of the aforementioned properties of PCA—especially computing squared MD from PCA—no longer hold when $\mathbf{X}$ is rank deficient (i.e., $\mathbf{X}$ is $I \times J$ where $I < J$):

1. $\mathbf{U}\mathbf{U}^T = \mathbf{I} \times (\frac{1}{I-1})$ and
2. $\mathbf{V}\mathbf{V}^T \neq \mathbf{I}$.

We can now see why MD is undefined for rank deficient data: $\mathbf{U}\mathbf{U}^T$ produces identical values for all items. However, even in the case of rank deficient data, there is a way to estimate a MD-like value through a predictive framework.

2.4 Predicted MD for any rank

With PCA we can compute MD of supplementary (a.k.a. left-out) observations. Here we generally use the nomenclature from the French literature (Holmes and Josse, 2017; Abdi and Williams, 2010) where "active data" or "active observations" refers to analyzed observations and "supplementary data" or "supplementary observations" refers to external observations measured on the same variables as the active set.

Given two data sets: (1) the active data set $\mathbf{X}$ that is $I \times J$ (2) the supplementary data set $\mathbf{Y}$ which is $K \times J$; the J columns (variables) in the active and supplementary

data are the same but the I and K rows (observations) are assumed to be unique. In general we require that $\mathbf{Y}$ has been centered and/or normalized by the same center or normalization factors as $\mathbf{X}$ (e.g., $\mathbf{Y}$ is column-wise centered by the column means of $\mathbf{X}$). We can compute "supplementary component scores" for $\mathbf{Y}$ akin to Eq. 3 as:

$$\mathbf{F}_K = \mathbf{YV}. \tag{6}$$

We can also compute "supplementary (left) singular vectors" for $\mathbf{Y}$ akin to Eq. 4 as:

$$\widehat{\mathbf{U}}_\mathbf{Y} = \mathbf{F}_K\mathbf{\Sigma}^{-1} = \mathbf{YV}\mathbf{\Sigma}^{-1}. \tag{7}$$

Thus we can also compute the "supplementary MD" of $\mathbf{Y}$ as $\widehat{\mathbf{m}'}_\mathbf{Y} = \text{diag}\{\widehat{\mathbf{M}'}_\mathbf{Y}\}$ where $\widehat{\mathbf{M}'}_\mathbf{Y} = \widehat{\mathbf{U}}_\mathbf{Y}\widehat{\mathbf{U}}_\mathbf{Y}^T$. This formulation is equivalent to the standard MD computation for supplementary observations, under the usual assumption that $\mathbf{X}$ is full rank:

$$\begin{aligned}
\widehat{\mathbf{M}'}_\mathbf{Y} = \mathbf{Y}\mathbf{C}'^{-1}\mathbf{Y}^T = \\
\mathbf{Y}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y}^T = \\
\mathbf{Y}(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^{-1}\mathbf{Y}^T = \\
\mathbf{Y}(\mathbf{V}\mathbf{\Sigma}\mathbf{\Sigma}\mathbf{V}^T)^{-1}\mathbf{Y}^T = \\
\mathbf{YV}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}^{-1}\mathbf{V}^T\mathbf{Y}^T = \\
(\mathbf{YV}\mathbf{\Sigma}^{-1})(\mathbf{YV}\mathbf{\Sigma}^{-1})^T = \widehat{\mathbf{U}}_\mathbf{Y}\widehat{\mathbf{U}}_\mathbf{Y}^T.
\end{aligned} \tag{8}$$

where $\widehat{\mathbf{U}}_\mathbf{Y} = \mathbf{YV}\mathbf{\Sigma}^{-1}$ is Eq. 7 and thus Eq. 8 shows we can compute supplementary (predicted) squared MDs of $\mathbf{Y}$ as $\widehat{\mathbf{m}'}_\mathbf{Y} = \text{diag}\{\mathbf{M}'_\mathbf{Y}\}$ from $\widehat{\mathbf{U}}_\mathbf{Y}$.

In the case of $\mathbf{M}'_\mathbf{Y} = \mathbf{Y}\mathbf{C}'^{-1}\mathbf{Y}^T$ we generally assume that $\mathbf{C}'$ is invertible and thus $\mathbf{X}$ is full rank. However, the use of PCA here shows that we do not require full rank for either $\mathbf{X}$ or $\mathbf{Y}$. Let us assume that $\mathbf{X}$ is rank deficient; the rank of $\mathbf{Y}$ is irrelevant. With PCA we know that the squared MDs of $\mathbf{X}$ are all equal: $\mathbf{M}' = \mathbf{U}\mathbf{U}^T = \mathbf{I} \times (\frac{1}{I-1})$ and thus not informative. However Eq. 8 shows that we can compute unique predicted squared MDs for $\mathbf{Y}$ when $\mathbf{X}$ is rank-deficient because the use of PCA here is effectively the use of the pseudo-inverse. Say we have some general matrix $\mathbf{A}$. We compute the psuedo-inverse of $\mathbf{A}$ through the SVD as:

$$\mathbf{A}^+ = \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{V}^T \tag{9}$$

Therefore we can consider Eq. 8 a more general approach where:

$$\begin{aligned}
\widehat{\mathbf{M}'}_\mathbf{Y} = \widehat{\mathbf{U}}_\mathbf{Y}\widehat{\mathbf{U}}_\mathbf{Y}^T = \\
(\mathbf{YV}\mathbf{\Sigma}^{-1})(\mathbf{YV}\mathbf{\Sigma}^{-1})^T = \\
\mathbf{YV}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}^{-1}\mathbf{V}^T\mathbf{Y}^T = \\
\mathbf{YV}\mathbf{\Lambda}^{-1}\mathbf{V}^T\mathbf{Y}^T = \\
\mathbf{Y}(\mathbf{X}^T\mathbf{X})^+\mathbf{Y}^T = \\
\mathbf{Y}\mathbf{C}'^+\mathbf{Y}^T.
\end{aligned} \tag{10}$$

Recall that $\mathbf{C}' = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ and though $\mathbf{X}$ may not be full rank and thus not invertible, we can use the pseudo-inverse—as defined in Eq. 9—for $\mathbf{C}'$ in Eq. 10 in order to compute predicted squared MDs for $\mathbf{Y}$.

In summary, with PCA we can compute:

1. squared MDs for full rank data directly from the (left) singular vectors, and
2. predicted squared MDs for supplementary data ($\mathbf{Y}$) of any rank conditional to active data ($\mathbf{X}$) of any rank.

PCA provides the first of two constituent pieces of a framework to compute predictive squared MDs even in the presence of rank-deficient data. In order to obtain predicted MD estimates—regardless of rank—a second constituent piece of this framework: split-half resampling.

2.4.1 Predicted MD from split-half resampling

Split-half resampling (SHR) is akin to a repeated two-fold cross-validation (Strother et al., 2002). First we introduce SHR as a single two-fold cross-validation step to highlight its connection with the predicted MDs outlined in Eq. 10. To illustrate split-half resampling we first show a single split-half PCA step. We first arbitrarily split $\mathbf{X}$ into two (approximately) equal subsets of observations and refer to each split-half of $\mathbf{X}$ as $\mathbf{X}_1$ ("split 1") and $\mathbf{X}_2$ ("split 2"). At this point, the rank of $\mathbf{X}_1$ and $\mathbf{X}_2$ are irrelevant. In order to compute a predicted squared MD for each half from the other half, we first need to perform the PCA on each split half separately:

$$\begin{aligned}\mathbf{X}_1 &= \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T \\ \mathbf{X}_2 &= \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^T\end{aligned} \tag{11}$$

Next, we center and normalize each split by the centering and scaling factors of the other split. Let us assume that we have only performed column-wise mean centering: split 1 is now centered by split 2's center and vice versa. These splits that are normalized by each other's factors are referred to as $\widetilde{\mathbf{X}}_1$ ($\mathbf{X}_1$ centered and normalized by $\mathbf{X}_2$'s center and scale) and $\widetilde{\mathbf{X}}_2$ ($\mathbf{X}_2$ centered and normalized by $\mathbf{X}_1$'s center and scale). We can compute predicted squared MDs for each split by way of Eqs. 3, 4, 8, and 10:

$$\begin{aligned}\widehat{\mathbf{M}'}_1 &= (\widetilde{\mathbf{X}}_1\mathbf{V}_2\mathbf{\Sigma}_2^{-1})(\mathbf{X}_1\mathbf{V}_2\mathbf{\Sigma}_2^{-1})^T = \mathbf{X}_1\mathbf{V}_2\mathbf{\Lambda}_2^{-1}\mathbf{V}_2^T\mathbf{X}_1^T \\ \widehat{\mathbf{M}'}_2 &= (\widetilde{\mathbf{X}}_2\mathbf{V}_1\mathbf{\Sigma}_1^{-1})(\mathbf{X}_2\mathbf{V}_1\mathbf{\Sigma}_1^{-1})^T = \mathbf{X}_2\mathbf{V}_1\mathbf{\Lambda}_1^{-1}\mathbf{V}_1^T\mathbf{X}_2^T\end{aligned} \tag{12}$$

where $\widehat{\mathbf{m}'}_1 = \text{diag}\{\widehat{\mathbf{M}'}_1\}$ and $\widehat{\mathbf{m}'}_2 = \text{diag}\{\widehat{\mathbf{M}'}_2\}$ are the predicted MDs from each set and $\widehat{\mathbf{m}'} = \begin{bmatrix}\widehat{\mathbf{m}'}_1 \\ \widehat{\mathbf{m}'}_2\end{bmatrix}$. Regardless of rank we can compute a predicted squared MD through this splitting procedure. However, a single instance of split-halves to predict squared MDs is not sufficient; we require resampling. Like other repeated resampling or cross-validation schemes, SHR is a performed many times (e.g., 1,000). If we want to compute predicted squared MDs in data of any rank, we can generate many (e.g., 1,000) predicted estimates and build distributions. With the distributions provided by SHR we can compute a single estimate (e.g., mean or median) of predicted MDs or use the entire distributions (e.g., percentile or confidence intervals). The same splitting and prediction procedure as presented in Eq. 12 can be used to compute predicted component scores (i.e., $\widehat{F_I}$) and predicted SDs, referred to as $\widehat{\mathbf{s}} = \begin{bmatrix}\widehat{\mathbf{s}}_1 \\ \widehat{\mathbf{s}}_2\end{bmatrix}$ (cf. Eq. 3).

2.5 Reproducibility estimates and orthogonal distances

With SHR we can also obtain reproducibility estimates to identify a robust subspace, and from the robust subspace we can compute orthogonal distances (OD). For each iteration, we can compute a predicted set of (row) singular vectors (cf. Eqs. 4, 7, and 12): $\widehat{\mathbf{U}} = \begin{bmatrix} \widehat{\mathbf{U}}_1 = \widetilde{\mathbf{X}}_1 \mathbf{V}_2 \boldsymbol{\Sigma}_2^{-1} \\ \widehat{\mathbf{U}}_2 = \widetilde{\mathbf{X}}_2 \mathbf{V}_1 \boldsymbol{\Sigma}_1^{-1} \end{bmatrix}$. To compute reproducibility (i.e., similarity between predicted and original values or between splits) we have two approaches: (1) compute the correlations between the original components and the SHR predicted components—i.e., cor($\mathbf{U}$, $\widehat{\mathbf{U}}$) or (2) compute the correlations between the split loadings—i.e., cor($\mathbf{V}_1$, $\mathbf{V}_2$). Through either approach, we can obtain the average (mean or median) similarity between components and thus identify which components were most reproducible over the course of all SHR iterations. We can then use this robust subspace—i.e., first N reproducible components where $N < L$—to compute orthogonal distances (OD; Hubert et al. (2005)). With a reproducible subspace, we can compute a robust version of $\mathbf{X}$ via Eq. 2:

$$\mathbf{X}_{1:N} = \mathbf{U}_{1:N} \boldsymbol{\Sigma}_{1:N} \mathbf{V}_{1:N}^T \tag{13}$$

computed from the first N reproducible components, and then compute OD as

$$\mathbf{o} = \{(\mathbf{X} - \mathbf{X}_{1:N}) \odot (\mathbf{X} - \mathbf{X}_{1:N})\}\mathbf{1} = (\mathbf{X}_{N:L} \odot \mathbf{X}_{N:L})\mathbf{1} \tag{14}$$

where $\mathbf{1}$ is a conformable matrix of ones where $\mathbf{X}_{N:L} = \mathbf{U}_{N:L} \boldsymbol{\Sigma}_{N:L} \mathbf{V}_{N:L}^T$; $\mathbf{X}_{N:L}$ can be thought of as the residuals after the removal of a robust subspace. The OD is the distance between an observation (row) and itself as the difference between the original data (i.e., $\mathbf{X}$) and a robust representation of the data (i.e., $\mathbf{X}_{1:N}$); alternatively OD is the row-wise sum of the squared residuals after the removal of a robust subspace (i.e., $\mathbf{o} = (\mathbf{X}_{N:L} \odot \mathbf{X}_{N:L})\mathbf{1}$).

2.6 PCA + SHR summary

In summary, PCA with SHR provides us three separate distance metrics—Mahalanobis, score, and orthogonal distances—that each provide a unique perspective to identify outliers. Furthermore, with SHR we can identify a robust subspace and effectively identify the components that are most likely signal (as opposed to noise). Finally, because SHR is a resampling technique, everything that we generate during the SHR procedure produces distributions (e.g., MDs and SDs) and we can use those distributions in a variety of ways to assess outlyingness of observations (e.g., mean predicted MD, width of confidence intervals for SD).

3. Results

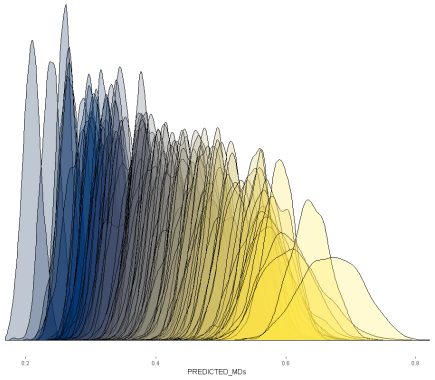
We illustrate the PCA+SHR framework on rsfMRI data in ONDRI from the Alzheimer's disease/mild cognitive impairment (ADMCI) cohort. Preliminary rsfMRI data from the ADMCI cohort included 109 individuals. All participant's rsfMRI data were warped to a common template. The 4D rsfMRI data were preprocessed with Optimization of Preprocessing Pipelines for NeuroImaging (OPPNI; Churchill et al. (2012a,b)) for "seed" based connectivity analyses. The seed was the posterior cingulate cortex. All connectivity values were transformed to Z-scores for each voxel. Each participant's (3D) brain was then vectorized and aggregated into a matrix

with individual participants on the rows (I) and vectorized voxels on the columns (J). The resulting matrix was $I = 109 \times J = 32,768$.

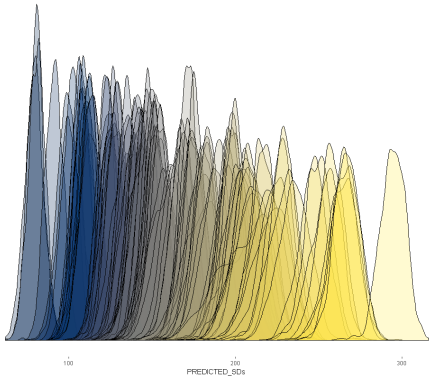
All analyses and visualizations were performed in `R` (R Core Team, 2017). Visualizations were created with `base` graphics as well as the `corrplot` (Wei and Simko, 2017), `psych` (Revelle, 2018), and `ggplot2` (Wickham, 2009) packages. We have made PCA+SHR available in the *Ou*tliers and *R*obust *S*tructures (`OuRS`) R package, available here: https://www.github.com/derekbeaton/ours (Sunderland and Beaton, 2018). First, we show the results of PCA+SHR with respect to the distance and reproducibility estimates produced. Next, we discuss several strategies to determine which individuals should be regarded as outliers. Finally, we compare the distances produced from PCA+SHR to ROBPCA via the `robustbase` package (Maechler et al., 2017) and leverage via the `CLEVER` package (Mejia, 2018). For all analyses, the data were column-wise centered but not scaled.

3.1 PCA+SHR in action

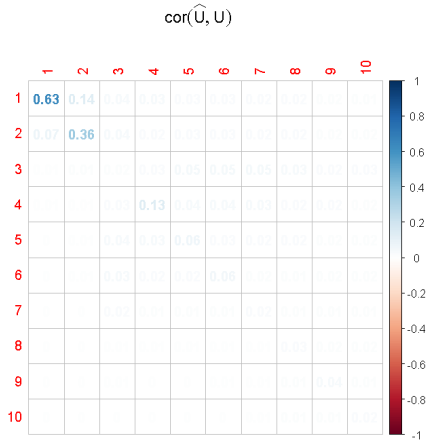
Resampling for PCA+SHR was performed $1,000$ times. Results for PCA+SHR are visualized in Figure 1. PCA+SHR produces $1,000$ predicted MDs and $1,000$ predicted SD values for each observation. These are visualized as distributions in Figures 1a and 1b. PCA+SHR also produces $1,000$ reproducibility estimates *per component*. For simplicity, here we only show and discuss the reproducibility estimates produced for the observations: $\text{cor}(\mathbf{U}, \widehat{\mathbf{U}})$. We visualize the reproducibility estimates as the median absolute value of correlations between all components in Figure 1c. There are no standards or rules to determine *a priori* a cutoff for the number of robust components (e.g., akin to α), thus analysts must rely on expert knowledge about the data and its expected structure. In Fig. 1c the reproducibility estimates suggest that there are two robust components that we should retain. Finally, we can compute the ODs as the difference between the original data—$\mathbf{X}$—and a robust version of the data, which was reconstructed from just the first two components—$\mathbf{X}_{1:2}$. ODs are visualized in a scatterplot matrix and compared to the median MDs and median SDs, shown in Figure 1d.



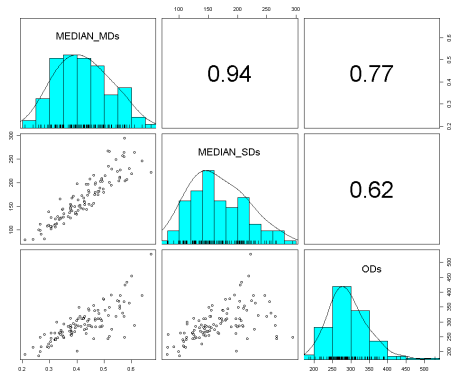
(a) Predicted MD distributions for every observation. Observations rank ordered by median predicted MD from left to right.



(b) Predicted SD distributions for every observation. Observations rank ordered by median predicted SD from left to right.



(c) Reproducibility estimates. Only the first two components have median absolute values above a reasonable threshold given the type of data and the expected multivariate structure for rsfMRI.



(d) The orthogonal distances between the orginal data and low-rank robust version of the data

Figure 1: An overview of the three distance estimates and one of the reproducibility estimates produced by PCA+SHR.

3.2 Outlier identification

Figures 1a and 1b show distributions for predicted MD and SD generally have wider distributions when distances tend to be larger (relative to the rest of the sample). To identify outliers from predicted MD and SD, we use the entire set of predicted distances to identify a cut-off. We can then define outliers based on quantile estimates from these distributions. In a similar fashion, we use quantile estimates to identify OD outliers from the ODs. In the case of ODs we have two options: (1) use the set of observations and compute the quantiles directly, or (2) bootstrap these distances and build a distribution akin to the SHR distributions for MD and SD; we prefer the latter. By default we use the 75%-ile of all distances: any of the observation (OD), or any individual observations' distribution (MD and SD), that are above the cutoff are considered outliers for that distance. Observations can be characterized as outliers for one or several metrics. We prefer to retain

the outlier estimates for all distances but conservatively identify observations as outliers if they are outliers for any of the distances. However, given the amount of information PCA+SHR provides, there are numerous alternatives to identify outliers; for examples quantiles of mean (or median) estimates, or even through the use of all pairwise distances between distributions (e.g., Kolmogorov-Smirnov or Hellinger distances).

3.3 Comparisons

Here we compare and contrast the results of PCA+SHR with two other high dimensional outlier techniques: ROBPCA (Hubert et al., 2005) and the Mejia et al., leverage approach (a.k.a. "CLEVER"). Though the three approaches (ROBPCA, CLEVER, PCA+SHR) are different, we use default parameters of their respective software for these comparisons.

Figure 2 shows a comparison of the distance estimates provided by each technique. In general, low similarity is observed between the two ROBPCA distances, as well as between the CLEVER distances and distances from ROBPCA and PCA+SHR. The MD and SD estimates from PCA+SHR for this example are highly similar ($r = .94$).

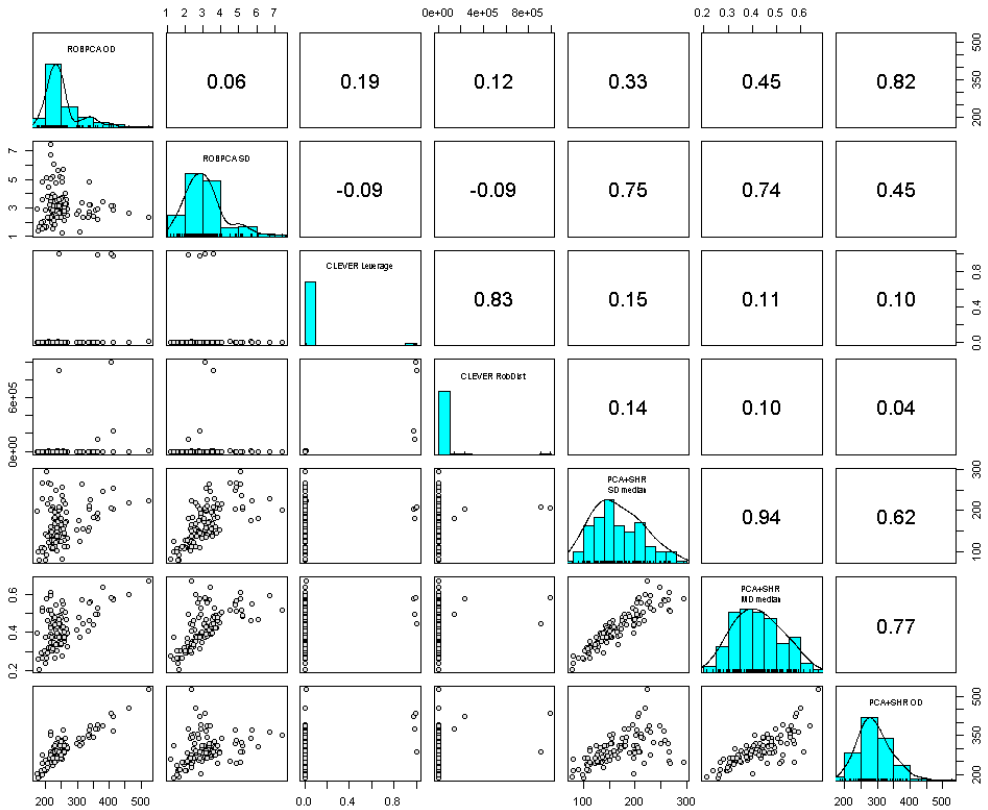


Figure 2: Scatterplot matrix of distances from ROBPCA, CLEVER, and PCA+SHR. Correlations are Spearman rank correlations. Distances are ordered as follows: ROBPCA orthogonal distance (OD), ROBPCA score distance (SD), CLEVER leverage, CLEVER robust distance, PCA+SHR SD, PCA+SHR Mahalanobis distance (MD), and PCA+SHR OD.

Table 1 shows the number of common outliers across the techniques. Because these techniques work in different ways, they also have different approaches to determine outliers; however, most of those are through some form of a distributional cutoff. ROBPCA and CLEVER use parametric estimates whereas PCA+SHR uses quantile estimates from the resampling procedure. Similar to the distance estimates, the two ROBPCA approaches find different outliers. However, the CLEVER robust distance (RobDist) estimate is a subset of the CLEVER leverage estimate. ROBPCA shares fewer outliers overall than with the other methods. Finally, PCA+SHR OD has the most overlap with the other approaches which suggests that PCA+SHR OD is a suitable proxy for all other outlier estimates.

Table 1: The number of common outliers across ROBPCA, CLEVER, and PCA+SHR. Default parameters used for the packages. Upper and lower triangles are identical but included for convenience. The highest number of outliers were identified by CLEVER leverage estimates. ROBPCA has little overlap with other methods including and especially itself. PCA+SHR OD has the most overlap across all outlier thresholds.

	ROBPCA		CLEVER		PCA+SHR		
	SD	OD	Leverage	RobDist	SD	MD	OD
ROBPCA SD	14	1	6	3	9	9	7
ROBPCA OD	1	20	12	8	8	12	18
CLEVER Leverage	6	12	44	26	10	10	15
CLEVER RobDist	3	8	26	26	4	5	11
PCA+SHR SD	9	8	10	4	22	17	12
PCA+SHR MD	9	12	10	5	17	22	16
PCA+SHR OD	7	18	15	11	12	16	28

4. Discussion

We presented a novel outlier identification approach that uses both PCA and SHR to build distributions of predicted MD and SD. Also, through the resampling process PCA+SHR provides reproducibility estimates per component which allows us to compute a robust version of the data and thus ODs (between the original and robust data). Because of the distributions created through resampling, PCA+SHR provides an important and detailed information to help analysts and researchers identify outliers and anomalies which could impart undue influence because of unique patterns or even because of data errors.

Our work here had two aims: (1) introduce, formalize, and explain how to use PCA+SHR, and (2) compare PCA+SHR with other outlier techniques for high dimensional data; in this case we chose ROBPCA (a well-established approach) and CLEVER (a relatively novel multivariate outlier technique). We compared sets of distances and outlier thresholds across PCA+SHR, ROBPCA, and CLEVER. In general we found that (1) distance types within ROBPCA do not identify the same observations, (2) CLEVER is relatively conservative and identifies many individuals where the "robust distance" approach is a subset of the "leverage" approach (for our data), and (3) the OD from PCA+SHR is an effective general purpose approach to robust distance estimates and outlier detection.

PCA+SHR has one major limitation: it is relatively slow (by comparison to ROBPCA and CLEVER), but PCA+SHR could be sped up through paralleliza-

tion. Though PCA+SHR is comparably slow, the information obtained from the procedure is far more detailed than other procedures and thus valuable for outlier identification.

There are two future directions for PCA+SHR: (1) introduce a parameter to allow for the splits to be of any size in a two-fold framework: instead of approximately 50% for each split, SHR could be expanded to have uneven splits such as 60/40 or even 90/10, and (2) expand the PCA+SHR framework to mixed data types through correspondence analysis, akin to our recent work to generalize the minimum covariance determinant algorithm (Beaton et al., 2018).

With large scale and complex data, the inclusion of any outlier procedure is vital in order to maintain data integrity. Such techniques make easier the discovery of errors and anomalies. Though we illustrated PCA+SHR on high dimensional data, PCA+SHR works for data of any dimensionality. Typically with multivariate outlier detection, the researcher or analyst must make a decision to use techniques designed for full-rank data (e.g., minimum covariance determinant) or for more complex HDLSS data (e.g., ROBPCA). Because PCA+SHR works for any dimensionality, PCA+SHR is thus a general-purpose approach for outlier detection which makes it a better candidate for inclusion in quality control piplines than other techniques.

References

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

Beaton, D., Sunderland, K. M., Adni, Levine, B., Mandzia, J., Masellis, M., Swartz, R. H., Troyer, A. K., Ondri, Binns, M. A., Abdi, H., and Strother, S. C. (2018). Generalization of the minimum covariance determinant algorithm for categorical and mixed data types. *bioRxiv*, page 333005.

Churchill, N. W., Oder, A., Abdi, H., Tam, F., Lee, W., Thomas, C., Ween, J. E., Graham, S. J., and Strother, S. C. (2012a). Optimizing preprocessing and analysis pipelines for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods. *Human Brain Mapping*, 33(3):609–627.

Churchill, N. W., Yourganov, G., Oder, A., Tam, F., Graham, S. J., and Strother, S. C. (2012b). Optimizing Preprocessing and Analysis Pipelines for Single-Subject fMRI: 2. Interactions with ICA, PCA, Task Contrast and Inter-Subject Heterogeneity. *PLOS ONE*, 7(2):e31147.

Farhan, S. M. K., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., Greenberg, B., Grimes, D. A., Hegele, R. A., Hudson, C., Kleinstiver, P. W., Lang, A. E., Masellis, M., McIlroy, W. E., McLaughlin, P. M., Montero-Odasso, M., Munoz, D. G., Munoz, D. P., Strother, S., Swartz, R. H., Symons, S., Tartaglia, M. C., Zinman, L., and Strong, M. J. (2016a). The Ontario Neurodegenerative Disease Research Initiative (ONDRI). *Canadian Journal of Neurological Sciences*, pages 1–7.

Farhan, S. M. K., Dilliott, A. A., Ghani, M., Sato, C., Liang, E., Zhang, M., McIntyre, A. D., Cao, H., Racacho, L., Robinson, J. F., Strong, M. J., Masellis, M., St George-Hyslop, P., Bulman, D. E., Rogaeva, E., and Hegele, R. A. (2016b). The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *Genomic Medicine*, 1:16032.

Garthwaite, P. H. and Koch, I. (2016). Evaluating the Contributions of Individual Variables to a Quadratic Form. *Australian & New Zealand Journal of Statistics*, 58(1):99–119.

Holmes, S. and Josse, J. (2017). Discussion of "50 Years of Data Science". *Journal of Computational and Graphical Statistics*, 26(4):768–769.

Hubert, M. and Debruyne, M. (2010). Minimum covariance determinant. *Wiley interdisciplinary reviews: Computational statistics*, 2(1):36–43.

Hubert, M., Rousseeuw, P. J., and Branden, K. V. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, 47(1):64–79.

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., and Anna di Palma, M. (2017). *robustbase: Basic Robust Statistics (version 0.92-8)*. http://robustbase.r-forge.r-project.org/.

Mejia, A. (2018). *CLEVER: principal Components LEVERage and robust distance*. https://github.com/mandymejia/clever.

Mejia, A. F., Nebel, M. B., Eloyan, A., Caffo, B., and Lindquist, M. A. (2017). PCA leverage: outlier detection for high-dimensional functional magnetic resonance imaging data. *Biostatistics*, 18(3):521–536.

Montero-Odasso, M., Pieruccini-Faria, F., Bartha, R., Black, S. E., Finger, E., Freedman, M., Greenberg, B., Grimes, D. A., Hegele, R. A., Hudson, C., Kleinstiver, P. W., Lang, A. E., Masellis, M., McLaughlin, P. M., Munoz, D. P., Strother, S., Swartz, R. H., Symons, S., Tartaglia, M. C., Zinman, L., Strong, M. J., and McIlroy, W. (2017). Motor Phenotype in Neurodegenerative Disorders: Gait and Balance Platform Study Design Protocol for the Ontario Neurodegenerative Research Initiative (ONDRI). *Journal of Alzheimer's Disease*, 59(2):707–721.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research (version 1.8.3)*. Northwestern University, Evanston, Illinois. https://CRAN.R-project.org/package=psych.

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., and Rottenberg, D. (2002). The Quantitative Evaluation of Functional Neuroimaging Experiments: The NPAIRS Data Analysis Framework. *NeuroImage*, 15(4):747–771.

Sunderland, K. and Beaton, D. (2018). *OuRS: Outliers and Robust Structures*. http://github.com/derekbeaton/ours.

Wei, T. and Simko, V. (2017). *R package "corrplot": Visualization of a Correlation Matrix*. https://github.com/taiyun/corrplot.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52.