Generalization of the minimum covariance determinant algorithm for categorical and mixed data types

Derek Beaton[1], Kelly M. Sunderland[1], ADNI[3], Brian Levine[1,9], Jennifer Mandzia[4], Mario Masellis[5,6], Richard H. Swartz[5,6], Angela K. Troyer[2,5], ONDRI[7], Malcolm A. Binns[1,5], Hervé Abdi[8], & Stephen C. Strother[1,5]

[1] Rotman Research Institute at Baycrest Health Sciences

[2] Baycrest Health Sciences

[3] ADNI

[4] Department of Clinical Neurological Sciences, Western University

[5] University of Toronto

[6] Department of Medicine (Neurology), Sunnybrook Health Sciences Centre

[7] ONDRI

[8] The University of Texas at Dallas

[9] Departments of Psychology and Medicine (Neurology), University of Toronto

Author Note

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.ucla.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Correspondence concerning this article should be addressed to Derek Beaton, 3560 Bathurst Street Toronto, Ontario Canada M6A 2E1. E-mail: dbeaton@research.baycrest.org

Abstract

The minimum covariance determinant (MCD) algorithm is one of the most common techniques to detect anomalous or outlying observations. The MCD algorithm depends on two features of multivariate data: the determinant of a matrix (i.e., geometric mean of the eigenvalues) and Mahalanobis distances (MD). While the MCD algorithm is commonly used, and has many extensions, the MCD is limited to analyses of quantitative data and more specifically data assumed to be continuous. One reason why the MCD does not extend to other data types such as categorical or ordinal data is because there is not a well-defined MD for data types other than continuous data. To address the lack of MCD-like techniques for categorical or mixed data we present a generalization of the MCD. To do so, we rely on a multivariate technique called correspondence analysis (CA). Through CA we can define MD via singular vectors and also compute the determinant from CA's eigenvalues. Here we define and illustrate a generalized MCD on categorical data and then show how our generalized MCD extends beyond categorical data to accommodate mixed data types (e.g., categorical, ordinal, and continuous). We illustrate this generalized MCD on data from two large scale projects: the Ontario Neurodegenerative Disease Research Initiative (ONDRI) and the Alzheimer's Disease Neuroimaging Initiative (ADNI), with genetics (categorical), clinical instruments and surveys (categorical or ordinal), and neuroimaging (continuous) data. We also make `R` code and toy data available in order to illustrate our generalized MCD.

Keywords: Minimum covariance determinant, correspondence analysis, categorical data, outliers, robust, neuroinformatics

Generalization of the minimum covariance determinant algorithm for categorical and mixed data types

Introduction

Today's large scale studies contain a wide array of data from multiple sources that span a variety of data types. For examples, certain types of genetic data (single nucleotide polymorphisms) are categorical, surveys and clinical instruments are often ordinal, and other measures, such as reaction times or neuroimaging, are continuous. Such data sets are typically open or broadly accessible, and thus analyzed by hundreds or thousands of researchers. With many analyses at stake, it is critical to ensure the integrity and accuracy of data as well as the robustness of results. In the Ontario Neurodegenerative Disease Research Initiative (ONDRI), we employ a quality control step as part of the release process to identify anomalies (Sunderland et al., 2019). Our quality control step uses a common and reliable outlier detection technique called the minimum covariance determinant (MCD; Hubert & Debruyne, 2010), which sees use across numerous disciplines (Hadi, Imon, & Werner, 2009; Leys, Klein, Dominicy, & Ley, 2018; Magnotti & Billor, 2014; Verity et al., 2017). However, the MCD can only identify outliers in continuous data. In ONDRI—and many studies like ours—we have data sets comprised entirely of categorical or ordinal data, and complex data sets with mixtures of data types. Clearly we need an outlier detection technique that can accommodate various data types and, preferably, work in much the same way the MCD technique works.

The MCD produces robust estimates for multivariate location (mean) and scatter (covariance). Given a matrix of observations (rows) and variables (columns), the MCD finds a subset of observations that have minimum scatter. In the MCD, scatter is defined by the determinant, and thus the search for a minimum determinant from a (covariance) matrix. Because the MCD algorithm searches for the smallest scatter of observations, it also helps identify multivariate outliers via robust Mahalanobis distances. Since the introduction of the

Fast-MCD approach (Rousseeuw & Van Driessen, 1999), there have been many improvements and variations on the technique, such as robust PCA (Hubert, Rousseeuw, & Branden, 2005), deterministic MCD (Hubert, Rousseeuw, & Verdonck, 2012), and extensions to address high-dimensional data through regularization (Boudt, Rousseeuw, Vanduffel, & Verdonck, 2017). See Hubert, Debruyne, and Rousseeuw (2017) for an overview on the MCD and various extensions thereof. Though the MCD and its variants are standard techniques to identify both robust covariance structures and outliers, MCD approaches work only for data assumed to be continuous. So, how can we adapt MCD approaches to non-quantitative data? The major barrier to adapt the MCD for non-quantitative data is because there are no standard definitions of Mahalanobis distances for non-quantitative data.

Goodall (1966) noted that there are generally two issues with the application of standard computations and rules of Mahalanobis distances to non-quantitative data: (1) "If [...] quantitative and binary attributes are included in the same index, these procedures will generally give the latter excessive weight." (p. 883), and (2) "indices appropriate to quantitative [data have] been applied to ordered, non-quantitative attributes [...] by arbitrarily assigning metric values to their different ordered states." (p. 883). Clearly, we do not want categorical data to provide undue influence on Mahalanobis distance estimates, nor should we arbitrarily apply metric values to categorical or ordinal data. While there are many Mahalanobis-like distances or alternative measures of similarity for non-quantitative data (Bar-hen & Daudin, 1995; Bedrick, Lapidus, & Powell, 2000; Boriah, Chandola, & Kumar, 2008; Leon & Carrière, 2005; McCane & Albert, 2008) many still have the drawbacks noted by Goodall (1966). If we want to develop a MCD algorithm for non-quantitative data, we need to define a Mahalanobis distance that neither imposes undue influence nor arbitrarily assigns values.

Here we introduce an approach that generalizes the MCD for use with a variety of data types: categorical, ordinal, continuous, and mixtures thereof. Our generalization of the MCD

is possible through the use of Correspondence Analysis (CA). By way of CA, we can obtain both the determinant of a matrix from its eigenvalues, and Mahalanobis distances from its singular vectors. Our paper is outlined as follows. We first show how the SVD provides the two key requirements for the MCD algorithm: the determinant of a covariance matrix and Mahlanobis distances. Then, we show how we can generalize these properties of the SVD in order to compute Mahalanobis distances for categorical data through CA. With Mahalanobis distances for categorical data, we can then formalize a generalized MCD for categorical data via Multiple CA. Specifically, our generalized MCD makes use of a technique called "subset multiple correspondence analysis". Next we illustrate our method on a toy (simulated) genetics dataset of categorical variables. We then show how our generalized MCD works at scale. First we use the MCD for survey data from ONDRI. Because survey data can be treated as either categorical or ordinal data, we apply our generalized MCD to the same survey data under categorical and ordinal assumptions. Coding the survey data both ways allows us to compare and contrast the behavior of our generalized MCD under each condition. We then apply our generalized MCD to a mixed data set from the Alzheimer's Disease Neuroimaging Initiative (ADNI). In the ADNI example we show how our generalized MCD extends to mixed data with categorical (genetics), ordinal (a clinical instrument), and continuous (brain volumes) data. In the ADNI example we highlight particular outliers identified from our generalized MCD, and then show how to understand why those individuals are outliers. Finally we discuss the technique with respect to the results, as well as provide future directions, limitations, and concluding remarks for our generalized MCD.

A Generalization of the MCD for non-quantitative data

Software and notation

We used the `papaja` package (Aust & Barth, 2018) to write this manuscript via RMarkdown. We make our software, code, and some examples of the GMCD available via the *Ou*tliers and *R*obust *S*tructures (OuRS) package at

https://github.com/derekbeaton/ours.

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 struture. Vectors are assumed to be column vectors unless otherwise specified. Two matrices side-by-side denotes standard matrix multiplication (e.g., $\mathbf{XY}$), where $\odot$ denotes element-wise multiplication. The matrix $\mathbf{I}$ denotes the identity matrix. Superscript ^{T} denotes the transpose operation, superscript $^{-1}$ denotes standard matrix inversion, and superscript $^+$ denotes the Moore-Penrose pseudo-inverse. Recall that for the generalized inverse: $\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$, $(\mathbf{X}^+)^+ = \mathbf{X}$, and $(\mathbf{X}^T)^+ = (\mathbf{X}^+)^T$. 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. Say $\mathbf{D}$ is some diagonal matrix and recall that the transpose of a diagonal matrix is itself and thus $\mathbf{D}^T = \mathbf{D}$. We denote $\lfloor . \rfloor$ as the floor function. We use det$\{\}$ to indicate the determinant of a square symmetric matrix. Finally, we reserve some letters to have very specific meanings: (1) $\mathbf{F}$ denotes component (sometimes called factor) scores where, for example, $\mathbf{F}_I$ denotes the component scores associated with the I items, (2) $\mathbf{O}$ and $\mathbf{E}$ mean "observed" and "expected", respectively, as used in standard Pearson's X^2 analyses, where for example $\mathbf{O}_\mathbf{X}$ denotes the "observed matrix derived from $\mathbf{X}$", (3) $\mathbf{Z}$ means a centered and/or normalized matrix derived from another, e.g., $\mathbf{Z}_\mathbf{X}$ is a centered and/or normalized version of $\mathbf{X}$, and (4) blackboard bold letters mean some *a priori* or known values, for example $\mathbb{E}_\mathbf{X}$ would reflect some previously known expected values associated with $\mathbf{X}$ but not necessarily derived from $\mathbf{X}$. From this point forward we also refer to Mahalanobis distance as MD.

Overview of MCD algorithm

Given a matrix $\mathbf{X}$ with I rows and J columns where $J < I$ and $\mathbf{X}$ is full rank. The MCD requires an initialization of subset of size H—a random subsample from I rows of $\mathbf{X}$—where $\lfloor (I + J + 1)/2 \rfloor \leq H \leq I$. From $\mathbf{X}_H$ we compute the column means as $\boldsymbol{\mu}_H$. The column-wise centered (on $\boldsymbol{\mu}_H$) version of $\mathbf{X}_H$ is denoted as $\mathbf{Z}_H$. The MCD algorithm works as follows:

1. Compute the covariance matrix $\mathbf{S}_H = (\mathbf{Z}_H^T \mathbf{Z}_H) \times (H - 1)^{-1}$.
2. Compute the determinant of $\mathbf{S}_H$.
3. Compute the squared MD for each observation in $\mathbf{X}$ as $m_i = (\mathbf{x}_i - \boldsymbol{\mu}_H)\mathbf{S}_H^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_H)^T$.
4. Set $\mathbf{X}_H$ as the subset of H observations with the smallest MDs in Step 3.

The size of H is controlled by α to determine the subsample size; α exists between 0.5 and 1. The above steps are repeated until we find a minimum determinant either through optimal search or within a specified number of iterations. We denote the H subset with the minimum determinant as Ω. For the Ω subset, we obtain a robust mean vector $\boldsymbol{\mu}_\Omega$, a robust covariance $\mathbf{S}_\Omega = (\mathbf{Z}_\Omega^T \mathbf{Z}_\Omega) \times (\Omega - 1)^{-1}$, and robust square MDs for each i observation from $\mathbf{X}$ as $m_i = (\mathbf{x}_i - \boldsymbol{\mu}_\Omega)\mathbf{S}_\Omega^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_\Omega)^T$.

The two key elements for MCD algorithms are: (1) the determinant of the subsample covariance matrix, which helps identify the smallest scatter of observations and (2) the squared MDs for all observations derived from the subsample covariance matrix, which both helps identify outliers and aids in the search for a minimum determinant. Both the determinant of a matrix and the squared MDs for observations can be computed from the singular value decomposition (SVD): (1) the determinant can be computed as the geometric mean of the eigenvalues (Boudt et al., 2017; SenGupta, 1987) and (2) squared MD can be computed as the sum of the squared row elements of the singular vectors (Barkmeijer, Bouttier, & Van Gijzen, 1998; Brereton, 2015; Stephenson, 1997). In the following sections

we first show how to compute what is needed (i.e., determinants, squared MDs for subsamples and out-of-sample data) only through the SVD, which provides the basis of how to generalize the MCD to additional data types.

Mahalanobis distances via PCA for continuous data

Given $\mathbf{X}$ with I rows and J columns, and have $\mathbf{Z_X}$ as the centered and/or normalized version of $\mathbf{X}$. Assume that $\mathbf{Z_X}$ is full rank, where L is the rank of $\mathbf{Z_X}$ and $L \leq J < I$. First we define the sample covariance matrix of $\mathbf{X}$ as $\mathbf{S} = (\mathbf{Z_X^T Z_X}) \times (I-1)^{-1}$. Squared MD is $\mathbf{m} = (\mathbf{Z_X S}^{-1} \odot \mathbf{Z_X})\mathbf{1}$ where $\mathbf{1}$ is a conformable $J \times 1$ vector of 1s and $\mathbf{m}$ is a $I \times 1$ vector of squared MDs. If we relax the definition and exclude the degrees of freedom scaling factor $(I-1)$:

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

where $\mathbf{m}' = \text{diag}\{\mathbf{M}'\}$ and $\mathbf{m}' = \mathbf{m} \times (I-1)^{-1}$. Henceforth we refer to squared MD as $\mathbf{m}' = \text{diag}\{\mathbf{Z_X}(\mathbf{Z_X^T Z_X})^{-1}\mathbf{Z_X^T}\}$. Note that $\mathbf{Z_X}(\mathbf{Z_X^T Z_X})^{-1}\mathbf{Z_X^T}$ is also called the "orthogonal projection matrix" (Yanai, Takeuchi, & Takane, 2011), and that $\text{diag}\{\mathbf{Z_X}(\mathbf{Z_X^T Z_X})^{-1}\mathbf{Z_X^T}\}$ is equal to the influence measure of "leverage" in regression and PCA (Mejia, Nebel, Eloyan, Caffo, & Lindquist, 2017; Wold, Esbensen, & Geladi, 1987). We can also define squared MDs by way of principal components analysis (PCA). We can perform the PCA of $\mathbf{X}$ through the SVD as:

$$\mathbf{Z_X} = \mathbf{U\Delta V}^T \tag{2}$$

where (1) $\mathbf{U}$ and $\mathbf{V}$ are orthonormal left and right singular vectors of sizes $I \times L$ and $J \times L$, respectively with $\mathbf{U}^T\mathbf{U} = \mathbf{I} = \mathbf{V}^T\mathbf{V}$ and $\mathbf{U}^T = \mathbf{U}^+$ and $\mathbf{V}^T = \mathbf{V}^+$, and (2) $\mathbf{\Delta}$ is the $L \times L$ diagonal matrix of singular values and $\mathbf{\Lambda} = \mathbf{\Delta}^2$ which is a diagonal matrix of eigenvalues

(squared singular values). With Eq. (2) we can compute the determinant from the eigenvalues, and then the squared MDs as:

$$\begin{aligned}
\mathbf{M}' = \mathbf{Z_X}(\mathbf{Z_X^T}\mathbf{Z_X})^{-1}\mathbf{Z_X^T} = \\
\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T(\mathbf{V}\boldsymbol{\Delta}\mathbf{U}^T\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T)^{-1}\mathbf{V}\boldsymbol{\Delta}\mathbf{U}^T = \\
\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T(\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T)^{-1}\mathbf{V}\boldsymbol{\Delta}\mathbf{U}^T = \\
\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T\mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^T\mathbf{V}\boldsymbol{\Delta}\mathbf{U}^T = \\
\mathbf{U}\boldsymbol{\Delta}\Lambda^{-1}\boldsymbol{\Delta}\mathbf{U}^T = \mathbf{U}\mathbf{U}^T,
\end{aligned} \tag{3}$$

where $\mathbf{m}' = \text{diag}\{\mathbf{Z_X}(\mathbf{Z_X^T}\mathbf{Z_X})^{-1}\mathbf{Z_X^T}\} = \text{diag}\{\mathbf{U}\mathbf{U}^T\}$. In PCA, component scores are scaled (by the singular values) versions of the vectors with $\mathbf{F}_I = \mathbf{U}\boldsymbol{\Delta}$ defined as row component scores. We can also define row component scores through projection (or rotation) as:

$$\mathbf{F}_I = \mathbf{U}\boldsymbol{\Delta} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T\mathbf{V} = \mathbf{Z_X}\mathbf{V}, \tag{4}$$

where $\mathbf{G}_I = \mathbf{F}_I(\mathbf{F}_I^T\mathbf{F}_I)^{-1}\mathbf{F}_I^T$ and each row item's leverage is an element in $\mathbf{g}_I = \text{diag}\{\mathbf{G}_I\}$. If we expand and substitute $\mathbf{U}\boldsymbol{\Delta}$ for $\mathbf{F}_I$, we see that $\mathbf{G}_I = \mathbf{U}\boldsymbol{\Delta}(\boldsymbol{\Delta}\mathbf{U}^T\mathbf{U}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}\mathbf{U}^T = \mathbf{U}\mathbf{U}^T$, thus $\mathbf{m}' = \mathbf{g}_I = \text{diag}\{\mathbf{U}\mathbf{U}^T\}$.

MCD algorithms find a subsample of observations that has a minimum determinant and produces a robust covariance structure. From this robust structure, squared MDs are computed for both the subsample and the excluded—or out of sample—observations. We can compute the squared MDs on the subsample and out of sample observations through projection of data onto the space spanned by a given set of components. First, we refer to a $K \times J$ matrix $\mathbf{Y}$ as a new set of observations with the same variables (columns), and preprocessed in the same way as $\mathbf{X}$. Thus we have $\mathbf{Z_Y}$ which has been, for example, centered by the column-wise center of $\mathbf{X}$. For $\mathbf{Z_Y}$, the squared MDs are $\mathbf{M}'_Y = \mathbf{Z_Y}(\mathbf{Z_X^T}\mathbf{Z_X})^{-1}\mathbf{Z_Y^T}$. Alternatively, we can project $\mathbf{Z_Y}$ onto previously defined components, as established in Eq.

(4). To compute component scores for $\mathbf{Z_Y}$ with respect to the PCA of $\mathbf{Z_X}$ as $\mathbf{F}_K = \mathbf{Z_Y V}$. Through that projection, we can compute the squared MDs for $\mathbf{Z_Y}$ with only $\mathbf{V}$ and $\mathbf{\Delta}$ from the SVD of $\mathbf{Z_X}$ as defined in Eq. (2) as

$$\begin{aligned}
\mathbf{M}'_Y = \mathbf{Z_Y}(\mathbf{Z_X}^T\mathbf{Z_X})^{-1}\mathbf{Z_Y}^T = \\
\mathbf{Z_Y}(\mathbf{V\Delta U}^T\mathbf{U\Delta V}^T)^{-1}\mathbf{Z_Y}^T = \\
\mathbf{Z_Y}(\mathbf{V\Lambda V}^T)^{-1}\mathbf{Z_Y}^T = \\
\mathbf{Z_Y V\Lambda}^{-1}\mathbf{V}^T\mathbf{Z_Y}^T = \\
\mathbf{F}_K\mathbf{\Lambda}^{-1}\mathbf{F}_K^T,
\end{aligned} \tag{5}$$

where the squared MDs are $\mathbf{m}'_Y = \text{diag}\{\mathbf{M}'\} = \text{diag}\{\mathbf{F}_K\mathbf{\Lambda}^{-1}\mathbf{F}_K^T\}$. Now let us assume that $\mathbf{Y}$ is comprised of both included and excluded observations, or in terms of subsamples: the H included samples and the $\bar{H}$ excluded samples. Then $\mathbf{Y} = \begin{bmatrix}\mathbf{X}_H \\ \mathbf{X}_{\bar{H}}\end{bmatrix}$. In this case, $\mathbf{Z_Y}$ is, for example, centered on the column-wise mean of $\mathbf{X}_H$. We can compute squared MDs for both the H included and the $\bar{H}$ excluded samples as $\mathbf{m'_Y} = \text{diag}\{\begin{bmatrix}\mathbf{UU}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_Y V\Lambda}^{-1}\mathbf{V}^T\mathbf{Z_Y}^T\end{bmatrix}\}$.

Next we show how to use the SVD under different assumptions and metrics so that we can define a MCD algorithm for various data types. We show how to do so through an analog of PCA designed for categorical data called multiple correspondence analysis.

Multiple correspondence analysis

Multiple correspondence analysis (MCA) generalizes both (1) PCA to categorical data (with observations on the rows) and (2) standard correspondence analysis (CA) from two-way contingency to N-way contingency tables (Abdi & Valentin, 2007; Greenacre, 1984; Greenacre & Blasius, 2006; Lebart, Morineau, & Warwick, 1984). However, MCA can accommodate many other data types – which we cover in detail in the next section. MCA operates under the assumption of independence (i.e., χ^2).

Say we have an $I \times A$ matrix called $\mathbf{A}$, as in Table 1a, where the rows are observations and the columns are categorical variables where each cell contains a nominal level of each variable. These data can be represented in complete disjunctive coding—as in Table 1b—where each variable is represented by a vector of length N_a, where N_a is the number of levels or nominal values for the ath variable. We refer to the disjunctive of $\mathbf{A}$ as an $I \times J$ matrix called $\mathbf{X}$.

MCA requires specific preprocessing and constraints applied to the rows and columns in order to decompose a matrix under the assumption of independence. To perform MCA, we first compute the row and column weights:

$$\mathbf{r} = (\mathbf{1}^T\mathbf{X}\mathbf{1})^{-1} \times \mathbf{X}\mathbf{1} \text{ and } \mathbf{c} = (\mathbf{1}^T\mathbf{X}\mathbf{1})^{-1} \times \mathbf{X}^T\mathbf{1}, \tag{6}$$

where $\mathbf{r}$ and $\mathbf{c}$ are the row and column marginal sums, respectively, divided by the total sum. In the case of purely disjunctive data, each value in $\mathbf{r}$ is identical as each row contains the same number of 1s (see Table 1b): Each element in $\mathbf{r}$ is simply the total number of columns in $\mathbf{A}$ divided by the sum of all elements of $\mathbf{X}$. Similarly, $\mathbf{c}$ contains the column sums of $\mathbf{X}$ divided by the total sum. Next we compute observed ($\mathbf{O_X}$) and expected ($\mathbf{E_X}$) matrices of $\mathbf{X}$ as $\mathbf{O_X} = (\mathbf{1}^T\mathbf{X}\mathbf{1})^{-1} \times \mathbf{X}$ and $\mathbf{E_X} = \mathbf{r}\mathbf{c}^T$, respectively, and deviations computed as $\mathbf{Z_X} = \mathbf{O_X} - \mathbf{E_X}$. MCA is performed with the generalized SVD (GSVD) of $\mathbf{Z_X}$:

$$\begin{gathered} \mathbf{Z_X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T, \text{where} \\ \mathbf{P}^T\mathbf{W}_I\mathbf{P} = \mathbf{I} = \mathbf{Q}^T\mathbf{W}_J\mathbf{Q}, \end{gathered} \tag{7}$$

and where $\mathbf{W}_I = \text{diag}\{\mathbf{r}\}^{-1}$ and $\mathbf{W}_J = \text{diag}\{\mathbf{c}\}^{-1}$. The *generalized* singular vectors are computed as $\mathbf{P} = \mathbf{W}_I^{-\frac{1}{2}}\mathbf{U}$ and $\mathbf{Q} = \mathbf{W}_J^{-\frac{1}{2}}\mathbf{V}$. We obtain the results of the GSVD of $\mathbf{Z_X}$ through the SVD of $\widetilde{\mathbf{Z}}_\mathbf{X} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{Z_X}\mathbf{W}_J^{\frac{1}{2}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T$ as in Eq. (2). We can see the relationship between $\mathbf{Z_X}$ and $\widetilde{\mathbf{Z}}_\mathbf{X}$ via substitution:

Table 1

Example categorical and disjunctive data.

(a) *Categorical representation of two variables.*

	Var. 1	Var. 2
*Subj.*1	B	YES
*Subj.*2	A	YES
...	...	...
Subj.I-1	A	NO
Subj.I	C	YES

(b) *Disjunctive representation of two variables.*

	Var. 1			Var. 2	
	A	B	C	YES	NO
*Subj.*1	0	1	0	1	0
*Subj.*2	1	0	0	1	0
...	...	...	...	...	...
Subj.I-1	1	0	0	0	1
Subj.I	0	0	1	1	0

Note. Hypothetical example of categorical data. (a) shows two variables Variable 1 and Variable 2 with 3 and 2 levels respectively. (b) shows the disjunctive code for these data. Note that the sum of each row is equal to the total number of original variables, and that the sum within each set of disjunctive columns per variable is equal to the total number of rows. The column sums indicate the frequency of a particular level within the data.

$$\begin{aligned}\mathbf{Z_X} = \mathbf{W}_I^{-\frac{1}{2}}\widetilde{\mathbf{Z}}_\mathbf{X}\mathbf{W}_J^{-\frac{1}{2}} &= \\ \mathbf{W}_I^{-\frac{1}{2}}\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T\mathbf{W}_J^{-\frac{1}{2}} &= \\ (\mathbf{W}_I^{-\frac{1}{2}}\mathbf{U})\boldsymbol{\Delta}(\mathbf{W}_J^{-\frac{1}{2}}\mathbf{V})^\mathbf{T} &= \\ \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T.\end{aligned} \tag{8}$$

For simplicity, we refer to the GSVD where needed in triplet notation as GSVD($\mathbf{W}_I, \mathbf{Z_X}, \mathbf{W}_J$) with row constraints (e.g., $\mathbf{W}_I$), data (e.g., $\mathbf{Z_X}$), and column constraints (e.g., $\mathbf{W}_J$). To note, we have taken liberty with the standard GSVD triplet notation (see Holmes, 2008) and present the triplet more akin to its multiplication steps.

A computational note on MCA. Because of the way we represent data for MCA—see the disjunctive coding example in Table 1—and because of how the SVD works, MCA produces "null dimensions" (Greenacre, 1984). These "null dimensions" are produced through the use of the SVD (or eigendecomposition) where the singular (or eigen) vectors are undetermined and the singular (or eigen) values are effectively zero. These "null dimensions" are to be discarded. In all subsequent sections, our use of the MCA, CA, and related methods also excludes these "null dimensions". In effect, we only retain the dimensions with non-zero eigenvalues.

Mahalanobis distances via MCA for categorical data

We showed for continuous data that we can obtain the squared MD from the diagonal of the crossproduct of left singular vectors as diag$\{\mathbf{U}\mathbf{U}^T\}$. We use the same premise to define a squared MD for categorical data by way of MCA. MCA is performed as GSVD($\mathbf{W}_I, \mathbf{Z_X}, \mathbf{W}_J$) where we apply the SVD as $\widetilde{\mathbf{Z}}_\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T$ (see Eqs. (7) and (8). Recall that $\widetilde{\mathbf{Z}}_\mathbf{X} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{Z_X}\mathbf{W}_J^{\frac{1}{2}} \Longleftrightarrow \mathbf{Z_X} = \mathbf{W}_I^{-\frac{1}{2}}\widetilde{\mathbf{Z}}_\mathbf{X}\mathbf{W}_J^{-\frac{1}{2}}$ where $\mathbf{Z_X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T$ (see Eq. (8)). In the case of MCA, $\widetilde{\mathbf{Z}}_\mathbf{X}$ would be the analog of $\mathbf{Z_X}$ in PCA (see Eq. (2)), thus via the SVD we still have $\mathbf{U}\mathbf{U}^T = \widetilde{\mathbf{Z}}_\mathbf{X}(\widetilde{\mathbf{Z}}_\mathbf{X}^T\widetilde{\mathbf{Z}}_\mathbf{X})^+\widetilde{\mathbf{Z}}_\mathbf{X}^T$. We replace the standard inversion with the pseudo-inverse—and only use non-null dimensions—because of the inherent collinearity in

$(\widetilde{\mathbf{Z}}_{\mathbf{X}}^T\widetilde{\mathbf{Z}}_{\mathbf{X}})$. Furthermore $\widetilde{\mathbf{Z}}_{\mathbf{X}}(\widetilde{\mathbf{Z}}_{\mathbf{X}}^T\widetilde{\mathbf{Z}}_{\mathbf{X}})^+\widetilde{\mathbf{Z}}_{\mathbf{X}}^T = \mathbf{Z}_{\mathbf{X}}(\mathbf{Z}_{\mathbf{X}}^T\mathbf{Z}_{\mathbf{X}})^+\mathbf{Z}_{\mathbf{X}}^T$, which leads to

$$\begin{aligned}
\mathbf{M}'_{\mathbf{Z_X}} = \mathbf{Z}_{\mathbf{X}}(\mathbf{Z}_{\mathbf{X}}^T\mathbf{Z}_{\mathbf{X}})^+\mathbf{Z}_{\mathbf{X}}^T = \\
\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T\{\mathbf{Q}\boldsymbol{\Delta}\mathbf{P}^T\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T\}^+\mathbf{Q}\boldsymbol{\Delta}\mathbf{P}^T = \\
\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T\{\mathbf{Q}\boldsymbol{\Delta}\mathbf{W}_I^{-\frac{1}{2}}\mathbf{W}_I^{-\frac{1}{2}}\boldsymbol{\Delta}\mathbf{Q}^T\}^+\mathbf{Q}\boldsymbol{\Delta}\mathbf{P}^T = \\
\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T(\mathbf{Q}^+)^T\boldsymbol{\Delta}^+\mathbf{W}_I^{\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}\boldsymbol{\Delta}^+\mathbf{Q}^+\mathbf{Q}\boldsymbol{\Delta}\mathbf{P}^T = \\
\mathbf{P}\boldsymbol{\Delta}\boldsymbol{\Delta}^+\mathbf{W}_I^{\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}\boldsymbol{\Delta}^+\boldsymbol{\Delta}\mathbf{P}^T = \\
\mathbf{P}\mathbf{W}_I^{\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}\mathbf{P}^T = \\
\mathbf{W}_I^{-\frac{1}{2}}\mathbf{U}\mathbf{W}_I^{\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}\mathbf{U}^T\mathbf{W}_I^{-\frac{1}{2}} = \mathbf{U}\mathbf{U}^T.
\end{aligned} \tag{9}$$

Thus $\mathbf{M}' = \widetilde{\mathbf{Z}}_{\mathbf{X}}(\widetilde{\mathbf{Z}}_{\mathbf{X}}^T\widetilde{\mathbf{Z}}_{\mathbf{X}})^+\widetilde{\mathbf{Z}}_{\mathbf{X}}^T = \mathbf{Z}_{\mathbf{X}}(\mathbf{Z}_{\mathbf{X}}^T\mathbf{Z}_{\mathbf{X}})^+\mathbf{Z}_{\mathbf{X}}^T = \mathbf{U}\mathbf{U}^T$ is a squared MD for categorical data via MCA, à la Eqs. (1) and (3). Because of the way we define squared MD here—the singular vectors for the observations (i.e., $\mathbf{U}$)—there exists no ambiguity nor arbitrary decisions in the definition of squared MD for categorical data (à la, Goodall, 1966).

We can compute MCA row component scores as $\mathbf{F}_I = \mathbf{W}_I\mathbf{P}\boldsymbol{\Delta}$ or through projections. For MCA component scores through projections, it is easier to use "profile" matrices where $\boldsymbol{\Phi}_I = \mathbf{W}_I\mathbf{O}_{\mathbf{X}}$ are the row profiles. For profile matrices, each element in $\mathbf{X}$ divided by its respective row sum for the row profile matrices. MCA row component scores via projection are:

$$\begin{aligned}
\mathbf{F}_I = \boldsymbol{\Phi}_I\mathbf{F}_J\boldsymbol{\Delta}^{-1} = \boldsymbol{\Phi}_I\mathbf{W}_J\mathbf{Q}\boldsymbol{\Delta}\boldsymbol{\Delta}^{-1} = \\
\boldsymbol{\Phi}_I\mathbf{W}_J\mathbf{Q} = \boldsymbol{\Phi}_I\mathbf{W}_J\mathbf{W}_J^{-\frac{1}{2}}\mathbf{V} = \\
\boldsymbol{\Phi}_I\mathbf{W}_J^{\frac{1}{2}}\mathbf{V} = \mathbf{W}_I\mathbf{O}_{\mathbf{X}}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V}.
\end{aligned} \tag{10}$$

Note that $\text{diag}\{\mathbf{F}_I\mathbf{F}_I^T\}$ are referred to as χ^2-distances. We can compute sqaured MDs through row component scores where
$\mathbf{F}_I = \mathbf{W}_I\mathbf{P}\boldsymbol{\Delta} = \mathbf{W}_I\mathbf{O}_{\mathbf{X}}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V} = \mathbf{W}_I\mathbf{W}_I^{-\frac{1}{2}}\mathbf{U}\boldsymbol{\Delta} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{U}\boldsymbol{\Delta}$ and so:

$$\mathbf{U} = \mathbf{W}_I^{-\frac{1}{2}}(\mathbf{W}_I^{\frac{1}{2}}\mathbf{U}\boldsymbol{\Delta})\boldsymbol{\Delta}^{-1} \text{ and}$$
$$\mathbf{U} = \mathbf{W}_I^{-\frac{1}{2}}(\mathbf{W}_I\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V})\boldsymbol{\Delta}^{-1} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V}\boldsymbol{\Delta}^{-1}. \tag{11}$$

While Eq. (10) and Eq. (11) suggest ways to compute squared MD for sub- or out-of sample categorical data—akin to Eq. (5)—there exists a major barrier to the computation of *correct* squared MD for sub- or out-of sample categorical data. The sub-sampling procedure cannot be performed on the original categorical data $\mathbf{A}$ nor on the initial disjunctive table $\mathbf{X}$. We cannot subsample $\mathbf{A}$ and $\mathbf{X}$ because levels within categorical variables will likely be excluded in the subsample. Exclusion of levels produces incorrect (i.e., lower) estimates of squared MDs for the excluded subsample. To correctly compute MDs, we need to maintain the representation of all levels. To do so, we require a specific form of MCA called subset MCA (Greenacre, 2017; Greenacre & Pardo, 2006), which is a specific case of Generalized CA (Escofier, 1983, 1984).

Escofier's Generalized Correspondence Analysis and Subset MCA for the MCD.

Hiding in the broader CA literature is "generalized correspondence analysis" (GCA: Escofier, 1983, 1984)—as referred to by Grassi and Visentin (1994). Escofier established the idea of GCA as "(DATA - MODEL) / MARGINS". GCA has many uses well beyond the standard applications in CA and MCA such as for missing data and assumptions that deviate from the typical assumptions of independence (χ^2) for CA (e.g., quasi-CA with "quasi-independence" and "quasi-margins"; Leeuw & Heijden, 1988; Van der Heijden, De Falguerolles, & Leeuw, 1989). With a complete disjunctive data set $\mathbf{X}$, for the GCA perspective of MCA, we characterize each cell of $\widetilde{\mathbf{Z}}_\mathbf{X}$ with the generic element $\widetilde{z}_{\mathbf{X}_{ij}} = \frac{(o_{\mathbf{X}_{ij}} - e_{ij})}{\sqrt{w_{I_{ii}} w_{J_{jj}}}}$. The model ($\mathbb{E}$) and the margins ($\mathbb{W}_I$ and $\mathbb{W}_J$), could be defined in almost any way; it is not required to compute the model (expected values) nor margins from the data. To help simplify the notation, we can use the GSVD to represent the GCA approach:

$\mathbb{Z}_{\mathbf{X}} = \mathbf{O}_{\mathbf{X}} - \mathbb{E}$ and GSVD$(\mathbb{W}_I, \mathbb{Z}_{\mathbf{X}}, \mathbb{W}_J)$ (see Eqs. (7) and (8)). For our problem, we only need a simple change to the generic element $\tilde{z}_{\mathbf{X}_{ij}}$. First we compute all matrices required for MCA exactly as initially defined in Eqs. (7) and (8): $\mathbf{O}_{\mathbf{X}}$, $\mathbf{E}_{\mathbf{X}}$, $\mathbf{r}$ and $\mathbf{c}$.

Let us assume that $\mathbf{O}_{\mathbf{X}}$, $\mathbf{E}_{\mathbf{X}}$, and $\mathbf{r}$ are constructed as partitions defined by the H and $\bar{H}$ subsets as: $\mathbf{O}_{\mathbf{X}} = \begin{bmatrix} \mathbf{O}_{\mathbf{X}_H} \\ \mathbf{O}_{\mathbf{X}_{\bar{H}}} \end{bmatrix}$, $\mathbf{E}_{\mathbf{X}} = \begin{bmatrix} \mathbf{E}_{\mathbf{X}_H} \\ \mathbf{E}_{\mathbf{X}_{\bar{H}}} \end{bmatrix}$, and $\mathbf{r} = \begin{bmatrix} \mathbf{w}_{I_H} \\ \mathbf{w}_{I_{\bar{H}}} \end{bmatrix}$. To maintain the structure of the data and to be able to perform the necessary steps for a generalized MCD we define a generalized expected matrix as $\mathbb{E} = \begin{bmatrix} \mathbf{E}_{\mathbf{X}_H} \\ \mathbf{O}_{\mathbf{X}_{\bar{H}}} \end{bmatrix}$; that is, we replace part of the expected matrix with part of our observed matrix. Our assumption is that the expected values are the observed, thus $\mathbb{Z}_{\mathbf{X}} = \begin{bmatrix} \mathbf{Z}_{\mathbf{X}_H} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{O}_{\mathbf{X}_H} \\ \mathbf{O}_{\mathbf{X}_{\bar{H}}} \end{bmatrix} - \begin{bmatrix} \mathbf{E}_{\mathbf{X}_H} \\ \mathbf{O}_{\mathbf{X}_{\bar{H}}} \end{bmatrix}$. Any subsample of H and $\bar{H}$ guarantee that we maintain the same shape (rows and columns) as $\mathbf{X}$ when we use $\mathbb{E} = \begin{bmatrix} \mathbf{E}_{\mathbf{X}_H} \\ \mathbf{O}_{\mathbf{X}_{\bar{H}}} \end{bmatrix}$.

Like in Eqs. (7) and (8), we have $\begin{bmatrix} \mathbf{Z}_{\mathbf{X}_H} \\ \mathbf{0} \end{bmatrix} = \mathbb{Z}_{\mathbf{X}} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T = \begin{bmatrix} \mathbf{P}_H \\ \mathbf{0} \end{bmatrix} \boldsymbol{\Delta}\mathbf{Q}^T$, where $\mathbf{P}^T\mathbf{W}_I\mathbf{P} = \mathbf{I} = \mathbf{Q}^T\mathbf{W}_J\mathbf{Q}$ and $\mathbf{P}^T\mathbf{W}_I\mathbf{P} = \begin{bmatrix} \mathbf{P}_H \\ \mathbf{0} \end{bmatrix}^T \mathbf{W}_I \begin{bmatrix} \mathbf{P}_H \\ \mathbf{0} \end{bmatrix} = \mathbf{I}$. Furthermore, recall that $\tilde{\mathbb{Z}}_{\mathbf{X}} = \mathbf{W}_I^{\frac{1}{2}}\mathbb{Z}_{\mathbf{X}}\mathbf{W}_J^{\frac{1}{2}}$ and thus $\mathbf{W}_I^{\frac{1}{2}}\mathbb{Z}_{\mathbf{X}}\mathbf{W}_J^{\frac{1}{2}} = \mathbf{W}_I^{\frac{1}{2}} \begin{bmatrix} \mathbf{Z}_{\mathbf{X}_H} \\ \mathbf{0} \end{bmatrix} \mathbf{W}_J^{\frac{1}{2}} = \begin{bmatrix} \tilde{\mathbf{Z}}_{\mathbf{X}_H} \\ \mathbf{0} \end{bmatrix} = \tilde{\mathbb{Z}}_{\mathbf{X}}$. Just as in Eq. (8) we would apply the SVD as:

$$\tilde{\mathbb{Z}}_{\mathbf{X}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T = \begin{bmatrix} \mathbf{U}_H \\ \mathbf{0} \end{bmatrix} \boldsymbol{\Delta}\mathbf{V}^T, \tag{12}$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{I} = \mathbf{V}^T\mathbf{V}$ and $\mathbf{U}^T\mathbf{U} = \begin{bmatrix} \mathbf{U}_H \\ \mathbf{0} \end{bmatrix}^T \begin{bmatrix} \mathbf{U}_H \\ \mathbf{0} \end{bmatrix} = \mathbf{U}_H^T\mathbf{U}_H = \mathbf{I}$.

We can simplify this approach through a specific application of GCA called "subset MCA" (Greenacre, 2017; Greenacre & Pardo, 2006; Hendry, North, Zewotir, & Naidoo,

2014). To compute subset MCA for use in the MCD we use the matrices defined in Eqs. (7) and (8). Specifically $\widetilde{\mathbf{Z}}_{\mathbf{X}} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{Z}_{\mathbf{X}}\mathbf{W}_J^{\frac{1}{2}} \Longleftrightarrow \mathbf{Z}_{\mathbf{X}} = \mathbf{W}_I^{-\frac{1}{2}}\widetilde{\mathbf{Z}}_{\mathbf{X}}\mathbf{W}_J^{-\frac{1}{2}}$. In subset MCA we require $\widetilde{\mathbf{Z}}_{\mathbf{X}_H}$ which is the H subsample of $\widetilde{\mathbf{Z}}_{\mathbf{X}}$ and perform the SVD as

$$\widetilde{\mathbf{Z}}_{\mathbf{X}_H} = \mathbf{U}_H\boldsymbol{\Delta}_H\mathbf{V}_H^T. \tag{13}$$

Which leaves us with the complement $\bar{H}$ where $\widetilde{\mathbf{Z}}_{\mathbf{X}_{\bar{H}}} = \mathbf{U}_{\bar{H}}\boldsymbol{\Delta}_{\bar{H}}\mathbf{V}_{\bar{H}}^T$.

We can compute the determinant for H subsamples via the geometric mean of the eigenvalues ($\boldsymbol{\Lambda}_H = \boldsymbol{\Delta}_H^2$) and we can compute robust squared MDs in subset MCA with the row profiles as in Eqs. (10) and (11). We compute the component scores for the full I set from subset MCA as

$$\begin{bmatrix}\mathbf{F}_H \\ \widehat{\mathbf{F}}_{\bar{H}}\end{bmatrix} = \boldsymbol{\Phi}_I\mathbf{F}_{J_H}\boldsymbol{\Delta}_H^{-1} \tag{14}$$

which allows us to compute $\mathbf{U} = \mathbf{W}_I^{-\frac{1}{2}}\boldsymbol{\Phi}_I\mathbf{F}_J\boldsymbol{\Lambda}^{-1}$ because $\mathbf{W}_I^{-\frac{1}{2}}\boldsymbol{\Phi}_I\mathbf{F}_J\boldsymbol{\Lambda}^{-1} = \mathbf{W}_I^{-\frac{1}{2}}\boldsymbol{\Phi}_I\mathbf{W}_J\mathbf{Q}\boldsymbol{\Delta}\boldsymbol{\Lambda}^{-1} = \mathbf{W}_I^{-\frac{1}{2}}\boldsymbol{\Phi}_I\mathbf{W}_J\mathbf{W}_J^{-\frac{1}{2}}\mathbf{V}\boldsymbol{\Delta}\boldsymbol{\Lambda}^{-1} = \mathbf{W}_I^{-\frac{1}{2}}\boldsymbol{\Phi}_I\mathbf{W}_J\mathbf{W}_J^{-\frac{1}{2}}\mathbf{V}\boldsymbol{\Delta}^{-1} = \mathbf{W}_I^{-\frac{1}{2}}\mathbf{W}_I\mathbf{O}_{\mathbf{X}}\mathbf{W}_J\mathbf{W}_J^{-\frac{1}{2}}\mathbf{V}\boldsymbol{\Delta}^{-1} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{O}_{\mathbf{X}}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V}\boldsymbol{\Delta}^{-1}$. We can compute the squared MDs for the analyzed and excluded subsamples from the subset MCA results as:

$$\begin{bmatrix}\mathbf{U}_H \\ \widehat{\mathbf{U}}_{\bar{H}}\end{bmatrix} = \mathbf{W}_I^{-\frac{1}{2}}\boldsymbol{\Phi}_I\mathbf{F}_{J_H}\boldsymbol{\Lambda}_H^{-1} \tag{15}$$

where the $\mathbf{U}_H$ in Eq. (15) is exactly the $\mathbf{U}_H$ in Eqs. (12) and (13). Alternatively we can compute Eq. (15) from (14) as $\begin{bmatrix}\mathbf{U}_H \\ \widehat{\mathbf{U}}_{\bar{H}}\end{bmatrix} = \mathbf{W}_I^{-\frac{1}{2}}\begin{bmatrix}\mathbf{F}_H \\ \widehat{\mathbf{F}}_{\bar{H}}\end{bmatrix}\boldsymbol{\Delta}^{-1}$, and thus squared MDs are computed as $\text{diag}\{\begin{bmatrix}\mathbf{U}_H \\ \widehat{\mathbf{U}}_{\bar{H}}\end{bmatrix}\begin{bmatrix}\mathbf{U}_H \\ \widehat{\mathbf{U}}_{\bar{H}}\end{bmatrix}^T\}$.

MCD algorithm for categorical data

Here we outline the core steps required for a MCD via subset MCA. First the MCD requires an initial random subsample of size H where $\lfloor(I+J+1)/2\rfloor \leq H \leq I$. The H

subset of $\widetilde{\mathbf{Z}}_{\mathbf{X}}$ is denoted as $\widetilde{\mathbf{Z}}_{\mathbf{X}_H}$.

1. Apply the SVD: $\widetilde{\mathbf{Z}}_{\mathbf{X}_H} = \mathbf{U}_H \boldsymbol{\Delta}_H \mathbf{V}_H^T$

2. Compute the determinant of $\widetilde{\mathbf{Z}}_{\mathbf{X}_H}$ as the geometric mean of $\boldsymbol{\Lambda}_H$

3. Compute the squared MDs for the full I sample as $\begin{bmatrix} \mathbf{U}_H \\ \widehat{\mathbf{U}}_{\bar{H}} \end{bmatrix} = \mathbf{W}_I^{-\frac{1}{2}} \boldsymbol{\Phi}_I \mathbf{F}_{J_H} \boldsymbol{\Lambda}_H^{-1}$

4. Set $\widetilde{\mathbf{Z}}_{\mathbf{X}_H}$ as the subset of H observations with the smallest squared MDs computed from Step 3.

These steps are repeated until a we find a minimum determinant (Step 2) or stop at a given number of iterations. Like in the traditional MCD, Step 3 helps find a minimum determinant because the subset of H observations with the smallest squared MDs produce a smaller or equal determinant in subsequent iterations.

Differences between traditional MCD and our proposed MCD. Primarily, our approach to the generalization of a MCD for categorical data maintains the two most important features of the MCD algorithm: (1) search for a minimum determinant to identify a robust covariance structure and (2) compute robust squared MDs with respect to that structure, in order to identify outliers. In effect, like with the standard MCD approach, ours also helps in diagnostic steps that could be susceptible to masking effects. However there are some key differences between our approach to a MCD algorithm for categorical data and the traditional MCD (which is, generally, only defined for continous data). These differences stem from our approach via MCA and the constraints imposed on the problem by categorical data. The most considerable issue with categorical data is that rare response levels are likely sources of an individual's outlierness. But those levels cannot be removed, else the individuals with those response levels end up with *smaller* MDs simply because those rare levels are no longer represented. We retain all the key computational aspects, while relaxing some of the theoretical.

First, our approach uses what should be considered a pseudo-determinant. Both $\det\{\widetilde{\mathbf{Z}}_{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{Z}}_{\mathbf{X}}\}$ and $\det\{\mathbf{Z}_{\mathbf{X}}^{\mathrm{T}}\mathbf{Z}_{\mathbf{X}}\}$ are by definition zero because the multiple columns used to represent a single variable within the disjunctive table (see Table 1) are linear combinations of one another. Computationally, we use the full set of non-null dimensions to compute the pseudo-determinant (i.e., all components with non-zero eigenvalues). Related to this is that it is possible for different subsets, in the search for a minimum determinant, to have different dimensionality; specifically, subsets could span fewer dimensions than the original data. However in practice, this is not much of an issue because fewer dimenions does not necessarily produce a smaller pseudo-determinant (i.e., geometric mean of the eigenvalues for the existing dimensions). With respect to dimensionality in MCA, the "true" dimensionality is typically much smaller than just the non-null dimensions (i.e., effectively zero eigenvalue). Typically in MCA, a correction is applied to retain fewer dimensions and then adjust the eigenvalues through, for example, the Benzecri (Benzécri, 1979) or Greenacre (Greenacre, 2017) corrections. While this is typical for MCA, we recommend against this for our MCD procedure. Generally, the "masking effect" occurs because of anomalous values in the smaller dimensions (i.e., typically dimensions with small eigenvalues).

In the traditional MCD, a new center is computed for each H subsample in order to center that H subsample on its respective column means. In turn, the subsample identified for the minimum determinant produces a "robust center". Our generalized approach to MCD does not produce a robust center, per se. Rather, via the search procedure in our approach, we detect both the center and minimum scatter (smallest determinant) from a set of points *within the existing space*, as opposed to defining a new space from the subsample. As previously noted, this is a limitation imposed by categorical data: if we were to drop levels of variables, then the MD of individuals with those levels goes down; this is because those individuals would have exclusively zeros in the disjunctive table. However there is one way—under certain conditions—where a robust center could be computed. A robust center—i.e., the expected matrix $\mathbf{E}_{\mathbf{X}}$—could be computed from different *row weights*, say,

from only the rows included in the H subsample. However, this is only possible (or relevant) when the rows (observations) have different weights. In our approach, all rows have weights of $\frac{1}{I}$. Even with different row weights, our approach leverages the subset CA procedure and thus still searches for a center and minimum scatter within an existing space (defined under assumptions of independence).

Examples, applications, and extensions

We provide three sets of examples to illustrate the use of our generalized MCD. Because our approach is based on CA, we can capitalize on particular properties of CA to allow for the analyses of non-categorical data such as ordinal, continuous, or mixed ("heterogeneous") data. We show how to accomodate different data types—including mixed—through a data recoding technique called data "doubling" (a.k.a. fuzzy coding, bipolar coding, or "Escofier transform"). We first illustrate our generalized MCD with a small toy genetic data set of single nucleotide polymorphisms (SNPs; see supplemental section from Beaton, Dunlop, and Abdi (2016)) which are entirely categorical. Next we analyze the miniature survey of autobiographical memory (MSAM; adapted from Palombo, Williams, Abdi, & Levine, 2013), from the Ontario Neurodegernative Disease Research Initiative (ONDRI). While the MSAM is comprised entirely of ordinal responses, we apply our approach to two versions of the MSAM: (1) where each question is treated as a categorical variable, and (2) an approach that preserves the ordinality of the data. We compare and contrast the results from these applications to highlight how GMCD results change. Finally, we provide an analysis of a data set from the Alzheimer's Disease Neuroimaging Initiative (ADNI) that spans numerous types: categorical data (SNPs), ordinal data (clinical dementia rating), and continuous data (volumetric estimates from brain regions).

Toy data

The toy data are comprised of SNPs, which are categorical variables. The toy data available as part of the `OuRS` package: https://github.com/derekbeaton/ours. Each SNP

contains three possible genotypes: “AA”, “Aa”, and “aa” which are the major homozygote, the heterozygote, and the minor homozygote, respectively. In this example, we set $\alpha = 0.5$.

The toy data are of size $I = 60 \times C = 9$ (observations by variables). These data were transformed into complete disjunctive coding (as in Table 1) that are of size $I = 60 \times J = 25$ (observations by columns); note that some SNPs have 3 observed genotypes where as others only have 2 observed genotypes. Figure 1 shows GMCD applied to the toy data set over the course of $1,000$ iterations. Figure 1a shows the search for a minimum determinant (a decrease over iterations). In Figures 1b-d, individuals in red are those within the H subsample and individuals in blue are those excluded from the H subsample (i.e., the subsample with the smallest determinant). Figure 1b shows the observed squared MD (horizontal) and χ^2-distances (vertical). The χ^2-distances are the squared component scores (*cf.* Eq. (10)). We have denoted two individuals with “X” circumscribed by a circle. These two individuals are the only two with “aa” for one particular SNP. Finally, we show the robust squared MD, which helps reveal “masking” effects: Figure 1c shows the observed vs. robust squared MD. Note the substantial increase of the two previously denoted outliers. Figure 1d is the same as Figure 1d except without the two most extreme individuals to provide a better view of the observed vs. robust squared MDs.

ONDRI Data

The Ontario Neurodegenerative Disease Research Initiative (ONDRI; http://ondri.ca/) is a longitudinal, multi-site, “deep-phenotyping” study across multiple neurodegenerative disease cohorts (Farhan et al., 2017). The data presented here are preliminary data, and used for illustrative purposes. As part of the quality control process, various subsets of data are subjected to an outlier analyses pipeline (Sunderland et al., 2019) via the MCD and CorrMax (Garthwaite & Koch, 2016) procedures. In ONDRI, many individual variables or entire instruments are not quantitative, such as the MSAM. The MSAM is a short version of the survey of autobiographical memory (SAM). The MSAM is comprised of 10 questions

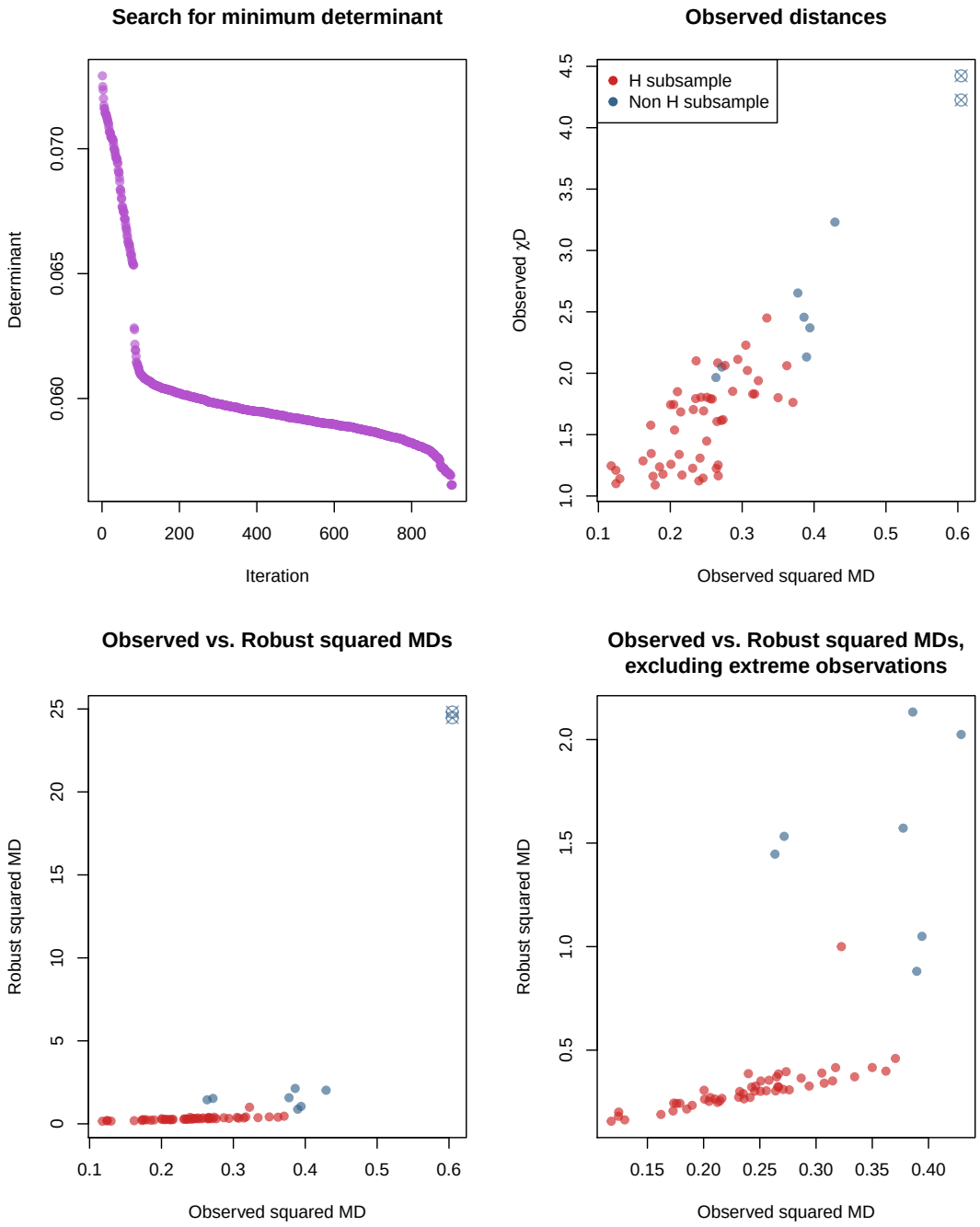

Figure 1. Application of generalized MCD to categorical data. We illustrate the generalized MCD here first on a small data set of single nucleotide polymorphisms (SNPs). (a) shows the search for the determinant. (b) shows the observed distances with the final H subset colored in red. (c) shows the observed vs. robust squared Mahalanobis distances (MDs). (d) shows the same as (c) but excludes the two extreme individuals (which were the only two to have 'aa' for one particular SNP.

where each question has five possible responses: two responses indicate disagreement with a question ("strongly disagree", "disagree somewhat"), two responses indicate agreement with a question ("agree somewhat", "strongly agree") and one neutral respone ("neither"). Here we apply two strategies to the data—as categorical, then as ordinal—on $N = 300$ from two cohorts (cerebrovascular disease $n = 161$ and Parkinson's disease $n = 139$). We compare and contrast the results of the two strategies and at three α levels: $\alpha = 0.5$, $\alpha = 0.75$, $\alpha = 0.9$ (for a total of six applications of the GMCD).

First, we treat the data as categorical and apply our technique as previously outlined. Next, we use a specific recoding scheme for ordinal data which has many names: "bipolar coding" (Greenacre, 1984), fuzzy coding (Goldfarb & Pardoux, 2006), or "doubling" (Greenacre, 2014; Lebart et al., 1984). We refer to the doubling approach for ordinal data as "fuzzy coding" (described below). Fuzzy coding transforms variables into "pseudo-disjunctive" in that, from the perspective of CA and MCA, the data tables behave like disjuntive tables (see Table 1): the sums of the rows equal the number of (original) variables, the sum of variables (each pair of columns) equal the number of rows, and the sum of the table equals the number of rows $\times$ the number of (original) variables. With fuzzy coding, we represent each original variable with two columns that reflect the distance between an ordinal response and its expected minimum or expected maximum. The pair of columns for each variable, referred to here as "$-$" and "$+$", sum to 1. So "$-$" $= \frac{\max - x}{\max - \min}$ reflects how close to the minimum of the scale the observed (x) value is, and "$+$" $= \frac{x - \min}{\max - \min}$ reflects how close to the maximum of the scale the observed (x) value is. The MSAM has 5 possible levels (1, 2, 3, 4, 5), thus our expected—but not necessarily observed—minimum and maximum for each response is 1 and 5, respectively. We illustrate fuzzy coding in Table 2.

As in the Fast-MCD algorithm, α controls the proportion of samples to identify as the H subset. The size of H is computed as $H = \lfloor (2 \times H_I) - I + (2 \times (I - H_I) \times \alpha) \rfloor$ where $H_I = \mod((I + J + 1), 2)$ (which is the same subsample size computation as the `rrcov`

Table 2

Example of fuzzy coding

(a) *CDR ordinal example*

	Q1: Specific events	Q2: Recall objects
OND01_TWH_1100	2	4
OND01_SBH_1200	1	4
...	...	...
OND01_HDH_5100	5	4
OND01_HDH_5201	3	5

(b) *Doubling for ordinal data ("fuzzy coding") with normalization so each variable sums to 1*

	Q1−	Q1+	Q2−	Q2+
OND01_TWH_1100	$\frac{5-2}{5-1} = 0.75$	$\frac{2-1}{5-1} = 0.25$	$\frac{5-4}{5-1} = 0.25$	$\frac{4-1}{5-1} = 0.75$
OND01_SBH_1200	$\frac{5-1}{5-1} = 1.00$	$\frac{1-1}{5-1} = 0.00$	$\frac{5-4}{5-1} = 0.25$	$\frac{4-1}{5-1} = 0.75$
...	...	...	...	...
OND01_HDH_5100	$\frac{5-5}{5-1} = 0.00$	$\frac{5-1}{5-1} = 1.00$	$\frac{5-4}{5-1} = 0.25$	$\frac{4-1}{5-1} = 0.75$
OND01_HDH_5201	$\frac{5-3}{5-1} = 0.50$	$\frac{3-1}{5-1} = 0.50$	$\frac{5-5}{5-1} = 0.00$	$\frac{5-1}{5-1} = 1.00$

Note. Example of data doubling via fuzzy coding for ordinal data with the miniature survey of autobiographical memory (MSAM). Note that the sum of each row is equal to the total number of original variables, and that the sum within each pair of columns per variable is equal to the total number of rows.

and `robustbase` packages in `R` via the `h.alpha.n` function). Because the categorical and the ordinal versions of the data sets have a different number of columns (J) after transformation into disjunctive or pseudo-disjunctive data, the same α produced a different size H for each data set. Each application of these analyses had an upper limit of $1,000$ iterations to search for a minimum determinant.

Figure 2 shows all six applications of the generalized MCD applied to the MSAM. To note in Figure 2 we show the MDs and robust MDs (not the squared MDs or robust MDs), because this better highlights the differences on a smaller scale for the axes. Generally, when α is low, we see greater exaggeration of outlying individuals with respect to the "inliers" and the H subset. However this behavior essentially disappears as α increases. We also see that the "ordinal" version of the data behaves quite differently from the categorical version, especially as α increases. In particular, we see considerable changes in the robust MDs for the individuals excluded from the H subset.

There are differences in how the robust MDs express themselves with respect to the robust covariance structure in each analysis. We can see that the categorical version exaggerates the robust MD across all versions, which highlights the masking effect. We still see this effect in the ordinal version, but only with the lower α level – once α is increased, the robust and standard MDs are quite similar. However, even though the robust MDs express varying patterns (see Fig. 2), the identification of the H subsamples across the different analyses are nearly identical. The overlap of the H subsamples (see Table 3) shows that each approach to the data (categorical or ordinal) identifies many of the same H subsample. The main difference between the categorical and ordinal approaches is the magnitude of the robust MDs with respect to their robust covariance structures.

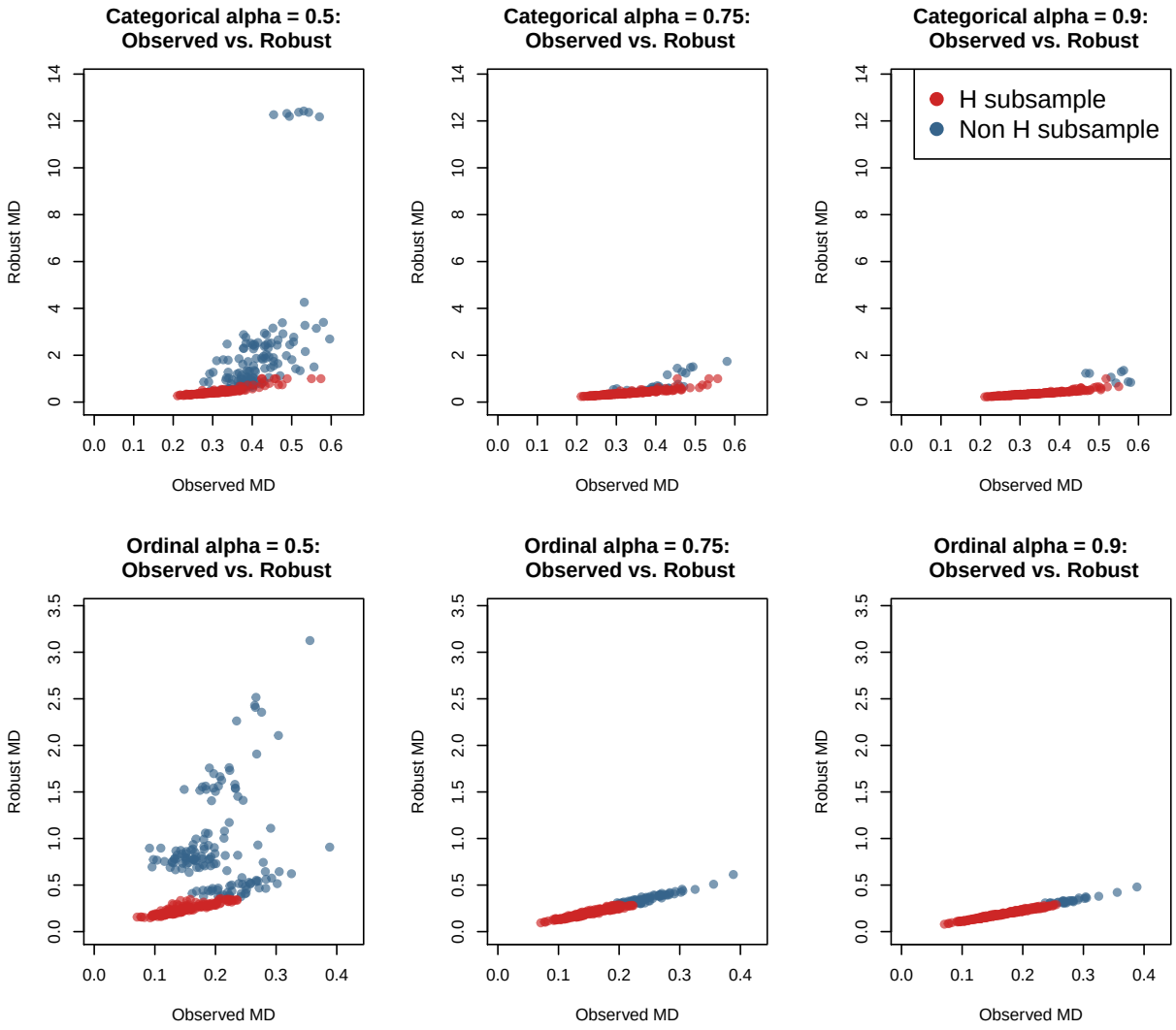

Figure 2. Generalized MCD applied to the miniature survey of autobiographical memory (MSAM) data from ONDRI. Left to right is the alpha level and top to bottom is how the data were coded. All plots show observed vs. robust Mahalanobis distances (MD). Observations in red are the H subsample, observations in blue are not in the H subsample. Axes are scaled to the same values within each recoding to highlight the changes in robust MD.

Table 3

Generalized MCD subsample sizes.

	C .5	C .75	C .9	O .5	O .75	O .9
C .5	175					
C .75	165	237				
C .9	168	228	275			
O .5	127	149	155	160		
O .75	159	205	223	150	230	
O .9	172	229	262	160	230	272

Note. C = Categorical, O = Ordinal. Only the lower triangle and diagonal are shown. The diagonal indicates the size of the H subset out of N = 300. Off-diagonal values indicate the overlap in subsets between each version of the miniature survey on autobiographical memory (MSAM) analysis. As the proportion for subsample sizes increase, overlap becomes more likely.

ADNI Data

Data used in the preparation of this article come from Phase 1 of the ADNI database (adni.loni.usc.edu). ADNI was launched in 2003 as a public-private funding partnership and includes public funding by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and the Food and Drug Administration. The primary goal of ADNI has been to test a wide variety of measures to assess the progression of mild cognitive impairment and early Alzheimer's disease. The ADNI project is the result of efforts of many coinvestigators from a broad range of academic institutions and private

corporations. Michael W. Weiner (VA Medical Center, and University of California-San Francisco) is the ADNI Principal Investigator. Subjects have been recruited from over 50 sites across the United States and Canada (for up-to-date information, see www.adni-info.org).

We used three data sets from ADNI: SNPs (categorical data), the clinical dementia rating (CDR; as ordinal data), and volumetric estimates from five brain regions (continuous data). The ADNI sample for use here was $N = 613$ across three diagnostic groups: Alzheimer's disease (AD) = 264, Mild cognitive impairment (MCI) = 166, and a control group (CON) = 183. These data can be obtained in ADNI via the genome-wide data and the `ADNIMERGE R` package.

We used SNPs associated with APOE and TOMM40, which are strong genetic contributions to AD (Roses et al., 2010). SNPs were excluded if their minor allele frequency was $< 5\%$. Participants or SNPs with $> 5\%$ missingness were excluded. Genotypes with $< 5\%$ were recoded; generally the minor homozygote ("aa") had low frequency therefore in those cases we coded for the presence or absence of "a" (i.e., "AA" vs. presence of "a"; which is the dominant inheritance model). We were thus left with 13 SNPs that spanned 35 columns: 4 of our SNPs were recoded to combine "Aa" and "aa" (i.e., "AA" vs. "Aa+aa") and the remaining 9 SNPs each had three levels (i.e., "AA", "Aa", and "aa").

The CDR is a structured interview that has six domains: memory, orientation, judgement, communication, home and hobbies, and self-care (Morris (1993); see also http://alzheimer.wustl.edu/cdr/cdr.htm). After the interview ratings are applied to each category with the following possible responses: 0 (normal), 0.5 (very mild), 1 (mild), 2 (moderate), and 3 (severe). The CDR has ordinal responses and therefore we recoded the CDR with fuzzy coding.

We also included five brain volumetric estimates known to be impacted by AD:

ventricles, hippocampus, enthorinal cortex, fusiform gyrus, and medial temporal regions. The volumetric estimates here are treated as continuous data. We used a "data doubling" approach for continuous data so that it too behaves like disjunctive data. We refer to the transformation of continuous to pseudo-disjunctive data specifically as the "Escofier transform" (Beaton et al., 2016; Escofier, 1979). The Escofier transform has a lower and upper value like fuzzy coding: "$-$" $= \frac{1-x}{2}$ and "$+$" $= \frac{1+x}{2}$, where x is typically a normalized form of a continuous value (e.g., a Z-score). An example of "doubling" for continuous data can be seen in Table 4.

After preprocessing we had a matrix with $I = 613 \times J = 57$ (13 SNPs categorical that spanned 35 columns, 6 CDR domains that spanned 12 columns, and 5 brain regions that spanned 10 columns). In this example, we set $\alpha = 0.5$. Figure 3 shows our GMCD applied to the mixed data set, where individuals in red are those included in the H subsample and individuals in blue are those excluded from the H subsample. Figure 3a shows the observed squared MD and χ^2-distances for individuals (in red) that comprise the minimum determinant. Figure 3b shows the observed vs. robust MD. Figure 3c shows the log transformed observed vs. robust MD to help visualize the entire set of individuals. In Figure 3c we also include a horizontal line to help denote which individuals were susceptible to the "masking" effect.

While MCD algorithms are useful to identify outlying individuals in a multivariate space, these algorithms do not indicate why individuals are outliers. To help understand why these individuals are outliers we used a heatmap of the $\widetilde{\mathbf{Z}}_{\mathbf{X}}$ matrix (Figure 4). The rows in Figure 4 are the observations where the columns are the variables. For the columns, we excluded the "lower pole" from the data-doubling procedure of the ordinal and continuous values as these are redundant in the visualization.

The scale in Figure 4 comes from the maximum absolute value from the $\widetilde{\mathbf{Z}}_{\mathbf{X}}$ matrix. For the columns, we indicate which data type each column belongs to with color in the

Table 4

Example of Escofier transform

(a) *Volumetric brain regions continuous (Z-scores) example*

	Hippocampus	Fusiform
001_S_4321	0.34	-0.46
006_S_1234	-0.39	1.80
...	...	...
007_S_0880	-1.02	0.92
004_S_8866	-0.10	0.26

(b) *Escofier-style doubling for continuous data*

	Hippocampus−	Hippocampus+	Fusiform−	Fusiform+
001_S_4321	$\frac{1-0.34}{2} = 0.33$	$\frac{1+0.34}{2} = 0.67$	$\frac{1-(-0.46)}{2} = 0.73$	$\frac{1+(-0.46)}{2} = 0.27$
006_S_1234	$\frac{1-(-0.39)}{2} = 0.70$	$\frac{1+(-0.39)}{2} = 0.30$	$\frac{1-1.80}{2} = -0.40$	$\frac{1+1.80}{2} = -1.40$
...	...	...	...	...
007_S_0880	$\frac{1-(-1.02)}{2} = 1.01$	$\frac{1+(-1.02)}{2} = -0.01$	$\frac{1-0.92}{2} = 0.04$	$\frac{1+0.92}{2} = 0.96$
004_S_8866	$\frac{1-(-0.1)}{2} = 0.55$	$\frac{1+(-0.1)}{2} = 0.45$	$\frac{1-0.26}{2} = 0.37$	$\frac{1+0.26}{2} = 0.63$

Note. Example of data doubling via the Escofier transform for continous data with the volumetric estimates for (structural) brain imaging as an example. Note that the sum of each row is equal to the total number of original variables, and that the sum within each pair of columns per variable is equal to the total number of rows.

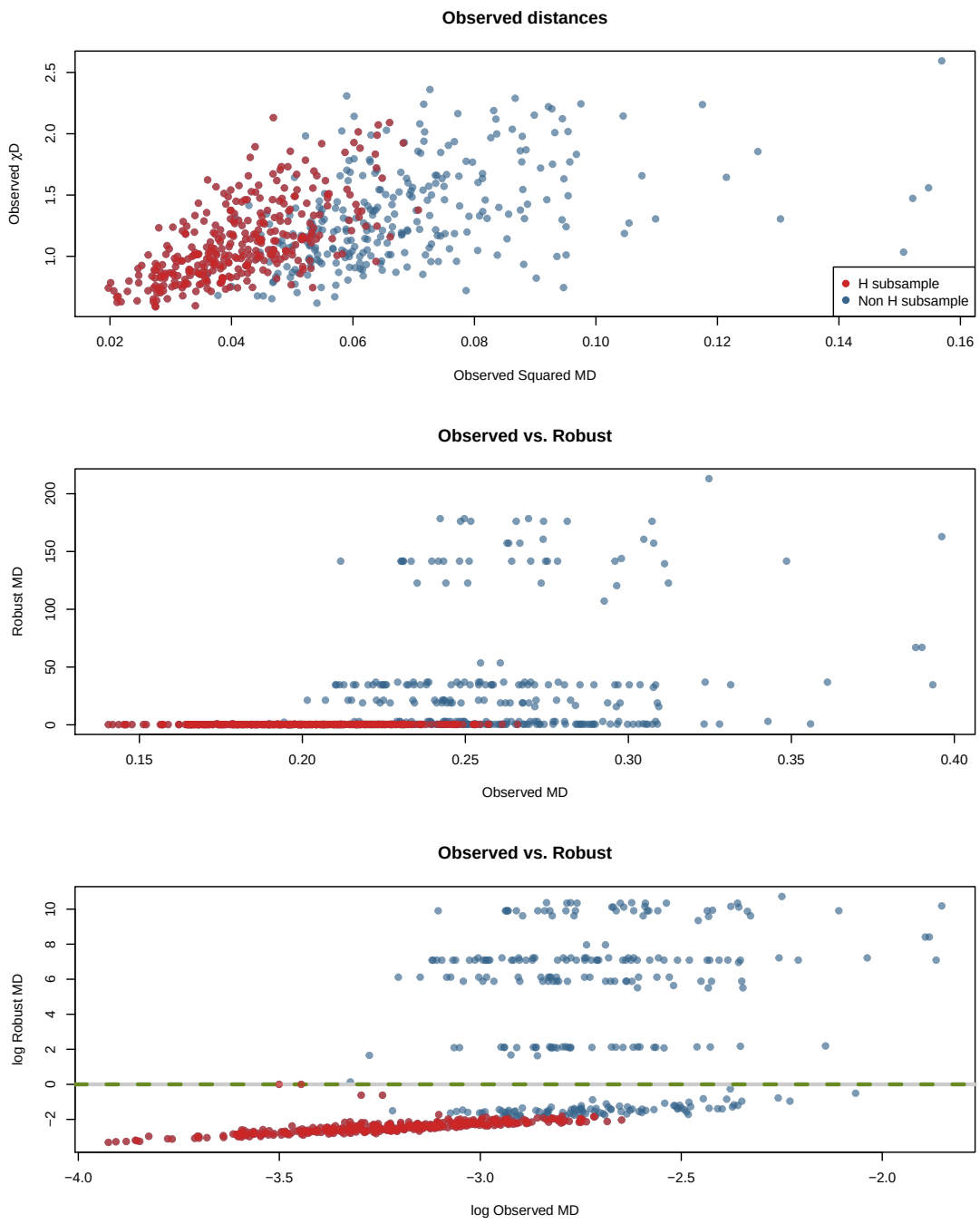

Figure 3. Application of generalized MCD to mixed data. Data are comprised of SNPs (categorical), clinical dementia rating (ordinal) and volumetric brain estimates (continuous) from ADNI. (a) shows the observed distances with the H subset colored in red, where (b) shows the observed vs. robust Mahalanobis distances (MDs), and (c) shows the same but with log transformed observed and robust distances. Panel (c) includes a horizontal line to highlight individuals susceptible to the 'masking' effect.

legend—brain volumes (continuous), clinical dementia rating (ordinal), and SNPs (categorical). For the rows, we provide three separate labels: (1) a continuous label that is the (natural) log transform of the robust MDs, (2) whether the row (observation) is above or below the horizontal line in Figure 3c, and (3) the diagnostic label of each participant where CON = control, MCI = mild cognitive impairment, and AD = Alzheimer's disease. We generally refer to the individuals "above" the line as outliers in this case.

The outliers in Figure 4 do not show a homogeneous pattern. That is, there exist multiple ways in which an individual could be an outlier. Generally, when compared to the "inliers" (those below the line) there exists clear patterns of outlierness. First, a particular genotype—the GG genotype in SNP rs387976—exists exclusively within the outlier set. However, not all individuals within the outlier set have the GG genotype in SNP rs387976. In the outlier set, we can see that there also exists patterns of rare genotypes. For example, quite a few individuals have both the the GG genotype for rs157580 and TT for rs439401. But there also exists additional patterns in the outlier set that are unexpected. Generally, as clinical symptoms of AD present, we also expect that ventricle volumes increase, while the other brain volumes decrease. Instead, we see individuals that have those brain patterns with low CDR ratings (i.e., no clinical indiciation of AD). We also see the opposite: higher CDR ratings with brain volume patterns that do not reflect atrophy or degeneration. In some cases, we see only specific high (or low) values for the CDR and brain volumes, which indicate specific unexpected values as opposed to unexpected patterns. Furthermore, some individuals in the outlier set express unique patterns across all measures, e.g., individuals with infrequent genotypes and unexpected high (or low) CDR scores with low (or high) brain volumes. The diagnostic label provides critical information. The outlier set is not comprised of any single diagnosis. Rather, the outlier set is diagnostically heterogeneous, which means that these individuals are atypical with respect to the entire sample and especially for their respective diagnostic groups.

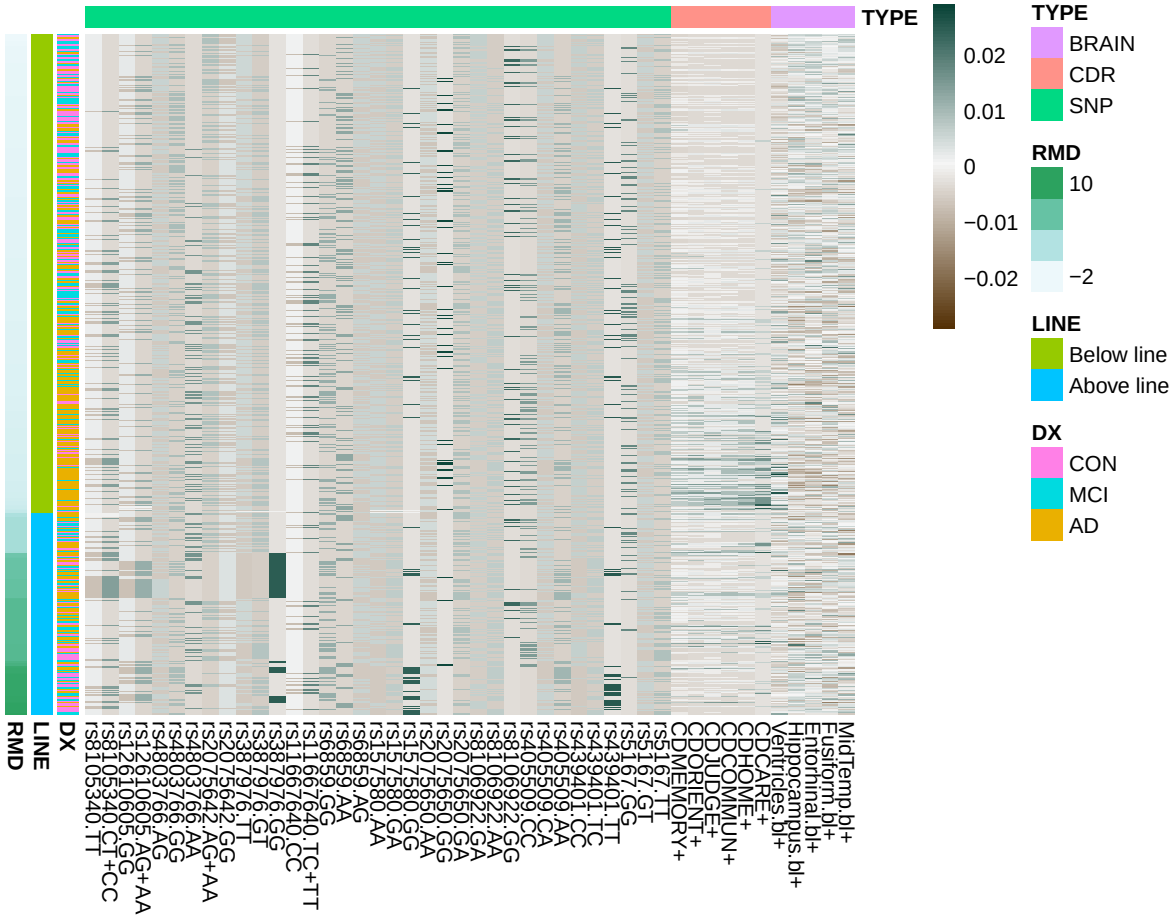

Figure 4. A heatmap of the $\widetilde{\mathbf{Z}}_{\mathbf{X}}$ for the ADNI data. The scale is based on the maximum absolute value in $\widetilde{\mathbf{Z}}_{\mathbf{X}}$. TYPE denotes which type of data each were in each column. To note, here we only show the 'upper pole' of the clinical dementia rating (CDR; ordinal) and brain volume (continuous) data. Rows of the heatmap are in descending order of robust Mahalanobis distance. The rows have three labels: (1) RMD which is the log transform of the robust Mahalanobis distance (as seen in the previous figure), (2) LINE which denotes which rows are above or below the horizontal line in the previous figure, and (3) DX which is the diagnosis of each participant.

Discussion

We presented a MCD-based approach that generalizes to virtually any data type (i.e., categorical, ordinal, continuous). The basis of our approach relies on Mahalnobis distances from CA via the SVD. Our definition of MDs avoids the issues—such as excessive weights and arbitrary metrics—of Mahalanobis-like distances that Goodall (1966) noted. And because our generalized MCD relies on CA, this allows for various types or mixtures of data through simple recoding approaches (see Tables 2 and 4). Our generalized MCD was designed around the necessity to identify robust structures and outliers in complex non-quantitative data. And like many MCD-based and -inspired approaches (Boudt et al., 2017; Fritsch, Varoquaux, Thyreau, Poline, & Thirion, 2012; Mejia et al., 2017; Rousseeuw & Van Driessen, 1999), our approach identifies both robust substructures and outliers (often hidden through "masking" effects).

Within the ONDRI study, we realized that many—if not all—existing outlier detection and robust covariance techniques could not accomodate our needs. We required a multivariate outlier technique that could easily work on categorical, ordinal, or mixed data. While there are other possible algorithms (e.g., Candès, Li, Ma, & Wright, 2011; Fan, Liao, & Mincheva, 2013) we chose to extend the MCD because it is a reliable and widely-used technique (Rousseeuw & Van Driessen, 1999 has been cited $2,238$ times at the time we wrote this paper), and available in languages routinely used for multivariate analyses (e.g., https://cran.r-project.org/web/packages/robustbase/index.html, https://cran.r-project.org/web/packages/rrcov/index.html, and https://wis.kuleuven.be/stat/robust/LIBRA). We have made our GMCD approach available in an R package designed to detect outliers and robust structures (see https://github.com/derekbeaton/ours).

Future directions and applications of GMCD

In our ADNI example we treated each variable as an independent measure. Though routine, that is somewhat unrealistic: each individual variable comes from larger sets of related variables (e.g., SNPs, the CDR, and brain imaging). Multiple Factor Analysis (MFA; Escofier & Pages, 1994; Abdi, Williams, & Valentin, 2013) is a technique designed to accomodate *a priori* groupings of variables (columns) with respect to a global analysis. Notably, MFA has been extended to mixtures of data types (Bécue-Bertaut & Pagès, 2008) based on the work by Escofier (1979) and Escofier and Pages (1994) that we use here. Thus, our GMCD approach could be applied for MFA-type problems.

Because our generalized MCD is based on GCA, as defined by Escofier (1983) and Escofier (1984), we do not depend on the $\bar{H}$ sample set to 0s in the iterative process. The use of GCA allows for any reasonable definition of weights applied to the rows or columns, as well as defining alternate expected values. For example, we could use population level estimates of the allele frequencies (from, e.g., dbGaP, HapMap or 1000 Genomes), which would change our column weights and expected matrix. Likewise, we could also use alternate row weights. In standard MCA—and here in our GMCD—all observations are assigned equal weights (i.e., $\frac{1}{I}$). But in standard CA, rows do not necessarily have equal weights. Rather the row weights are proportional to all elements in that row akin to the column weights (*cf.* Eq. (6)). Though these variations are possible, we have not explored them here and consider these as open questions.

As previously noted, in MCA we discard null dimensions (i.e., those with effectively 0 eigenvalues). However, in MCA it is also typical to discard many of the low-variance components with eigenvalues (Abdi & Valentin, 2007). In MCA, the reason to discard these low-variance components is because of a correction to the dimensionality of the space, which should reflect the C original variables as opposed to the J analyzed columns. The "Benzécri correction" is the most common approach to identify which low-variance components to

discard (Benzécri, 1979), which are components with eigenvalues smaller than $1/C$. The Benzécri correction suggests a way to retain a lower dimensional space, and also means we can compute Mahalanobis-like distances from some arbitrary set of high-variance components. This would be equivalent to a rudimentary regularization. However in our GMCD approach, we do not discard components via Benzécri correction, rather we retain all non-null dimensions for analyses.

Limitations

While our approach opens many avenues to extend the MCD there are some limitations. As just noted, the Benzécri correction (Benzécri, 1979) discards low-variance components. First, the Benzécri correction does not work well in practice for use in the GMCD. In most cases outliers typically express high values on low-variance components. If we were to discard low-variance components we might not identify a suitable robust structure nor identify outliers. The low variance components are typically where the masking effect exists. Furthermore, the dimensionality correction (Benzécri, 1979) of the space is only established for strictly categorical—and thus completely disjunctive—data. How to determine, and then use, the correct dimensionality of the space (especially for mixed variables) is an open question. We have opted for the more conservative approach: require full rank data $I > J$ and retain all components with non-zero eigenvalues.

As noted in our concluding points in our formulation of the GMCD, the use of multiple columns to represent each variable means there exists linear dependency between the columns within any given variable. By definition, the determinant of such a matrix is 0. Our GMCD solution relaxes the requirement of a determinant and makes use of all non-null dimensions from MCA, and thus is a pseudo-determinant. Likewise, our MD is also defined only from those non-null dimensions.

Finally, our algorithm does not compute a robust center from the subsample. To do so

would likely drop levels from variables (effectively zeroing out columns). When discarding specific columns, the MDs of individuals with those levels *decreases*. This decrease is incorrect, because often those levels are rare, and thus the individuals should have a high MD. Rather, our robust center exists within the χ^2-space as defined by the factor scores. This is because subset MCA "maintains the geometry [...] and χ^2-distances of the complete MCA [...]" (Greenacre & Pardo, 2006). For example if we were to apply our technique *as is* with continuous data (recoded as in Table 4) and compare against the standard MCD we would obtain slightly different results. However, at each step if we were to re-center the continuous data before applying the Escofier transform, we would produce exactly the same results as the standard MCD.

Note that we also do not provide any discussion on thresholds to identify outliers. In most applications of the MCD, a cutoff values is defined as some percentage (e.g., 97.5%) and relies on various distributions such as χ^2 (Hubert et al., 2017) or the F-distribution (Hardin & Rocke, 2005). Rather, we encourage the use of resampling methods and to use empirical cutoffs from the data set. This could be done with distributions of robust distances generated from, for examples, bootstrap or repeated subsampling techniques.

Conclusions

The MCD is a reliable and widely-used approach to identify robust multivariate structures and outliers. Until now, the MCD could only be used for data that are (or were assumed to be) continuous. Our approach uses CA to define Mahalanobis distances via the singular vectors and allows us to apply an MCD approach to virtually any data type. We believe that our generalized MCD via GCA opens many new avenues to develop multivariate robust and outlier detection approaches for the complex data sets we face today. Not only does GCA provide the basis for a new family of MCD techniques but our technique also suggests ways to approach other robust techniques (Candès et al., 2011; Fan et al., 2013). With studies such as ONDRI and ADNI, the boundaries of what constitutes a "data set" are

fluid and now inherently heterogeneous: researchers analyze multiple variables of different types from various sources. Our examples highlight these types of routine data sets. With complex studies that contain many mixed-variables, techniques like our generalized MCD will be essential to maintain data quality and integrity.

Acknowledgements

Funding. 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 SS are partially supported by a Canadian Institutes of Health Research grant (MOP 201403).

ONDRI. This work was completed on behalf of the Ontario Neurodegenerative Disease Research Initiative (ONDRI). The authors would like to acknowledge the ONDRI Founding Authors: Robert Bartha, Sandra E. Black, Michael Borrie, Dale Corbett, Elizabeth Finger, Morris Freedman, Barry Greenberg, David A. Grimes, Robert A. Hegele, Chris Hudson, Anthony E. Lang, Mario Masellis, William E. McIlroy, Paula M. McLaughlin, Manuel Montero-Odasso, David G. Munoz, Douglas P. Munoz, J.B. Orange, Michael J. Strong, Stephen C. Strother, Richard H. Swartz, Sean Symons, Maria Carmela Tartaglia, Angela Troyer, and Lorne Zinman.

ADNI. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd

and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Abdi, H., & Valentin, D. (2007). Multiple correspondence analysis. *Encyclopedia of Measurement and Statistics*, 651–657.

Abdi, H., Williams, L. J., & Valentin, D. (2013). Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(2), 149–179.

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bar-hen, A., & Daudin, J. J. (1995). Generalization of the Mahalanobis Distance in the Mixed Case. *Journal of Multivariate Analysis*, *53*(2), 332–342. https://doi.org/10.1006/jmva.1995.1040

Barkmeijer, J., Bouttier, F., & Van Gijzen, M. (1998). Singular vectors and estimates of the analysis-error covariance metric. *Quarterly Journal of the Royal Meteorological Society*, *124*(549), 1695–1713.

Beaton, D., Dunlop, J., & Abdi, H. (2016). Partial least squares correspondence analysis: A framework to simultaneously analyze behavioral and genetic data. *Psychological Methods*, *21*(4), 621.

Bedrick, E. J., Lapidus, J., & Powell, J. F. (2000). Estimating the Mahalanobis Distance from Mixed Continuous and Discrete Data. *Biometrics*, *56*(2), 394–401. https://doi.org/10.1111/j.0006-341X.2000.00394.x

Benzécri, J. P. (1979). Sur le calcul des taux d?Inertie dans l?Analyse d?Un questionnaire, addendum et erratum à [bin. MULT.]. *Cahiers de L?Analyse Des Données*, *4*(3), 377–378.

Bécue-Bertaut, M., & Pagès, J. (2008). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics & Data Analysis*, *52*(6), 3255–3268.

Boriah, S., Chandola, V., & Kumar, V. (2008). Similarity Measures for Categorical Data: A Comparative Evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining* (pp. 243–254). Society for Industrial; Applied Mathematics.

Boudt, K., Rousseeuw, P., Vanduffel, S., & Verdonck, T. (2017). The Minimum Regularized Covariance Determinant estimator. *arXiv:1701.07086 [Stat]*. Retrieved from http://arxiv.org/abs/1701.07086

Brereton, R. G. (2015). The Mahalanobis distance and its relationship to principal component scores. *Journal of Chemometrics*, *29*(3), 143–145. https://doi.org/10.1002/cem.2692

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

Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en analyse factorielle [simultaneous analysis of qualitative and quantitative variables in factor analysis]. *Les Cahiers de L?Analyse Des Données*, *4*, 137–146.

Escofier, B. (1983). Analyse de la différence entre deux mesures définies sur le produit de deux mêmes ensembles. *Cahiers de L'Analyse Des Données*, *8*(3), 325–329.

Escofier, B. (1984). Analyse factorielle en reférence à un modéle. Application à lanalyse de tableaux dechanges. *Revue de Statistique Appliquée*, *32*(4), 25–36.

Escofier, B., & Pages, J. (1994). Multiple factor analysis (afmult package). *Computational Statistics & Data Analysis*, *18*(1), 121–140.

Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(4), 603–680.

Farhan, S. M., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., ... others. (2017). The ontario neurodegenerative disease research initiative (ondri). *Canadian Journal of Neurological Sciences*, *44*(2), 196–202.

Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.-B., & Thirion, B. (2012). Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. *Medical Image Analysis*, *16*(7), 1359–1370.

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.

Goldfarb, B., & Pardoux, C. M. (2006). Exploring series of multivariate censored temporal data through fuzzy coding and correspondence analysis. *Statistics in Medicine*, *25*(10), 1741–1750.

Goodall, D. W. (1966). A New Similarity Index Based on Probability. *Biometrics*, *22*(4), 882–907. https://doi.org/10.2307/2528080

Grassi, M., & Visentin, S. (1994). Correspondence analysis applied to grouped cohort data. *Statistics in Medicine*, *13*(23-24), 2407–2425.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press. Retrieved from http://books.google.com/books?id=LsPaAAAAMAAJ

Greenacre, M. (2014). Data doubling and fuzzy coding. In J. Blasius & M. Greenacre (Eds.), *Visualization and verbalization of data* (pp. 239–253). Philadelphia, PA, USA: CRC Press.

Greenacre, M. (2017). *Correspondence analysis in practice.* CRC press.

Greenacre, M., & Blasius, J. (2006). *Multiple correspondence analysis and related methods.* CRC press.

Greenacre, M., & Pardo, R. (2006). Subset correspondence analysis: Visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods & Research*, *35*(2), 193–218.

Hadi, A. S., Imon, A., & Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 57–70.

Hardin, J., & Rocke, D. M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, *14*(4), 928–946.

Hendry, G., North, D., Zewotir, T., & Naidoo, R. (2014). The application of subset correspondence analysis to address the problem of missing data in a study on asthma severity in childhood. *Statistics in Medicine*, *33*(22), 3882–3893.

Holmes, S. (2008). Multivariate data analysis: The french way. In *Probability and statistics: Essays in honor of david a. Freedman* (pp. 219–233). Institute of Mathematical Statistics.

Hubert, M., & Debruyne, M. (2010). Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 36–43. https://doi.org/10.1002/wics.61

Hubert, M., Debruyne, M., & Rousseeuw, P. (2017). Minimum Covariance Determinant and Extensions. *arXiv:1709.07045 [Stat]*. Retrieved from http://arxiv.org/abs/1709.07045

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

https://doi.org/10.1198/004017004000000563

Hubert, M., Rousseeuw, P., & Verdonck, T. (2012). A Deterministic Algorithm for Robust Location and Scatter. *Journal of Computational and Graphical Statistics*, *21*(3), 618–637. https://doi.org/10.1080/10618600.2012.672100

Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate descriptive statistical analysis: Correspondence analysis and related techniques for large matrices*. Wiley.

Leeuw, J. de, & Heijden, P. G. van der. (1988). Correspondence analysis of incomplete contingency tables. *Psychometrika*, *53*(2), 223–233.

Leon, A. R. de, & Carrière, K. C. (2005). A generalized Mahalanobis distance for mixed data. *Journal of Multivariate Analysis*, *92*(1), 174–185. https://doi.org/10.1016/j.jmva.2003.08.006

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150–156. https://doi.org/10.1016/j.jesp.2017.09.011

Magnotti, J. F., & Billor, N. (2014). Finding multivariate outliers in fMRI time-series data. *Computers in Biology and Medicine*, *53*(Supplement C), 115–124. https://doi.org/10.1016/j.compbiomed.2014.05.010

McCane, B., & Albert, M. (2008). Distance functions for categorical and mixed variables. *Pattern Recognition Letters*, *29*(7), 986–993. https://doi.org/10.1016/j.patrec.2008.01.021

Mejia, A. F., Nebel, M. B., Eloyan, A., Caffo, B., & Lindquist, M. A. (2017). PCA leverage: Outlier detection for high-dimensional functional magnetic resonance imaging data. *Biostatistics*, kxw050.

Morris, J. C. (1993). The clinical dementia rating (cdr): Current version and scoring rules. *Neurology*.

Palombo, D. J., Williams, L. J., Abdi, H., & Levine, B. (2013). The survey of autobiographical memory (sam): A novel measure of trait mnemonics in everyday life. *Cortex*, *49*(6), 1526–1540.

Roses, A., Lutz, M., Amrine-Madsen, H., Saunders, A., Crenshaw, D., Sundseth, S., ... Reiman, E. (2010). A tomm40 variable-length polymorphism predicts the age of late-onset alzheimer's disease. *The Pharmacogenomics Journal*, *10*(5), 375–384.

Rousseeuw, P., & Van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, *41*(3), 212–223. https://doi.org/10.1080/00401706.1999.10485670

SenGupta, A. (1987). Tests for standardized generalized variances of multivariate normal populations of possibly different dimensions. *Journal of Multivariate Analysis*, *23*(2), 209–219.

Stephenson, D. (1997). Correlation of spatial climate/weather maps and the advantages of using the mahalanobis metric in predictions. *Tellus A*, *49*(5), 513–527.

Sunderland, K. M., Beaton, D., Fraser, J., Kwan, D., McLaughlin, P. M., Montero-Odasso, M., ... others. (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.

Van der Heijden, P. G., De Falguerolles, A., & Leeuw, J. de. (1989). A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *Applied Statistics*, 249–292.

Verity, R., Collins, C., Card, D. C., Schaal, S. M., Wang, L., & Lotterhos, K. E. (2017). Minotaur: A platform for the analysis and visualization of multivariate results from genome scans with R Shiny. *Molecular Ecology Resources*, *17*(1), 33–43. https://doi.org/10.1111/1755-0998.12579

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1-3), 37–52.

Yanai, H., Takeuchi, K., & Takane, Y. (2011). *Projection matrices, generalized inverse matrices, and singular value decomposition*. Springer Science & Business Media.