ONDRI Neuroinformatics & Biostatistics Methods

On behalf of the Neuroinformatics and Biostatistics (NIBS) platform

This methods document is organized into four sections:

1. Methods summary
2. Methods details
3. References
4. Contributions and contacts

This document is subject to change as necessary. To cite this document:

The Ontario Neurodegenerative Disease Research Initiative's Neuroinformatics and Biostatistics team (2021). *Methods.* Retrieved from https://github.com/ondri-nibs/documentation

Version: 1.1.0.0

Methods Summary

As part of ONDRI's curation protocol, all data were verified for adherence to ONDRI standards (Neuroinformatics & Biostatistics, n.d.). Most data were assessed with multivariate outlier detection techniques (Beaton et al., 2020; Sunderland et al., 2019) in order to identify anomalies, and in some cases make corrections to erroneous data. Software, documentation, examples, and materials for the standardization and outlier detection processes are available at https://github.com/ondri-nibs.

Methods Details

As part of ONDRI's curation protocol, all data were verified for adherence to ONDRI standards (Neuroinformatics & Biostatistics, n.d.). Most data were assessed with multivariate outlier detection techniques (Beaton et al., 2020; Sunderland et al., 2019) in order to identify anomalies and, in some cases, make corrections to erroneous data. Software, documentation, examples, and materials for the standardization and outlier detection processes are available at https://github.com/ondri-nibs. We detail each of the steps and discuss specific methods in the subsections below.

Standards

ONDRI's data standards are generally tabular or non-tabular based standards built around data packages. Each data package requires one or more data files and additional information files, such as dictionaries. Individual files are checked for adherence to formatting requirements, where data packages are more generally checked to ensure adequate, consistent, and accurate information, including but not limited to: (1) correct participants, visits, sites, and dates, (2) correct missing codes, shortcodes, and naming conventions, (3) minimum required set of files with appropriate information. See the ONDRI-NIBS data curation compendium and data standards documentation for more details. All data are subject to standardization checks with software created by the Neuroinformatics and Biostatics (NIBS) team (see https://github.com/ondri-nibs for standards documentation and software tools).

Outlier detection

Following confirmation of adherence to standards, the NIBS team performs at least one of the following multivariate outlier detection techniques, and reports the results to the platform: ordination, minimum covariance determinant, and CorrMax transformation. In some cases, a set of these analyses are performed twice: on the data as they are ("as is") and again but residualized for age and sex ("residualized"). See ***Residualization with respect to covariates*** (at the end) for more information. Those results report anomalous data, which could be participants (on the whole), variables (on the whole), or specific values (a specific variable for a specific participant). The corresponding platform team then verifies the data. If errors are found, the platform corrects the erroneous anomalies and this process is repeated. If no errors are detected, the data package proceeds to release. These outlier detection techniques apply to almost all tabular data packages, but non-tabular packages (i.e., file based such as neuroimaging, sensors, rare genomic variants) are not always subject to outlier detection by the NIBS team during this process. In these cases, platforms usually perform QA/QC and/or outlier analyses on their own, and/or with the guidance of NIBS.

Software, tools, and resources

To perform standards and outliers checks and analyses, we use a set of tools we have developed. The primary place for public-facing software, data examples, and materials is here: https://github.com/ondri-nibs. This Github organization page ("ONDRI-NIBS") houses multiple repositories each with different materials, tools, software, or resources. Any use of

any software, tools, or resources should be correctly cited, and most of the software automatically generates citations (via citation() in R). Please see these respective repositories for additional information on citations including but not limited to manuscripts and additional software. Much of the standards and outliers procedure are performed with the following software:

- A set of Shiny/R Apps:
    - The *Standards ShinyApp* to perform structural and project level checks via graphical user interface (GUI), run as a web browser-based ShinyApp: https://github.com/ondri-nibs/standards_app
    - The *DataPrep ShinyApp* to perform key and common data preparation steps used by the NIBS team, generally for outlier analyses (but can be used for other analyses); also run via GUI in a web browser-based ShinyApp: https://github.com/ondri-nibs/dataprep_app
    - The *Outliers ShinyApp* to perform a battery of multivariate outlier analyses (described below) and to produce visualizations and harmonized reports for review by platforms, curators, and other data or clinical experts within the project; also run via GUI in a web browser-based ShinyApp https://github.com/ondri-nibs/outliers_app
- A stand alone standards package that performs structural standards checks (not project checks), run via R code: https://github.com/ondri-nibs/standards_package
- The *Outliers and Robust Structures* (OuRS) package, which is the core of the outlier analyses that includes the methods listed below, i.e., ordination techniques, minimum covariance determinant approaches, and CorrMax transformations: https://github.com/derekbeaton/ours
- The *Generalized Partial Least Squares* (GPLS) package, which is the core of the residualization steps: https://github.com/derekbeaton/gpls

Also see the technical documentation we provide:
https://github.com/ondri-nibs/documentation

Data preparation and analyses

Data coding

Because data can be of several types, we sometimes require different ways of representing, recoding, or transforming data so that they can be analyzed. This is especially the case for data of mixed types (e.g., a data set with continuous, categorical, and ordinal variables). Here we list just a few ways and provide references for further reading. All of these types of coding and transformations can be found in the *OuRS* package, and are explained through documentation and publications on generalized MCD (Beaton et al., 2020) and generalized PLS (Beaton et al., 2019).

Z-scoring is for continuous variables, where each variable is transformed into a unitless variable, where it is first centered by its mean, and then divided (normalized) by its standard deviation.

Complete disjunctive coding is for categorical variables, where each variable (one column) is transformed into multiple columns comprised of 0s and 1s, where each category is mutually exclusive. This is sometimes called "one-hot encoding" or "dummy coding" when all variables are represented (but the intercept is excluded).

Thermometer coding is for ordinal variables, where each variable (one column) is transformed into *two* columns that represent the lower ("-") and upper ("+") ends of an ordinal scale. Together, the two columns indicate how far above a lower value ("-") and how far below an upper value ("+") the observed value is. By default, this assumes uniform differences and is based on the numeric (typically a rank) representation of the ordinal values. However, this can be used in a variety of ways. This is also sometimes called fuzzy coding or data doubling (M. Greenacre, 2014).

Escofier coding is for when continuous variables are to be analyzed in conjunction with ordinal and/or categorical variables. Like the *thermometer coding*, each variable (one column) is transformed into two columns: a lower and an upper distance from the *Z*-score, where the lower column is [(1 - *Z*)/2] and the upper column is [(1 + *Z*)/2]. This is also sometimes called fuzzy coding or data doubling (M. Greenacre, 2014), but we have elected to refer to it as *Escofier coding* (Beaton et al., 2016, 2019) based on the original works of Escofier (Escofier, 1979).

Ordination

Ordination analysis is principal component analysis (PCA) or correspondence analysis (CA). *Principal component analysis* (Abdi & Williams, 2010) is used when all data in a data set (for analysis) is numeric and can, generally, be assumed continuous. All columns are *Z*-scored so they have the same (unitless) scale (a "correlation PCA"). *Correspondence analysis* (M. J. Greenacre, 2010) is used when at least one variable is categorical or ordinal. CA can be used for data comprised entirely of categorical or ordinal data, or can be applied to data with mixtures of categorical, ordinal, and continuous variables.

Both techniques work in the same way and provide analogous results. However, each technique is applied to different types of data (mentioned above). Each technique creates components. Components are linear combinations of variables. In both PCA and CA the first component explains the maximum variance in the data. Subsequent components explain the maximum possible variance conditional to orthogonality to the previous components. The last two components explain the second least (second to last component) and least (last component) amount of variance. For ordination analysis, we provide graphical results (scatter plot-like figures) that show the first two and last two components. This is a visual inspection of the data to quickly identify any potentially problematic observations. The results show the configurations of participants and variables separately. Any particularly distant participants or variables on the first two components are considered "high variance" outliers: they tend to be unlike other data and could be overrepresented in the data. When these types of outliers are errors, they could be due to sign flips (negative values are positive and vice versa), specific very large values, or scaling/magnitude problems. Any particularly distant participants or

variables on the last two components are considered candidates for “masked” outliers (Saporta & Keita, 2009); we also use the minimum covariance determinant (MCD) to identify masked outliers.

Identify outlying participants

Outlier analysis is when we use multivariate techniques to identify outlying *participants* or *observations* (rows of a data matrix). We use either the minimum covariance determinant (MCD) or the generalized minimum covariance determinant (GMCD). *Minimum covariance determinant* (Hubert et al., 2018; Sunderland et al., 2019) is used when all data in a data set (for analysis) is numeric and can, generally, be assumed continuous. All columns are Z-scored so they have the same (unitless) scale. *Generalized minimum covariance determinant* (Beaton et al., 2020) is used when at least one variable is categorical or ordinal. GMCD can be used for data comprised entirely of categorical or ordinal data, or can be applied to data with mixtures of categorical, ordinal, and continuous variables.

Both MCD approaches work by identifying a robust covariance matrix—by way of subsampling—where said robust covariance matrix has a minimum determinant. Once a robust covariance matrix has been identified, it is then used to compute robust Mahalanobis distances. These robust Mahalanobis distances help uncover “masked” outliers: those that are difficult to identify within multivariate data. We present these results with standard Mahalanobis distances vs. robust Mahalanobis distances. Individuals with high values on both distances, and high values on robust Mahalanobis distances alone are participants/observations of concern.

Identify specific outlying values

Specific values are identified as possible sources of anomalies; that is, specific variables for specific participants are identified. We use either the CorrMax transformation or the Generalized CorrMax transformation. *CorrMax transformation* (Garthwaite & Koch, 2016) is used when all data in a data set (for analysis) is numeric and can, generally, be assumed continuous. All columns are Z-scored so they have the same (unitless) scale. *Generalized CorrMax transformation* is used when at least one variable is categorical or ordinal. Generalized CorrMax can be used for data comprised entirely of categorical or ordinal data, or can be applied to data with mixtures of categorical, ordinal, and continuous variables. ***NOTE:*** This method is currently not published, but the code is available in the *OuRS* package (https://github.com/derekbeaton/ours). The Generalized CorrMax works based on the same principles of covariance matrices we established for the GMCD method (Beaton et al., 2020).

Both CorrMax approaches work by (1) using the robust covariance structure from the MCD (for CorrMax) or GMCD (for Generalized CorrMax), and (2) applying a transformation to identify specific variables with high contributions to the robust covariance structure (and thus the robust Mahalanobis distances). Values are presented row-wise (for each participant or observation) so that the sum of each row is 1 (or 100%). The value for each variable across the row is the amount of total contribution to that row.

Residualization with respect to covariates

Partial least squares regression (Abdi, 2010; Tenenhaus, 1998) is used when all data in a data set (for analysis) is numeric and can, generally, be assumed continuous. All columns in the response variables are Z-scored so they have the same (unitless) scale. For the explanatory variables: age is z-scored and sex is recoded as a one-column binary variable (either MALE or FEMALE are coded as '1', and the other as '0').

Partial least squares-correspondence analysis-regression (Beaton et al., 2019) is used when at least one variable *in the response data set* is categorical or ordinal. PLSCAR can be used for data comprised entirely of categorical or ordinal data, or can be applied to (response variable) data with mixtures of categorical, ordinal, and continuous variables. Age is recoded/transformed with Escofier transform, and sex is coded as a two-column complete disjunctive data set (a.k.a. one-hot encoding, sometimes also a.k.a. dummy coding with all levels and no intercept).

In many cases, we perform analyses twice: (1) "as is" which is on the data as they are, and (2) "residualized" which is on data that are the residuals from a regression, this is also known as "correcting for" covariates or "orthogonalization" because the residuals are now orthogonal to the covariates. We use either partial least squares regression (PLSR) or partial least squares-correspondence analysis-regression (PLSCAR) for the residualization step. Both PLS approaches use a small set of covariates or explanatory variables, typically just age and sex, and regress these covariates out of the data set of interest. We only use age and sex because these are (1) the most common covariates in neurodegenerative and aging research, and (2) generally do explain some large effects or differences; sometimes to identify outliers we need to remove the effects of age and sex. We use PLS approaches because the problem is a multivariate regression: our responses are multivariate data sets. We use the residualized data from this step and submit it to all of the previous steps above (when relevant or possible). We report on the similarities and differences between the two types of analyses ("as is" and "residualized").

References

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 97–106. https://doi.org/10.1002/wics.51

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Beaton, D., Dunlop, J., Abdi, H., & Alzheimer's Disease Neuroimaging Initiative. (2016). Partial Least Squares Correspondence Analysis: A Framework to Simultaneously Analyze Behavioral and Genetic Data. *Psychological Methods*, *21*(4), 621–651. https://doi.org/10.1037/met0000053

Beaton, D., Saporta, G., Abdi, H., Initiative, A. D. N., & others. (2019). A generalization of partial least squares regression and correspondence analysis for categorical and mixed data: An application with the ADNI data. *BioRxiv*, 598888.

Beaton, D., Sunderland, K. M., ADNI, Levine, B., Mandzia, J., Masellis, M., Swartz, R. H., Troyer, A. K., ONDRI, Binns, M. A., Abdi, H., & Strother, S. C. (2020). Generalization of the minimum covariance determinant algorithm for categorical and mixed data types. *BioRxiv*, 333005. https://doi.org/10.1101/333005

Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. *Cahiers de l'Analyse Des Données*, *Tome 4*(2), 137–146.

Garthwaite, P. H., & Koch, I. (2016). Evaluating the Contributions of Individual Variables to a Quadratic Form. *Australian & New Zealand Journal of Statistics*, *58*(1), 99–119. https://doi.org/10.1111/anzs.12144

Greenacre, M. (2014). Data Doubling and Fuzzy Coding. In J. Blasius & M. Greenacre (Eds.), *Visualization and Verbalization of Data* (pp. 239–253). CRC Press.

Greenacre, M. J. (2010). Correspondence analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(5), 613–619. https://doi.org/10.1002/wics.114

Hubert, M., Debruyne, M., & Rousseeuw, P. J. (2018). Minimum covariance determinant and extensions. *WIREs Computational Statistics*, *10*(3), e1421. https://doi.org/10.1002/wics.1421

Neuroinformatics & Biostatistics. (n.d.). *The Ontario Neurodegenerative Disease Research Initiative's Data Standards*. https://github.com/ondri-nibs/documentation

Saporta, G., & Keita, N. N. (2009). *Principal component analysis: Application to statistical process control*. ISTE.

Sunderland, K. M., Beaton, D., Fraser, J., Kwan, D., McLaughlin, P. M., Montero-Odasso, M., Peltsch, A. J.,

Pieruccini-Faria, F., Sahlas, D. J., Swartz, R. H., Bartha, R., Black, S. E., Borrie, M., Corbett, D., Finger, E., Freedman, M., Greenberg, B., Grimes, D. A., Hegele, R. A., … ONDRI Investigators. (2019). The utility of multivariate outlier detection techniques for data quality evaluation in large studies: An application within the ONDRI project. *BMC Medical Research Methodology*, *19*(1), 102. https://doi.org/10.1186/s12874-019-0737-5

Tenenhaus, M. (1998). *La régression PLS: Théorie et pratique*. Editions TECHNIP.

Contributions and contacts

Creation and authorship this document is outlined below. There are four roles, three of which are defined by the CRediT system (https://casrai.org/credit/) with adaptations of the definitions as necessary, see also (https://www.pnas.org/content/115/11/2557):

- ***Conceptualization***: Ideas; formulation or evolution of overarching goals and aims.
- ***Writing - original draft***: Preparation, creation and/or presentation of the published work, specifically writing the initial draft.
- ***Writing - review & editing***: Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision
- ***Additional contributions***: Contributions to this document not otherwise captured in Conceptualization, Writing (draft), or Writing (reviews/edits).

Conceptualization, Writing (draft), and Writing (review/edit) are not exclusive from one another but are exclusive from Additional Contributions. If one appears in Contributions they do not appear in Conceptualization, Writing (draft), and Writing (review/edit) and vice versa.

Conceptualization	*Writing (initial draft)*	*Writing (review/edit)*	*Additional Contributions*
Derek Beaton	Derek Beaton	Derek Beaton Kelly Sunderland	

For information on NIBS standards and outliers pipelines, please contact the Neuroinformatics & Biostatistics team.