Robust Outlier Detection

for Low and High-Dimensional Neuroimaging Data with Principal Components Analysis and Split-Half Resampling

Derek Beaton

Rotman Research Institute, Baycrest Health Sciences

Twitter: @derek__beaton (that's two underscores!)
Github for today: http://github.com/derekbeaton/ours

August 01, 2018

The other authors

Kelly M Sunderland, Abiramy Uthirakumaran, Stephen R Arnott, Robert Bartha, Sandra E Black, Leanne Casaubon, Morris Freedman, Richard H Swartz, Sean Symons, ONDRI Investigators, Malcolm A Binns, and Stephen C Strother

Acknowledgements

ONTARIO BRAIN INSTITUTE INSTITUT ONTARIEN DU CERVEAU

Some of this research was conducted with the support of the Ontario Brain Institute, an independent non-profit corporation, funded partially by the Ontario government. The opinions, results, and conclusions are those of the authors and no endorsement by the Ontario Brain Institute is intended or should be inferred. DB and SCS are partly supported by a Canadian Institutes of Health Research grant (MOP 201403).

Acknowledgements

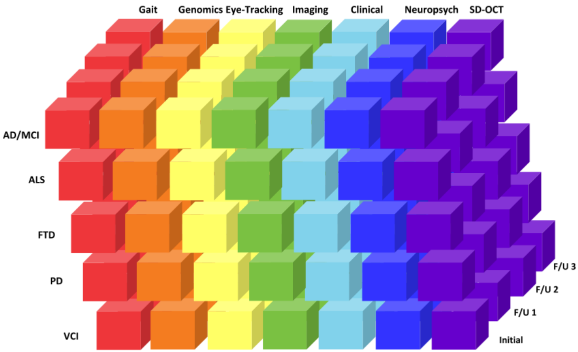
ONTARIO
NEURODEGENERATIVE
DISEASE RESEARCH
INITIATIVE

Ontario neurodegenerative disease research initiative (ONDRI)

1

Introduction

ONDRI



The ONDRI "cube" (Farhan et al., 2017). Ontario-wide, multi-site, longitudinal, multi-cohort, "deep-phenotyping". Today's focus: Alzheimer's (AD/MCI) and Vascular Cognitive Impairment (VCI)

ONDRI

- Most platforms have many modalities.

ONDRI

- Most platforms have many modalities.
  - Today's focus: structural and functional neuroimaging

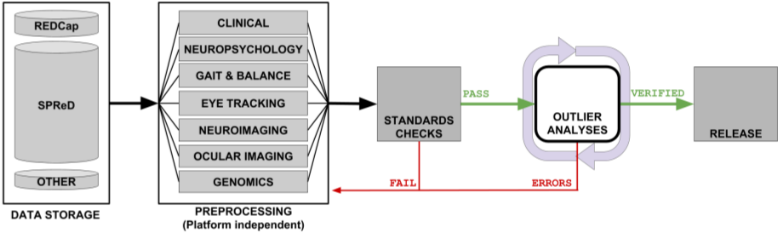
ONDRI

- Most platforms have many modalities.
  - Today's focus: structural and functional neuroimaging
- Almost everything multivariate with varying complexities

ONDRI

- Most platforms have many modalities.
  - Today's focus: structural and functional neuroimaging
- Almost everything multivariate with varying complexities
- How to ensure data are of highest quality?

Outlier detection



The ONDRI preprocessing to release pipeline

Outlier detection

- ONDRI's Neuroinformatics & Biostatistics team performs multivariate outlier detection on all data sets (see Sunderland et al., in prep).

Outlier detection

- ONDRI's Neuroinformatics & Biostatistics team performs multivariate outlier detection on all data sets (see Sunderland et al., in prep).
- Identify anomalous data points; can be

Outlier detection

- ONDRI's Neuroinformatics & Biostatistics team performs multivariate outlier detection on all data sets (see Sunderland et al., in prep).
- Identify anomalous data points; can be
  - observations that deviate from sample

Outlier detection

- ONDRI's Neuroinformatics & Biostatistics team performs multivariate outlier detection on all data sets (see Sunderland et al., in prep).
- Identify anomalous data points; can be
  - observations that deviate from sample
  - interesting patterns (e.g., co- or multi-morbid)

Outlier detection

- ONDRI's Neuroinformatics & Biostatistics team performs multivariate outlier detection on all data sets (see Sunderland et al., in prep).
- Identify anomalous data points; can be
  - observations that deviate from sample
  - interesting patterns (e.g., co- or multi-morbid)
  - errors

Outlier detection

- ONDRI's Neuroinformatics & Biostatistics team performs multivariate outlier detection on all data sets (see Sunderland et al., in prep).
- Identify anomalous data points; can be
    - observations that deviate from sample
    - interesting patterns (e.g., co- or multi-morbid)
    - errors
- Why a new approach?

Outlier detection

- ONDRI's Neuroinformatics & Biostatistics team performs multivariate outlier detection on all data sets (see Sunderland et al., in prep).
- Identify anomalous data points; can be
  - observations that deviate from sample
  - interesting patterns (e.g., co- or multi-morbid)
  - errors
- Why a new approach?
  - There are substantial limitations of existing methods

2

PCA+SHR background

New framework

- New framework:

New framework

- New framework:
    - Principal components analysis (PCA) plus

New framework

- New framework:
  - Principal components analysis (PCA) plus
  - Split-half resampling (SHR)

New framework

- New framework:
  - Principal components analysis (PCA) plus
  - Split-half resampling (SHR)
- The BIG goal

New framework

- New framework:
    - Principal components analysis (PCA) plus
    - Split-half resampling (SHR)
- The BIG goal
    - Provide flexible & robust multivariate outlier detection

Distances

Distances

- Mahalanobis distances (MD)

Distances

- Mahalanobis distances (MD)
  - Distance (standard deviation) of observation from multivariate mean

Distances

- Mahalanobis distances (MD)
    - Distance (standard deviation) of observation from multivariate mean
- Score distances (SD)

Distances

- Mahalanobis distances (MD)
  - Distance (standard deviation) of observation from multivariate mean
- Score distances (SD)
  - MD scaled by explained variance per component

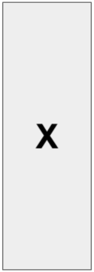
Distances

- Mahalanobis distances (MD)
  - Distance (standard deviation) of observation from multivariate mean
- Score distances (SD)
  - MD scaled by explained variance per component
- Orthogonal distances

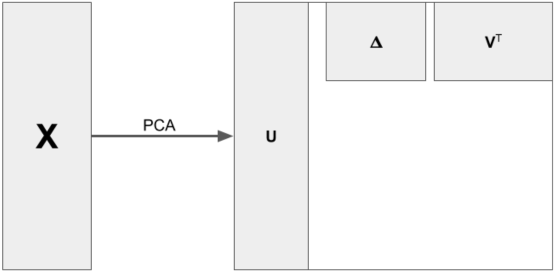
Distances

- Mahalanobis distances (MD)
  - Distance (standard deviation) of observation from multivariate mean
- Score distances (SD)
  - MD scaled by explained variance per component
- Orthogonal distances
  - We'll bring these up later

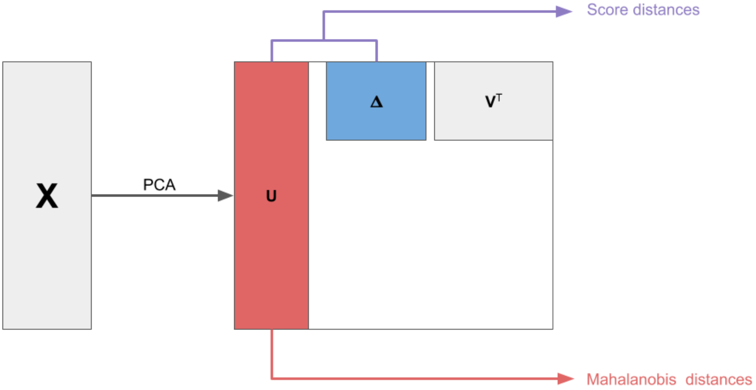
PCA



PCA



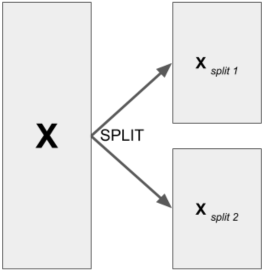
PCA



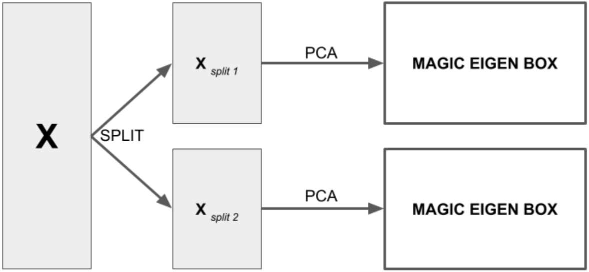
PCA


X
PCA
MAGIC EIGEN BOX
Score distances
Mahalanobis distances

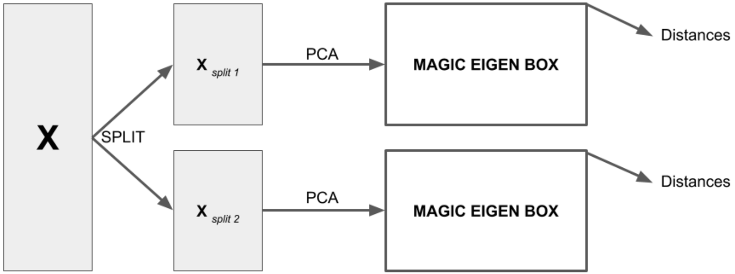
Single split PCA



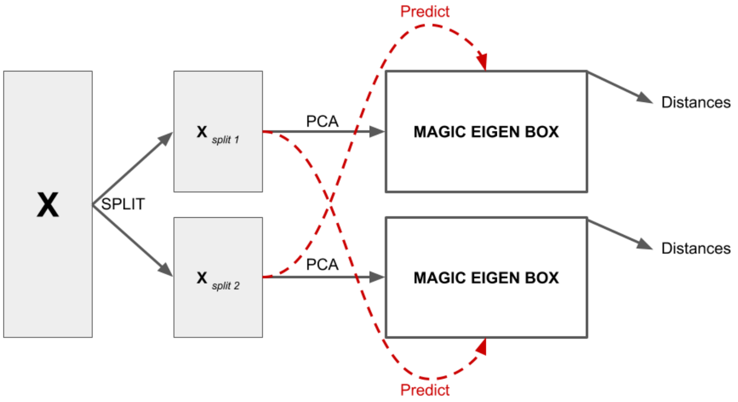
Single split PCA



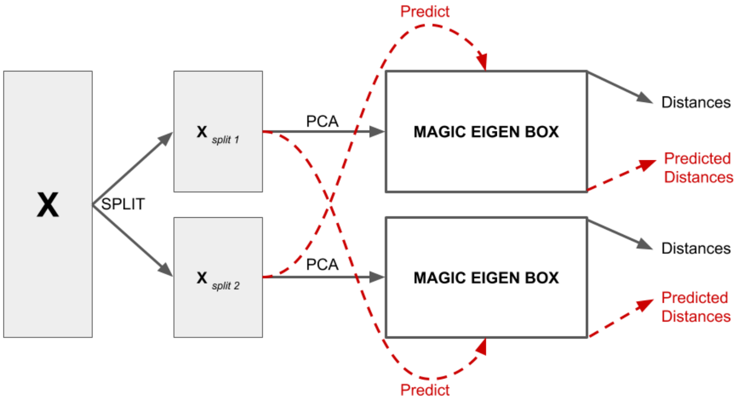
Single split PCA



Prediction from split PCAs



Prediction from split PCAs



Predicted distances

- When data are singular, collinear, or rank deficient (e.g., $I < J$):

Predicted distances

- When data are singular, collinear, or rank deficient (e.g., $I < J$):
  - Standard MD cannot be computed

Predicted distances

- When data are singular, collinear, or rank deficient (e.g., $I < J$):
  - Standard MD cannot be computed
  - Some techniques (e.g., MCD [4], ROBPCA [5]) no longer work

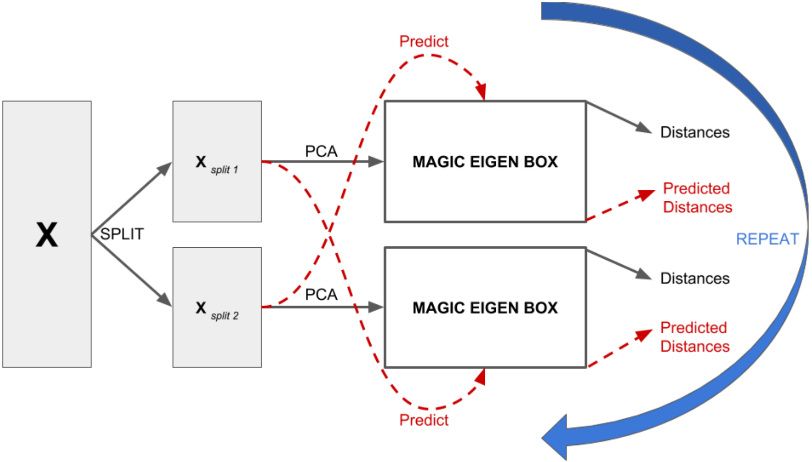
Predicted distances

- When data are singular, collinear, or rank deficient (e.g., $I < J$):
  - Standard MD cannot be computed
  - Some techniques (e.g., MCD [4], ROBPCA [5]) no longer work
  - **Predicted** distances can always be computed

Predicted distances

- When data are singular, collinear, or rank deficient (e.g., $I < J$):
    - Standard MD cannot be computed
    - Some techniques (e.g., MCD [4], ROBPCA [5]) no longer work
    - **Predicted** distances can always be computed
- But: just one pass at split PCA is not enough

Resampling


X
SPLIT
X split 1
X split 2
PCA
PCA
Predict
Predict
MAGIC EIGEN BOX
MAGIC EIGEN BOX
Distances
Predicted Distances
Distances
Predicted Distances
REPEAT

PCA+SHR

PCA+SHR

- Predicted distances can always be meaningfully computed.

PCA+SHR

- Predicted distances can always be meaningfully computed.
- Resampling provides distributions to estimate stability/spread and cutoffs.

PCA+SHR

- Predicted distances can always be meaningfully computed.
- Resampling provides distributions to estimate stability/spread and cutoffs.
- Estimates of reproducibility

PCA+SHR

- Predicted distances can always be meaningfully computed.
- Resampling provides distributions to estimate stability/spread and cutoffs.
- Estimates of reproducibility
    - Will come up later with orthogonal distances (OD).

3

PCA+SHR walkthrough

Data

- $I = 161$ observations: patients with vascular cognitive impairment

Data

- $I = 161$ observations: patients with vascular cognitive impairment
- $J = 10$ variables: volumetric brain measures from the SABRE-LE software (Ramirez et al., 2011)

Data

- $I = 161$ observations: patients with vascular cognitive impairment
- $J = 10$ variables: volumetric brain measures from the SABRE-LE software (Ramirez et al., 2011)
- Data are collinear

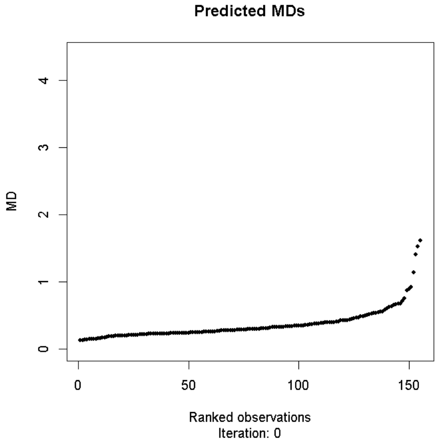
Data

- $I = 161$ observations: patients with vascular cognitive impairment
- $J = 10$ variables: volumetric brain measures from the SABRE-LE software (Ramirez et al., 2011)
- Data are collinear
    - Requires adjustment because of brain size

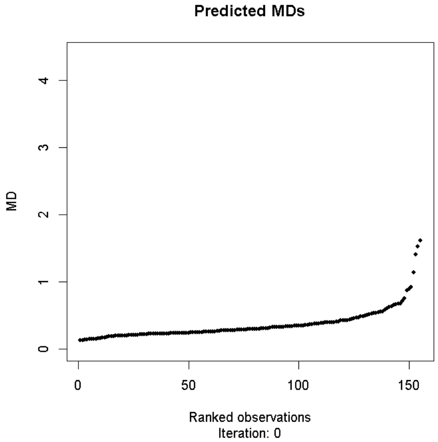
Data

- $I = 161$ observations: patients with vascular cognitive impairment
- $J = 10$ variables: volumetric brain measures from the SABRE-LE software (Ramirez et al., 2011)
- Data are collinear
  - Requires adjustment because of brain size
- $N = 1000$ iterations

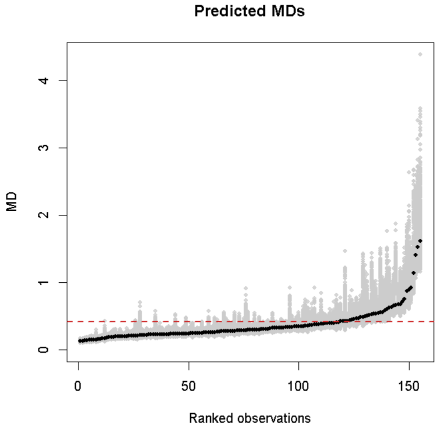
Predicted MD distributions



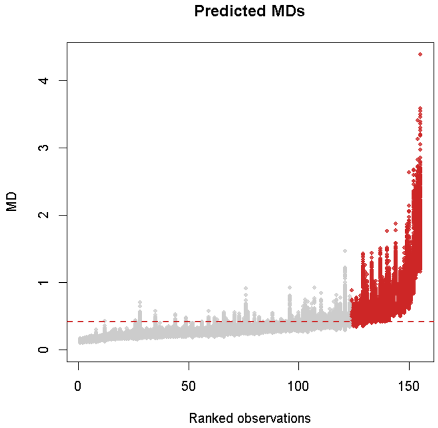
Predicted MD distributions



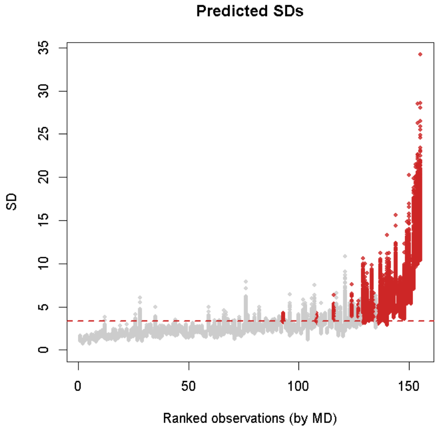
MD distribution threshold



Mahalanobis distribution outliers



Score distribution outliers



Reproducibility

- What else do we get?

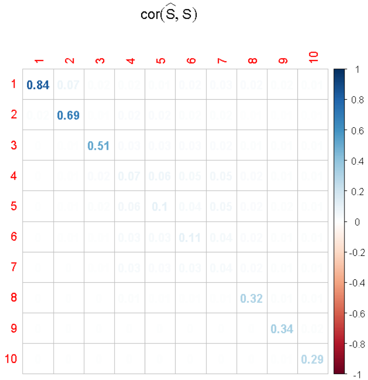
Reproducibility

- What else do we get?
  - Squared correlations between SDs and predicted SDs

Reproducibility

- What else do we get?
  - Squared correlations between SDs and predicted SDs
  - Find reproducible & robust subspace (components)

Reproducibility



Median R^2 between split scores and predicted scores over 1000 iterations

Resampling

- We can use those to rebuild a robust version of $\mathbf{X}$:

Resampling

- We can use those to rebuild a robust version of $\mathbf{X}$:
  - $\mathbf{X}$ - original data

Resampling

- We can use those to rebuild a robust version of $\mathbf{X}$:
  - $\mathbf{X}$ - original data
  - $\mathbf{X}'$ - reconstructed from just first three components

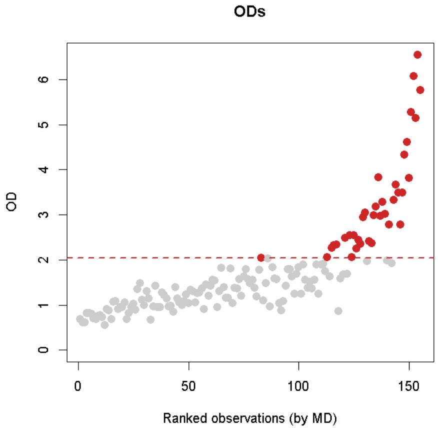
Resampling

- We can use those to rebuild a robust version of $\mathbf{X}$:
  - $\mathbf{X}$ - original data
  - $\mathbf{X}'$ - reconstructed from just first three components
- How much do observations change between $\mathbf{X}$ and robust $\mathbf{X}'$?

Orthogonal distance (OD)

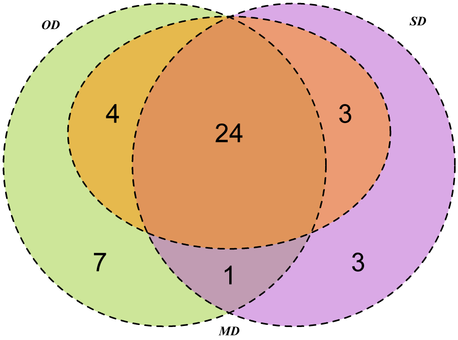
- Distance between observation and itself for $\mathbf{X}$ and robust $\mathbf{X}'$

Orthogonal distance outliers



Outliers

Outliers



$I = 161$, Outliers $= 42$.

Summary of PCA+SHR

- Build (predictive) distributions

Summary of PCA+SHR

- Build (predictive) distributions
- Find a reproducible subspace

Summary of PCA+SHR

- Build (predictive) distributions
- Find a reproducible subspace
- Three types of distances & outliers

What about large data?

What about large data?

- High dimensional, low sample size

What about large data?

- High dimensional, low sample size
    - Almost certainly rank deficient

What about large data?

- High dimensional, low sample size
  - Almost certainly rank deficient
  - Likely collinear

What about large data?

- High dimensional, low sample size
  - Almost certainly rank deficient
  - Likely collinear
  - Cannot compute MD

What about large data?

- High dimensional, low sample size
  - Almost certainly rank deficient
  - Likely collinear
  - Cannot compute MD
  - Difficult to find robust subspaces

What about large data?

- High dimensional, low sample size
    - Almost certainly rank deficient
    - Likely collinear
    - Cannot compute MD
    - Difficult to find robust subspaces
- PCA+SHR works the same regardless of size

Data

- $I = 109$ observations from the AD/MCI cohort

Data

- $I = 109$ observations from the AD/MCI cohort
- $J = 32,768$ voxels per person

Data

- $I = 109$ observations from the AD/MCI cohort
- $J = 32,768$ voxels per person
    - Resting state fMRI

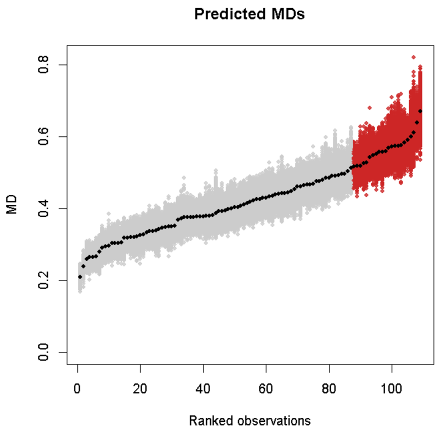
Data

- $I = 109$ observations from the AD/MCI cohort
- $J = 32,768$ voxels per person
    - Resting state fMRI
    - "Seed" voxel used to estimate whole-brain connectivity

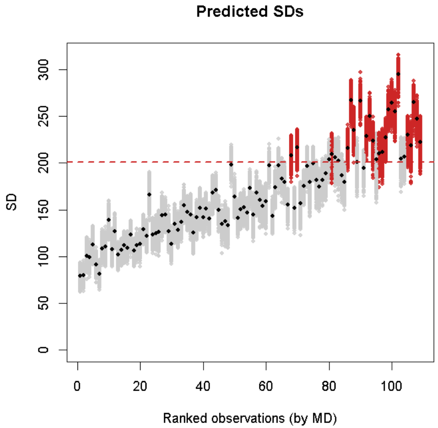
Data

- $I = 109$ observations from the AD/MCI cohort
- $J = 32,768$ voxels per person
  - Resting state fMRI
  - "Seed" voxel used to estimate whole-brain connectivity
  - OPPNI pipeline (Churchill et al., 2015)

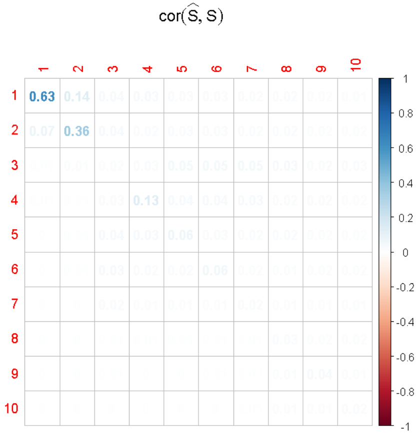
Outliers


Predicted MDs
MD
0.0
0.2
0.4
0.6
0.8
0
20
40
60
80
100
Ranked observations

Outliers

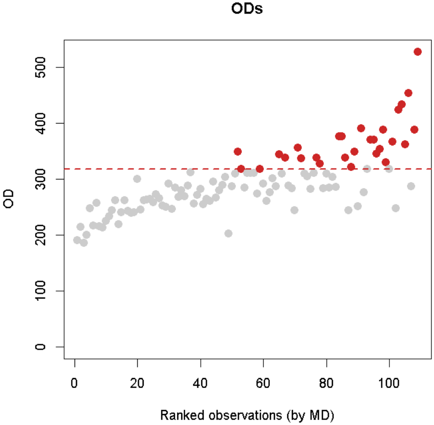

Predicted SDs
SD
0
50
100
150
200
250
300
0
20
40
60
80
100
Ranked observations (by MD)

Subspace



Only the first 10 of 108 Components shown

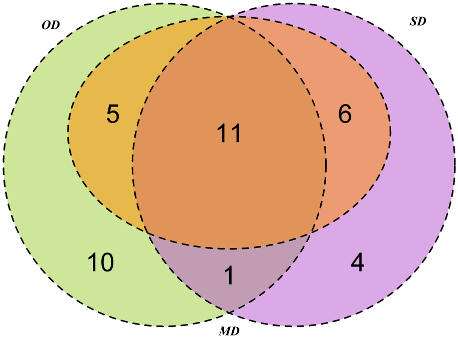
Outliers



OD computed from two components

Outliers

Outliers



$I = 109$, Outliers $= 37$.

4 Discussion

Conclusions

Conclusions

- Different distances provide different perspectives

Conclusions

- Different distances provide different perspectives
- Thresholds play a role in (non) overlap

Conclusions

- Different distances provide different perspectives
- Thresholds play a role in (non) overlap
  - High: Tends to intersect, find most outlying individuals

Conclusions

- Different distances provide different perspectives
- Thresholds play a role in (non) overlap
  - High: Tends to intersect, find most outlying individuals
  - Low: Find most outlying & unique outliers to each type

Benefits

Benefits

- PCA+SHR is flexible, overcomes limits of other methods

Benefits

- PCA+SHR is flexible, overcomes limits of other methods
    - Most methods don't work for high dimensional/low sample size data

Benefits

- PCA+SHR is flexible, overcomes limits of other methods
  - Most methods don't work for high dimensional/low sample size data
- A lot of information available

Benefits

- PCA+SHR is flexible, overcomes limits of other methods
  - Most methods don't work for high dimensional/low sample size data
- A lot of information available
  - Multiple distances, distributions to help make decisions

Limitations

- Lots of options

Limitations

- Lots of options
  - How to decide outlier identifcation?

Limitations

- Lots of options
    - How to decide outlier identifcation?
    - Distributions, point, spread?

Limitations

- Lots of options
  - How to decide outlier identifcation?
  - Distributions, point, spread?
  - But we provide defaults in the software that work well

Limitations

- Lots of options
    - How to decide outlier identifcation?
    - Distributions, point, spread?
    - But we provide defaults in the software that work well
- Can be slow

Current & future work

- Speed up (e.g., parallelization)

Current & future work

- Speed up (e.g., parallelization)
- Rolling this out within ONDRI

Current & future work

- Speed up (e.g., parallelization)
- Rolling this out within ONDRI
  - Examples online right now: http://github.com/derekbeaton/ours

Current & future work

- Speed up (e.g., parallelization)
- Rolling this out within ONDRI
  - Examples online right now: http://github.com/derekbeaton/ours
- PCA+SHR easily extends to virutally any data type

Current & future work

- Speed up (e.g., parallelization)
- Rolling this out within ONDRI
  - Examples online right now: http://github.com/derekbeaton/ours
- PCA+SHR easily extends to virutally any data type
  - Continuous, categorical, ordinal, or mixed

Current & future work

- Speed up (e.g., parallelization)
- Rolling this out within ONDRI
  - Examples online right now: http://github.com/derekbeaton/ours
- PCA+SHR easily extends to virutally any data type
  - Continuous, categorical, ordinal, or mixed
  - Beaton et al., (2018) - extended MCD [1] to any data type

References

[1] Beaton, D., Sunderland, K. M., Levine, B., Mandzia, J., Masellis, M., Swartz, R. H., ... & Strother, S. C. (2018). Generalization of the minimum covariance determinant algorithm for categorical and mixed data types. bioRxiv, 333005.

[2] Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., ... & Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. NeuroImage, 15(4), 747-771.

[3] Sunderland, K. M., Beaton D., Fraser J., Kwan, D., McLaughlin, P. M., Montero-Odasso, M., ... Binns, M. A. (in prep.). Using Multivariate Outlier Detection for Data Quality Evaluation in Large Studies: An application within the ONDRI project.

[4] Hubert, M., & Debruyne, M. (2010). Minimum covariance determinant. Wiley interdisciplinary reviews: Computational statistics, 2(1), 36-43.

[5] Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal component analysis. Technometrics, 47(1), 64-79.

[6] Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis?. Journal of the ACM (JACM), 58(3), 11.

[7] Churchill, N. W., Spring, R., Afshin-Pour, B., Dong, F., & Strother, S. C. (2015). An automated, adaptive framework for optimizing preprocessing pipelines in task-based functional MRI. PloS one, 10(7), e0131520.

[8] Farhan, S. M., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., ... & Kleinstiver, P. W. (2017). The ontario neurodegenerative disease research initiative (ONDRI). Canadian Journal of Neurological Sciences, 44(2), 196-202.

[9] Ramirez, J., Gibson, E., Quddus, A., Lobaugh, N. J., Feinstein, A., Levine, B., ... & Black, S. E. (2011). Lesion Explorer: a comprehensive segmentation and parcellation package to obtain regional volumetrics for subcortical hyperintensities and intracranial tissue. Neuroimage, 54(2), 963-973.

Questions, comments, complaints?

- Twitter: @derek__beaton (that's two underscores!)

Questions, comments, complaints?

- Twitter: @derek__beaton (that's two underscores!)
- Github for today: http://github.com/derekbeaton/ours

Questions, comments, complaints?

- Twitter: @derek__beaton (that's two underscores!)
- Github for today: http://github.com/derekbeaton/ours
- ONDRI Neuroinformatics & Biostatistics team will be in Montreal next week at **INCF: Neuroinformatics 2018**

Questions, comments, complaints?

- Twitter: @derek__beaton (that's two underscores!)
- Github for today: http://github.com/derekbeaton/ours
- ONDRI Neuroinformatics & Biostatistics team will be in Montreal next week at **INCF: Neuroinformatics 2018**
  - More detailed explanation of outlier process (Sunderland et al., in prep)

Questions, comments, complaints?

- Twitter: @derek__beaton (that's two underscores!)
- Github for today: http://github.com/derekbeaton/ours
- ONDRI Neuroinformatics & Biostatistics team will be in Montreal next week at **INCF: Neuroinformatics 2018**
  - More detailed explanation of outlier process (Sunderland et al., in prep)
  - Higher level overview of ONDRI curation-through-release pipeline

5 Bonus material

PCA

PCA

The SVD of a matrix $\mathbf{X}$ of size $I \times J$ (at least column-wise centered, i.e., covariance)

PCA

The SVD of a matrix $\mathbf{X}$ of size $I \times J$ (at least column-wise centered, i.e., covariance)

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathbf{T}} \tag{1}$$

PCA

The SVD of a matrix $\mathbf{X}$ of size $I \times J$ (at least column-wise centered, i.e., covariance)

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathbf{T}} \tag{1}$$

with the following properties

- Rank is L where $L \leq \min(I, J)$

PCA

The SVD of a matrix $\mathbf{X}$ of size $I \times J$ (at least column-wise centered, i.e., covariance)

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathbf{T}} \tag{1}$$

with the following properties

- Rank is L where $L \leq \min(I, J)$
- $\boldsymbol{\Delta}$ is $L \times L$ diagonal matrix of singular values

PCA

The SVD of a matrix $\mathbf{X}$ of size $I \times J$ (at least column-wise centered, i.e., covariance)

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathbf{T}} \tag{1}$$

with the following properties

- Rank is L where $L \leq \min(I, J)$
- $\boldsymbol{\Delta}$ is $L \times L$ diagonal matrix of singular values
- $\mathbf{U}$ is $I \times L$ (left singular vectors; rows of $\mathbf{X}$)

PCA

The SVD of a matrix $\mathbf{X}$ of size $I \times J$ (at least column-wise centered, i.e., covariance)

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathbf{T}} \tag{1}$$

with the following properties

- Rank is L where $L \leq \min(I, J)$
- $\boldsymbol{\Delta}$ is $L \times L$ diagonal matrix of singular values
- $\mathbf{U}$ is $I \times L$ (left singular vectors; rows of $\mathbf{X}$)
- $\mathbf{V}$ is $J \times L$ (right singular vectors; columns of $\mathbf{X}$)

PCA

The SVD of a matrix $\mathbf{X}$ of size $I \times J$ (at least column-wise centered, i.e., covariance)

$$\mathbf{X} = \mathbf{U \Delta V^T} \tag{1}$$

with the following properties

- Rank is L where $L \leq \min(I, J)$
- $\mathbf{\Delta}$ is $L \times L$ diagonal matrix of singular values
- $\mathbf{U}$ is $I \times L$ (left singular vectors; rows of $\mathbf{X}$)
- $\mathbf{V}$ is $J \times L$ (right singular vectors; columns of $\mathbf{X}$)
- $\mathbf{U}$ and $\mathbf{V}$ are orthonormal: $\mathbf{U^T U} = \mathbf{I} = \mathbf{V^T V}$

Score distances

Given $\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathsf{T}}$

Score distances

Given $\mathbf{X} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^{\mathbf{T}}$

- Sum of squared component (factor) scores

$$\mathbf{S} = \mathrm{diag}\{(\mathbf{U}\mathbf{\Delta})(\mathbf{U}\mathbf{\Delta})^{T}\} = \mathrm{diag}\{(\mathbf{X}\mathbf{V})(\mathbf{X}\mathbf{V})^{T}\} \quad (2)$$

Score distances

Given $\mathbf{X} = \mathbf{U\Delta V^T}$

- Sum of squared component (factor) scores
- Can be computed regardless of rank

$$\mathbf{S} = \mathrm{diag}\{(\mathbf{U\Delta})(\mathbf{U\Delta})^T\} = \mathrm{diag}\{(\mathbf{XV})(\mathbf{XV})^T\} \tag{2}$$

Score distances

Given $\mathbf{X} = \mathbf{U\Delta V^T}$

- Sum of squared component (factor) scores
- Can be computed regardless of rank
- Score distances (SD) are defined as:

$$\mathbf{S} = \mathrm{diag}\{(\mathbf{U\Delta})(\mathbf{U\Delta})^T\} = \mathrm{diag}\{(\mathbf{XV})(\mathbf{XV})^T\} \tag{2}$$

Mahalanobis

Given $\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathbf{T}}$

Mahalanobis

Given $\mathbf{X} = \mathbf{U\Delta V^T}$

- Sum of squared singular vectors

$$\begin{aligned}\mathbf{M} &= \mathrm{diag}\{\mathbf{UU}^T\} = \mathrm{diag}\{(\mathbf{XV\Delta}^{-1})(\mathbf{XV\Delta}^{-1})^T\} \\ &= \mathrm{diag}\{(\mathbf{S\Delta}^{-1})(\mathbf{S\Delta}^{-1})^T\}\end{aligned} \tag{3}$$

Mahalanobis

Given $\mathbf{X} = \mathbf{U\Delta V^T}$

- Sum of squared singular vectors
- Can only be computed when not rank deficient

$$\begin{aligned}\mathbf{M} = \mathrm{diag}\{\mathbf{UU}^{T}\} &= \mathrm{diag}\{(\mathbf{XV\Delta}^{-1})(\mathbf{XV\Delta}^{-1})^{T}\} \\ &= \mathrm{diag}\{(\mathbf{S\Delta}^{-1})(\mathbf{S\Delta}^{-1})^{T}\}\end{aligned} \tag{3}$$

Mahalanobis

Given $\mathbf{X} = \mathbf{U\Delta V^T}$

- Sum of squared singular vectors
- Can only be computed when not rank deficient
  - if $I >> L$; when $I \leq L$, $\mathbf{M} = \mathbf{I}c$ where $c = L \times I^{-1}$

$$\begin{aligned} \mathbf{M} = \mathrm{diag}\{\mathbf{UU}^T\} &= \mathrm{diag}\{(\mathbf{XV\Delta}^{-1})(\mathbf{XV\Delta}^{-1})^T\} \\ &= \mathrm{diag}\{(\mathbf{S\Delta}^{-1})(\mathbf{S\Delta}^{-1})^T\} \end{aligned} \tag{3}$$

Mahalanobis

Given $\mathbf{X} = \mathbf{U\Delta V^T}$

- Sum of squared singular vectors
- Can only be computed when not rank deficient
  - if $I >> L$; when $I \leq L$, $\mathbf{M} = \mathbf{I}c$ where $c = L \times I^{-1}$
- Mahalanobis distances (MD) are defined as:

$$\begin{aligned} \mathbf{M} = \mathrm{diag}\{\mathbf{UU}^T\} &= \mathrm{diag}\{(\mathbf{XV\Delta}^{-1})(\mathbf{XV\Delta}^{-1})^T\} \\ &= \mathrm{diag}\{(\mathbf{S\Delta}^{-1})(\mathbf{S\Delta}^{-1})^T\} \end{aligned} \tag{3}$$

Single split PCA

For some subset H and its complement $\bar{H}$ we have two PCAs:

$$\begin{aligned} \mathbf{X}_H &= \mathbf{U}_H \boldsymbol{\Delta}_H \mathbf{V}_H^T \\ \mathbf{X}_{\bar{H}} &= \mathbf{U}_{\bar{H}} \boldsymbol{\Delta}_{\bar{H}} \mathbf{V}_{\bar{H}}^T \end{aligned} \tag{4}$$

Single split PCA

For some subset H and its complement $\bar{H}$ we have two PCAs:

$$\begin{aligned} \mathbf{X}_H &= \mathbf{U}_H \boldsymbol{\Delta}_H \mathbf{V}_H^T \\ \mathbf{X}_{\bar{H}} &= \mathbf{U}_{\bar{H}} \boldsymbol{\Delta}_{\bar{H}} \mathbf{V}_{\bar{H}}^T \end{aligned} \tag{4}$$

Size of H could be $\alpha = .5$ (split half) or e.g., $\alpha = .9$ (90-10)

Predicted distances

Predicted distances

Predicted SD:

$$
\begin{aligned}
\widehat{\mathbf{S}}_H &= \operatorname{diag}\{(\mathbf{X}_H\mathbf{V}_{\bar{H}})(\mathbf{X}_H\mathbf{V}_{\bar{H}})^T\} \\
\widehat{\mathbf{S}}_{\bar{H}} &= \operatorname{diag}\{(\mathbf{X}_{\bar{H}}\mathbf{V}_H)(\mathbf{X}_{\bar{H}}\mathbf{V}_H)^T\}
\end{aligned}
\tag{5}
$$

Predicted distances

Predicted SD:

$$\begin{aligned}\widehat{\mathbf{S}}_H &= \mathrm{diag}\{(\mathbf{X}_H\mathbf{V}_{\bar{H}})(\mathbf{X}_H\mathbf{V}_{\bar{H}})^T\}\\ \widehat{\mathbf{S}}_{\bar{H}} &= \mathrm{diag}\{(\mathbf{X}_{\bar{H}}\mathbf{V}_H)(\mathbf{X}_{\bar{H}}\mathbf{V}_H)^T\}\end{aligned} \tag{5}$$

Predicted MD:

$$\begin{aligned}\widehat{\mathbf{M}}_H &= \mathrm{diag}\{(\widehat{\mathbf{S}}_H\mathbf{\Delta}_{\bar{H}})(\widehat{\mathbf{S}}_H\mathbf{\Delta}_{\bar{H}})^T\}\\ \widehat{\mathbf{M}}_{\bar{H}} &= \mathrm{diag}\{(\widehat{\mathbf{S}}_{\bar{H}}\mathbf{\Delta}_H)(\widehat{\mathbf{S}}_{\bar{H}}\mathbf{\Delta}_H)^T\}\end{aligned} \tag{6}$$

Predicted distances

Predicted SD:

$$\begin{aligned}\widehat{\mathbf{S}}_H &= \mathrm{diag}\{(\mathbf{X}_H\mathbf{V}_{\bar{H}})(\mathbf{X}_H\mathbf{V}_{\bar{H}})^T\} \\ \widehat{\mathbf{S}}_{\bar{H}} &= \mathrm{diag}\{(\mathbf{X}_{\bar{H}}\mathbf{V}_H)(\mathbf{X}_{\bar{H}}\mathbf{V}_H)^T\}\end{aligned} \tag{5}$$

Predicted MD:

$$\begin{aligned}\widehat{\mathbf{M}}_H &= \mathrm{diag}\{(\widehat{\mathbf{S}}_H\boldsymbol{\Delta}_{\bar{H}})(\widehat{\mathbf{S}}_H\boldsymbol{\Delta}_{\bar{H}})^T\} \\ \widehat{\mathbf{M}}_{\bar{H}} &= \mathrm{diag}\{(\widehat{\mathbf{S}}_{\bar{H}}\boldsymbol{\Delta}_H)(\widehat{\mathbf{S}}_{\bar{H}}\boldsymbol{\Delta}_H)^T\}\end{aligned} \tag{6}$$

Orthogonal distance

- We can use subspace to rebuild a robust version of $\mathbf{X}$:

Orthogonal distance

- We can use subspace to rebuild a robust version of $\mathbf{X}$:
  - $\mathbf{X}' = \mathbf{U}_{1:3}\mathbf{\Delta}_{1:3}\mathbf{V}_{1:3}^{T}$

Orthogonal distance

- We can use subspace to rebuild a robust version of $\mathbf{X}$:
  - $\mathbf{X}' = \mathbf{U}_{1:3}\mathbf{\Delta}_{1:3}\mathbf{V}_{1:3}^{T}$
- How much do observations change between $\mathbf{X}$ and robust $\mathbf{X}'$?

Orthogonal distance

Given two commensurate matrices, $\mathbf{X}$ and $\mathbf{X}'$, orthogonal distances (OD) are defined as:

$$\mathbf{O} = \| \mathbf{X} - \mathbf{X}' \| \tag{7}$$