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

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

[1] Rotman Research Institute

[2] Baycrest Health Sciences

[3] ADNI

[4] London Health Sciences Centre

[5] University of Toronto

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

[7] ONDRI

[8] The University of Texas at Dallas

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 we can 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 accomodate 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 data such as 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

1 Introduction

The minimum covariance determinant (MCD; Hubert & Debruyne, 2010) algorithm is a robust estimator for multivariate location (mean) and scatter (covariance). Given a matrix of observations and variables, MCD approaches, generally, find a subset of individuals that have minimum scatter where scatter is defined as the determinant of the covariance matrix. Robust estimates of mean and covariance are computed from the subset of individuals with minimum scatter (i.e., the smallest determinant from subsample covariance matrix).

In general, MCD techniques work based on two key features in order to detect the likely-smallest cloud of observations: (1) the determinant of a given covariance matrix and (2) the Mahalanobis distances (MDs) of all observations with respect to a that covariance matrix. Because MCD searches for the smallest and most homogeneous cloud of observations, it is also a common technique used for detection of multivariate outliers (Hadi, Imon, & Werner, 2009; Magnotti & Billor, 2014; Verity et al., 2017). Since the introduction of the MCD, there have been many improvements and variations on the technique, such as robust PCA (Hubert, Rousseeuw, & Branden, 2005) and deterministic MCD (Hubert, Rousseeuw, & Verdonck, 2012), where the most commonly used variant is probably the FAST-MCD approach (Rousseeuw & Van Driessen, 1999) See also Hubert, Debruyne, and Rousseeuw (2017) for an overview. Recently the MCD approach has been extended to address high-dimensional data through regularization (Boudt, Rousseeuw, Vanduffel, & Verdonck, 2017).

Though the MCD and its variants are standard techniques to identify both (1) robust structures and (2) outliers, there is one substantial gap: MCD approaches generally work only for data assumed to be continuous. The lack of a non-quantitative MCD is problematic because many multivariate data sets are inherently non-quantitative (e.g., categorical or ordinal) such as clinical ratings scales, questionnaires, and genetics. So how can we apply MCD approaches to non-quantitative data? The major barrier to apply the MCD on

non-quantitative data is the lack of a standard Mahalanobis distance (MD) estimate for non-quantitative data.

Goodall (1966) noted that there are generally two issues with the application of standard computations and rules of MDs 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 MD estimates, nor do we want to arbitrarily apply metric values to categorical or ordinal data. While there are many Mahalanobis-like distances and 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).

Therefore, if we want to develop a MCD algorithm for non-quantitative data, we must first define a MD that neither imposes undue influence nor arbitrarily assigned values. Arguably the most well understood and defined metric for non-quantitative data is the χ^2-distance, typically used for contingency and categorical data (Greenacre, 2017; Greenacre & Hastie, 1987). In our work here we first show that the MCD can be generalized to categorical data by defining MD under the assumptions of χ^2 (i.e., independence) through Correspondence Analysis (CA), which is a singular value decomposition (SVD)-based technique. We then show how our MCD approach generalizes to almost any data type, including mixed sets of variables, because of CA.

Our paper is outlined as follows. In *Notation and software* we provide the notation set used in this paper, the software used in and to write this paper, and a description of the code we provide for our generalized MCD. In *Determinant and Mahalanobis distances via the SVD* we first show how the SVD provides the two key requirements for the MCD algorithm: the determinant of a covariance matrix and MDs (specifically, through principal components

analysis; PCA). Then, we show how these properties of the SVD generalize in order to compute MDs for categorical data through CA. In *Establishing the MCD algorithm for categorical data* we then show how we only require PCA to perform the standard MCD technique but that a categorical version of the MCD requires a specific form of CA (via generalized CA, specifically "subset" CA) in order to to achieve the MCD for categorical data. Next in *Applications and Extensions* we illustrate our newly defined MCD to identify robust structures and outliers. We then show that particular recoding schemes allow for the use of other data types such as ordinal or continuous via our categorical MCD and thus our approach generalizes the MCD to any data type or even mixed data types (i.e., a matrix of categorical, oridinal, and continuous variables). In *Establishing the MCD algorithm for categorical data* we use toy (simulated) genetics data from the supplemental material of Beaton, Dunlop, and Abdi (2016) in order to formalize a categorical version of the MCD. In *Applications and Extensions* we illustrate our generalized MCD with data from the Ontario Neurodegenerative Disease Research Initiative (ONDRI) and the Alzheimer's Disease Neuroimaging Initiative (ADNI). Finally we discuss the technique with respect to the results, as well as provide concluding remarks and future directions for a generalized MCD in *Discussion*.

2 Software and notation

We used R (3.3.1, R Core Team, 2016) and the R-packages *ExPosition* (2.8.19, Beaton, Fatt, & Abdi, 2014), *MASS* (7.3.47, Venables & Ripley, 2002), and *papaja* (0.1.0.9492, Aust & Barth, 2017) for all our analyses and to write the manuscript itself (in RMarkdown with `papaja`). Software, code, and examples available at https://github.com/derekbeaton/ours.

2.1 Notation

Bold uppercase letters denotes 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: $\mathbf{XX}^+\mathbf{X} = \mathbf{X}$, $(\mathbf{X}^+)^+ = \mathbf{X}$, 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. Finally, we reserve some letters to have very specific meanings: (1) $\mathbf{F}$ denotes component (sometimes called factor) scores where, for example, $\mathbf{F}_I$ denote 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 $\mathbf{X}$", and (3) 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 $\mathbf{X}$. Furthermore, when we use a previously defined element as a subscript, for example $\mathbf{Z}_\mathbf{X}$, it denotes that the $\mathbf{Z}$ of $\mathbf{Z}_\mathbf{X}$ was derived from the element in its subscript, for example $\mathbf{Z}_\mathbf{X}$ is a centered and/or normalized version $\mathbf{X}$.

3 Determinant and Mahalanobis distances via the SVD

In this section we show that we only need principal components analysis (PCA) more specifically the SVD to compute both the determinant and MD. Through the SVD we can compute MD by way of "leverage" through the singular vectors. Thus the SVD provides the two key features we require for the MCD. Next we show how these ideas extend to categorical data via multiple correspondence analysis (MCA) and therefore we establish the fundamental

features (MDs and the determinant) required for the MCD applied to categorical data.

3.1 Overview of MCD algorithm

Say we have a matrix $\mathbf{X}$ with I rows and J columns where $J < I$ and we assume that $\mathbf{X}$ is full rank. The MCD requires an initialization of subset of size H which is a random subsample from $\mathbf{X}$ where $\lfloor (I + J + 1)/2 \rfloor \leq H \leq I$. In general, the MCD algorithm can be described as:

1. From $\mathbf{X}_H$ compute the column-wise mean as $\boldsymbol{\mu}_H$ and compute the covariance as $\mathbf{S}_H = (\mathbf{X}_H^T \mathbf{X}_H) \times (H-1)^{-1}$ where $\mathbf{X}_H$ has been centered (i.e., subtract the column-wise mean).
2. Compute the determinant of $\mathbf{S}_H$.
3. Compute the squared MD for each observation in $\mathbf{X}$ as $m_i = (x_i - \boldsymbol{\mu}_H)\mathbf{S}_H^{-1}(x_i - \boldsymbol{\mu}_H)^T$.
4. Set $\mathbf{X}_H$ as the subset of H observations with the smallest MDs computed from Step 3.

The size of H is controlled by α to determine the subsample size; α exists between .5 and 1. These steps are repeated until we find a minimum determinant either through optimal search or limited to a given number of iterations. We denote the H subset with the minimum determinant as Ω. For the Ω subset, we can obtain robust mean vector $\boldsymbol{\mu}_\Omega$, robust covariance $\mathbf{S}_\Omega = (\mathbf{X}_\Omega^T \mathbf{X}_\Omega) \times (\Omega - 1)^{-1}$ (where $\mathbf{X}_\Omega$ is column-wise centered), and a set of robust square MDs for each i observation as $m_i = (x_i - \boldsymbol{\mu}_\Omega)\mathbf{S}_\Omega^{-1}(x_i - \boldsymbol{\mu}_\Omega)$. The two most important features for MCD algorithms are: (1) the determinant of the subsample covariance matrix, which helps identify the smallest scatter of data and (2) the squared MDs for all observations derived from the subsample covariance matrix, which helps us find the minimum determinant.

Both the determinant of a matrix and the squared MDs for observations can be computed from the eigen- or singular value decompositions (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 next section we show how squared MD can be computed from the SVD in order to show how the singular vectors provide a generalized estimate of squared MD.

3.2 Mahalanobis distance(s) for continuous data

Let $\mathbf{X}$ be a centered data matrix with I rows and J columns. We will assume that $\mathbf{X}$ is full rank on its columns where generally $I > J$. First we define the sample covariance matrix of $\mathbf{X}$, per usual, as $\mathbf{S} = (\mathbf{X}^T\mathbf{X}) \times (I-1)^{-1}$. The squared MD is defined as:

$$\mathbf{m} = (\mathbf{X}\mathbf{S}^{-1} \odot \mathbf{X})\mathbf{1} \tag{1}$$

where $\mathbf{1}$ is a conformable $J \times 1$ vector of 1s and $\mathbf{m}$ is a $I \times 1$ column vector (where $\mathbf{m}^T$ is a row vector) of squared MDs. We can reformulate Eq. (1) in two ways. The first reformulation is:

$$\mathbf{M} = \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^T, \tag{2}$$

where $\mathbf{m} = \text{diag}\{\mathbf{M}\}$. If we relax the definition of squared MD to exclude the degrees of freedom scaling factor, the second reformulation is:

$$\mathbf{M} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \times (I-1), \tag{3}$$

where $\mathbf{m}' = \text{diag}\{\mathbf{M}\} \times (I-1)$. Henceforth when we refer to squared MD it is the definition we provide in Eq. (3), specifically, $\mathbf{m}' = \text{diag}\{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}$.

3.3 Principal components analysis

Let us now define the principal components analysis (PCA) of $\mathbf{X}$ which has I rows and J columns, where $I > J$. If $\mathbf{X}$ is only column-wise centered then it is a "covariance" PCA. If $\mathbf{X}$ is column-wise centered and scaled (e.g., z-scores or sums of squares equal to 1) it is a "correlation" PCA. For the following formulation we assume only a column-wise centered $\mathbf{X}$.

Assume that $\mathbf{X}$ is full rank, where L is the rank of $\mathbf{X}$ and $L \leq J < I$. We can perform the PCA of $\mathbf{X}$ through the singular value decomposition as:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{4}$$

where the following properties hold:

1. $\mathbf{U}$ and $\mathbf{V}$ are orthonormal left and right singular vectors of sizes $I \times L$ and $J \times L$, respectively with $\mathbf{U}^T\mathbf{U} = \mathbf{I} = \mathbf{V}^T\mathbf{V}$. Furthermore, $\mathbf{U}^T = \mathbf{U}^+$ and $\mathbf{V}^T = \mathbf{V}^+$ and when $\mathbf{X}$ is full rank $\mathbf{V}^T = \mathbf{V}^{-1}$.

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

In PCA there are two sets of component scores, one set for the rows and one set for the columns defined as $\mathbf{F}_I = \mathbf{U}\mathbf{\Sigma}$ and $\mathbf{F}_J = \mathbf{V}\mathbf{\Sigma}$, respectively. The component scores are the singular vectors of the rows or columns scaled by the singular values. To note the column component scores are the loadings multiplied by the singular values. The row and column scores can also be defined through projection (or rotation) as:

$$\begin{aligned} \mathbf{F}_I &= \mathbf{U}\mathbf{\Sigma} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V} = \mathbf{X}\mathbf{V} \\ \mathbf{F}_J &= \mathbf{V}\mathbf{\Sigma} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U} = \mathbf{X}^T\mathbf{U}. \end{aligned} \tag{5}$$

In the regression and PCA literature there exists an influence measure called "leverage" which is bounded between 0 and 1 (Wold, Esbensen, & Geladi, 1987). Leverage is extracted from an $I \times I$ matrix called a "hat matrix" in the regression literature and is more generally called an "orthogonal projection matrix" in the multivariate literature (Yanai, Takeuchi, & Takane, 2011). The hat or projection matrix is defined generally as $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ where leverage is defined as $\text{diag}\{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}$. Note that the definition of leverage is equivalent to our definition of a squared MD in Eq. (3).

With respect to PCA there are two types of leverage: one for the columns (typically variables) and one for the rows (typically observations) of a matrix. Here we only discuss leverage for the rows (observations). Both Wold et al. (1987) and Mejia, Nebel, Eloyan, Caffo, and Lindquist (2017) described leverage for the observations (rows) as based on the component scores $\mathbf{F}_I$:

$$\mathbf{G}_I = \mathbf{F}_I(\mathbf{F}_I^T\mathbf{F}_I)^{-1}\mathbf{F}_I^T \tag{6}$$

where each row item's leverage is an element in $\mathbf{g}_I = \text{diag}\{\mathbf{G}_I\}$. If we expand Eq. (6) and substitute $\mathbf{U\Sigma}$ for $\mathbf{F}_I$:

$$\mathbf{G}_I = \mathbf{U\Sigma}(\mathbf{\Sigma U}^T\mathbf{U\Sigma})^{-1}\mathbf{\Sigma U}^T = \mathbf{U\Sigma}(\mathbf{\Sigma I\Sigma})^{-1}\mathbf{\Sigma U}^T = \mathbf{U\Sigma\Lambda}^{-1}\mathbf{\Sigma U}^T = \mathbf{UIU}^T = \mathbf{UU}^T, \tag{7}$$

which indicates that $\mathbf{g}_I$ can be computed as:

$$\mathbf{g}_I = \text{diag}\{\mathbf{UU}^T\}. \tag{8}$$

With the definition of squared MD from Eq. (3) it is obvious that leverage and squared MD are the same: $\mathbf{m}' = \text{diag}\{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\} = \text{diag}\{\mathbf{UU}^T\} = \mathbf{g}_I$. However we show this relationship in full through the SVD as follows:

$$\begin{aligned}
\mathbf{M}' = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T &= \\
\mathbf{U\Sigma V}^T(\mathbf{V\Sigma U}^T\mathbf{U\Sigma V}^T)^{-1}\mathbf{V\Sigma U}^T &= \\
\mathbf{U\Sigma V}^T(\mathbf{V\Sigma I\Sigma V}^T)^{-1}\mathbf{V\Sigma U}^T &= \\
\mathbf{U\Sigma V}^T(\mathbf{V\Lambda V}^T)^{-1}\mathbf{V\Sigma U}^T &= \\
\mathbf{U\Sigma V}^T\mathbf{V\Lambda}^{-1}\mathbf{V}^T\mathbf{V\Sigma U}^T &= \\
\mathbf{U\Sigma I\Lambda}^{-1}\mathbf{I\Sigma U}^T &= \\
\mathbf{U\Sigma}\Lambda^{-1}\mathbf{\Sigma U}^T &= \\
\mathbf{UIU}^T = \mathbf{UU}^T = \mathbf{G}_I,
\end{aligned} \tag{9}$$

because $\mathbf{V}^T = \mathbf{V}^{-1}$ then $(\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T)^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^T$. The relationship established in Eq. (9) is critical to our formulation of MD for categorical data and thus a generalized MCD.

3.4 Representation of categorical data

Say we have an $I \times C$ matrix called $\mathbf{C}$, as in Table 1a, where the rows are observations and the columns are categorical variables typically assumed to be unordered thus each cell contains the nominal level of each variable. These data can be represented in disjunctive coding where each variable is represented by a vector of length N_c, where N_c is the number of levels for the cth variable, as in Table 1b. The disjunctive form of $\mathbf{C}$ is an $I \times J$ matrix $\mathbf{X}$. The multivariate analysis of a disjunctive matrix is usually performed with a technique called multiple correspondence analysis.

3.5 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) for two-way contingency tables to multiple variables, that is: N-way contingency tables (Abdi & Valentin, 2007; Greenacre, 1984; Greenacre & Blasius, 2006; Lebart, Morineau, & Warwick, 1984). MCA like PCA and CA is an SVD-based technique but with specific preprocessing, which requires constraints (weights) for the rows and columns and decomposes a matrix under the assumption of independence (i.e., χ^2). The general premise of Pearson's X^2 is to estimate how far *observed* values deviate from *expected* values. To perform MCA, we first compute the row and column weights:

$$\mathbf{w}_I = (\mathbf{1}^T\mathbf{X}\mathbf{1})^{-1} \times \mathbf{X}\mathbf{1} \text{ and } \mathbf{w}_J = (\mathbf{1}^T\mathbf{X}\mathbf{1})^{-1} \times \mathbf{X}^T\mathbf{1}, \tag{10}$$

where $\mathbf{w}_I$ and $\mathbf{w}_J$ are the row and column marginal sums, respectively, divided by the total sum. In the case of purely disjunctive data, each value in $\mathbf{w}_I$ is identical as each row contains the same number of 1s (see Table 1b): Each element in $\mathbf{w}_I$ is simply the total

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	*NA*	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	.25	.5	.25	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 one observation has missing data (NA). Missing data can be coded as barycentric (which is the mean of the variable).

number of columns in $\mathbf{C}$ divided by the sum of all elements of $\mathbf{X}$. Similarly, $\mathbf{w}_J$ contains the column sums of $\mathbf{X}$ divided by the total sum. Next, just as in Pearson's X^2 analyses, we compute observed ($\mathbf{O_X}$) and expected ($\mathbf{E_X}$) matrices of $\mathbf{X}$ as follows:

$$\mathbf{O_X} = (\mathbf{1}^T\mathbf{X}\mathbf{1})^{-1} \times \mathbf{X} \text{ and } \mathbf{E_X} = \mathbf{w}_I\mathbf{w}_J^T, \tag{11}$$

and deviations computed as

$$\mathbf{Z_X} = \mathbf{O_X} - \mathbf{E_X}. \tag{12}$$

MCA is performed with the generalized SVD (GSVD) of $\mathbf{Z_X}$:

$$\begin{aligned} \mathbf{Z_X} &= \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^T, \text{where} \\ \mathbf{P}^T\mathbf{W}_I\mathbf{P} &= \mathbf{I} = \mathbf{Q}^T\mathbf{W}_J\mathbf{Q}, \end{aligned} \tag{13}$$

and where $\mathbf{W}_I = \text{diag}\{\mathbf{w}_I^{-1}\}$ and $\mathbf{W}_J = \text{diag}\{\mathbf{w}_J^{-1}\}$. We obtain the results of the GSVD of $\mathbf{Z_X}$ through the SVD of $\widetilde{\mathbf{Z}}_\mathbf{X} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{Z_X}\mathbf{W}_J^{\frac{1}{2}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ as in Eq. (4) thus $\mathbf{U}^T\mathbf{U} = \mathbf{I} = \mathbf{V}^T\mathbf{V}$, $\mathbf{U}^T = \mathbf{U}^+$, and $\mathbf{V}^T = \mathbf{V}^+$.

We can see the relationship between $\mathbf{Z_X}$ and $\widetilde{\mathbf{Z}}_\mathbf{X}$ via substitution:

$$\begin{aligned} \mathbf{Z_X} = \mathbf{W}_I^{-\frac{1}{2}}\widetilde{\mathbf{Z}}_\mathbf{X}\mathbf{W}_J^{-\frac{1}{2}} &= \\ \mathbf{W}_I^{-\frac{1}{2}}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_J^{-\frac{1}{2}} &= \\ (\mathbf{W}_I^{-\frac{1}{2}}\mathbf{U})\mathbf{\Sigma}(\mathbf{V}^T\mathbf{W}_J^{-\frac{1}{2}}) &= \\ (\mathbf{W}_I^{-\frac{1}{2}}\mathbf{U})\mathbf{\Sigma}(\mathbf{W}_J^{-\frac{1}{2}}\mathbf{V})^\mathbf{T} &= \\ \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^T. \end{aligned} \tag{14}$$

The *generalized* singular vectors are computed as $\mathbf{P} = \mathbf{W}_I^{-\frac{1}{2}}\mathbf{U}$ and $\mathbf{Q} = \mathbf{W}_J^{-\frac{1}{2}}\mathbf{V}$. For simplicity, we will henceforth refer to the GSVD decomposition where needed in triplet notation, e.g., GSVD($\mathbf{W}_I, \mathbf{Z_X}, \mathbf{W}_J$). 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.

MCA component scores are computed as $\mathbf{F}_I = \mathbf{W}_I\mathbf{P}\boldsymbol{\Sigma}$ and $\mathbf{F}_J = \mathbf{W}_J\mathbf{Q}\boldsymbol{\Sigma}$. Alternatively, much like in PCA, we can define the MCA component scores through projections or rotations (see Eq. (5)). To define MCA component scores through projections, it is easier to work with what are called "profile" matrices where $\boldsymbol{\Phi}_I = \mathbf{W}_I\mathbf{O}_\mathbf{X}$ are the row profiles and $\boldsymbol{\Phi}_J = \mathbf{W}_J\mathbf{O}_\mathbf{X}^T$ are the column profiles. Both the row and column profile matrices can be thought of as each element in $\mathbf{X}$ divided by its respective row or column sum (for the row and column profile matrices, respectively). We can define the MCA row component scores via projection as:

$$\begin{aligned}
\mathbf{F}_I = \boldsymbol{\Phi}_I\mathbf{F}_J\boldsymbol{\Sigma}^{-1} &= \\
\boldsymbol{\Phi}_I\mathbf{W}_J\mathbf{Q}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} &= \\
\boldsymbol{\Phi}_I\mathbf{W}_J\mathbf{Q}\mathbf{I} &= \\
\boldsymbol{\Phi}_I\mathbf{W}_J\mathbf{Q} &= \\
\boldsymbol{\Phi}_I\mathbf{W}_J\mathbf{W}_J^{-\frac{1}{2}}\mathbf{V} &= \\
\boldsymbol{\Phi}_I\mathbf{W}_J^{\frac{1}{2}}\mathbf{V} &= \\
\mathbf{W}_I\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V}, &
\end{aligned} \tag{15}$$

and similarly the column component scores are computed as $\mathbf{F}_J = \boldsymbol{\Phi}_J\mathbf{F}_I\boldsymbol{\Sigma}^{-1} = \mathbf{W}_J\mathbf{O}_\mathbf{X}^T\mathbf{W}_I^{\frac{1}{2}}\mathbf{U}$.

3.6 Mahalanobis distance(s) for categorical data

In *Principal Components Analysis* we showed for continuous data the equivalence between leverage and squared MD (via the hat matrix). Both leverage and squared MD are obtained as the diagonal of the crossproduct of left singular vectors i.e., $\text{diag}\{\mathbf{U}\mathbf{U}^T\}$ or in a computationally simpler way: the row sums of the squared singular vectors $(\mathbf{U} \times \mathbf{U})\mathbf{1}$, where $\mathbf{1}$ is a conformable vector of 1s. Because leverage via the singular vectors is squared MD for continuous data, we use the same premise to define a squared MD for categorical data. In the case of continuous data we can define squared MD for continuous data via PCA and thus

we can define squared MD for categorical data via MCA.

As noted in Eqs. (13) and (14) MCA is performed as GSVD($\mathbf{W}_I, \mathbf{Z}_\mathbf{X}, \mathbf{W}_J$) wherein we apply the SVD as $\widetilde{\mathbf{Z}}_\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Thus in MCA the GSVD via the SVD provides $\mathbf{U}$. Recall the relationship between $\widetilde{\mathbf{Z}}_\mathbf{X}$ and $\mathbf{Z}_\mathbf{X}$: $\widetilde{\mathbf{Z}}_\mathbf{X} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{Z}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}} \Longleftrightarrow \mathbf{Z}_\mathbf{X} = \mathbf{W}_I^{-\frac{1}{2}}\widetilde{\mathbf{Z}}_\mathbf{X}\mathbf{W}_J^{-\frac{1}{2}}$ where $\mathbf{Z}_\mathbf{X} = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^T$ (see Eq. (14)). In the case of MCA, $\widetilde{\mathbf{Z}}_\mathbf{X}$ would be the analog of $\mathbf{X}$ in PCA (see Eq. (4)). Thus in MCA we know that $\mathbf{U}\mathbf{U}^T = \widetilde{\mathbf{Z}}_\mathbf{X}(\widetilde{\mathbf{Z}}_\mathbf{X}^T\widetilde{\mathbf{Z}}_\mathbf{X})^+\widetilde{\mathbf{Z}}_\mathbf{X}^T$. Note that here we replace the standard inversion with the pseudo-inverse because $(\widetilde{\mathbf{Z}}_\mathbf{X}^T\widetilde{\mathbf{Z}}_\mathbf{X})$ is a "group" matrix, where all columns that represent a single variable are collinear. The rank of $(\widetilde{\mathbf{Z}}_\mathbf{X}^T\widetilde{\mathbf{Z}}_\mathbf{X})$ is C, though it spans J columns where $J > C$, thus $(\widetilde{\mathbf{Z}}_\mathbf{X}^T\widetilde{\mathbf{Z}}_\mathbf{X})$ is not invertible. We have shown previously for continuous data that leverage is squared MD. We argue that leverage in MCA i.e., $\widetilde{\mathbf{Z}}_\mathbf{X}(\widetilde{\mathbf{Z}}_\mathbf{X}^T\widetilde{\mathbf{Z}}_\mathbf{X})^+\widetilde{\mathbf{Z}}_\mathbf{X}^T$ is an analog of squared MD for categorical data just as in PCA: $\mathbf{M}'_{\widetilde{\mathbf{Z}}_\mathbf{X}} = \widetilde{\mathbf{Z}}_\mathbf{X}(\widetilde{\mathbf{Z}}_\mathbf{X}^T\widetilde{\mathbf{Z}}_\mathbf{X})^+\widetilde{\mathbf{Z}}_\mathbf{X}^T = \mathbf{U}\mathbf{U}^T$, where $\mathbf{m}'_{\widetilde{\mathbf{Z}}_\mathbf{X}} = \text{diag}\{\mathbf{M}'_{\widetilde{\mathbf{Z}}_\mathbf{X}}\}$. However, here we also show that $\widetilde{\mathbf{Z}}_\mathbf{X}(\widetilde{\mathbf{Z}}_\mathbf{X}^T\widetilde{\mathbf{Z}}_\mathbf{X})^+\widetilde{\mathbf{Z}}_\mathbf{X}^T = \mathbf{Z}_\mathbf{X}(\mathbf{Z}_\mathbf{X}^T\mathbf{Z}_\mathbf{X})^+\mathbf{Z}_\mathbf{X}^T = \mathbf{U}\mathbf{U}^T$.

$$
\begin{aligned}
\mathbf{M}'_{\mathbf{Z_X}} = \mathbf{Z_X}(\mathbf{Z_X}^T\mathbf{Z_X})^+\mathbf{Z_X}^T &= \\
\mathbf{P\Sigma Q}^T\{\mathbf{Q\Sigma P}^T\mathbf{P\Sigma Q}^T\}^+\mathbf{Q\Sigma P}^T &= \\
\mathbf{P\Sigma Q}^T\{\mathbf{Q\Sigma}(\mathbf{UW}_I^{-\frac{1}{2}})^T(\mathbf{W}_I^{-\frac{1}{2}}\mathbf{U})\mathbf{\Sigma Q}^T\}^+\mathbf{Q\Sigma P}^T &= \\
\mathbf{P\Sigma Q}^T\{\mathbf{Q\Sigma W}_I^{-\frac{1}{2}}\mathbf{U}^T\mathbf{UW}_I^{-\frac{1}{2}}\mathbf{\Sigma Q}^T\}^+\mathbf{Q\Sigma P}^T &= \\
\mathbf{P\Sigma Q}^T\{\mathbf{Q\Sigma W}_I^{-\frac{1}{2}}\mathbf{IW}_I^{-\frac{1}{2}}\mathbf{\Sigma Q}^T\}^+\mathbf{Q\Sigma P}^T &= \\
\mathbf{P\Sigma Q}^T\{\mathbf{Q\Sigma W}_I^{-\frac{1}{2}}\mathbf{W}_I^{-\frac{1}{2}}\mathbf{\Sigma Q}^T\}^+\mathbf{Q\Sigma P}^T &= \\
\mathbf{P\Sigma Q}^T\{(\mathbf{Q}^+)^T\mathbf{\Sigma}^+\mathbf{W}_I^{\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}\mathbf{\Sigma}^+\mathbf{Q}^+\}\mathbf{Q\Sigma P}^T &= \\
\mathbf{P\Sigma Q}^T(\mathbf{Q}^+)^T\mathbf{\Sigma}^+\mathbf{W}_I^{\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}\mathbf{\Sigma}^+\mathbf{Q}^+\mathbf{Q\Sigma P}^T &= \\
\mathbf{P\Sigma I\Sigma}^+\mathbf{W}_I^{\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}\mathbf{\Sigma}^+\mathbf{I\Sigma P}^T &= \\
\mathbf{P\Sigma\Sigma}^+\mathbf{W}_I^{\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}\mathbf{\Sigma}^+\mathbf{\Sigma P}^T &= \\
\mathbf{PIW}_I^{\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}\mathbf{IP}^T &= \\
\mathbf{PW}_I^{\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}\mathbf{P}^T &= \\
(\mathbf{W}_I^{-\frac{1}{2}}\mathbf{U})\mathbf{W}_I^{\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}(\mathbf{W}_I^{-\frac{1}{2}}\mathbf{U})^T &= \\
\mathbf{W}_I^{-\frac{1}{2}}\mathbf{UW}_I^{\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}\mathbf{U}^T\mathbf{W}_I^{-\frac{1}{2}} &= \\
\mathbf{UU}^T &
\end{aligned}
\tag{16}
$$

Note that $(\mathbf{W}_I^{-\frac{1}{2}})^+ = \mathbf{W}_I^{\frac{1}{2}}$ and that $\mathbf{W}_I^{-\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}} = \mathbf{I}$. Thus squared MD for categorical data can be defined as $\mathbf{M}' = \widetilde{\mathbf{Z}}_\mathbf{X}(\widetilde{\mathbf{Z}}_\mathbf{X}^T\widetilde{\mathbf{Z}}_\mathbf{X})^+\widetilde{\mathbf{Z}}_\mathbf{X}^T = \mathbf{Z_X}(\mathbf{Z}_\mathbf{X}^T\mathbf{Z_X})^+\mathbf{Z}_\mathbf{X}^T = \mathbf{UU}^T$. Finally and most importantly: because we use the "leverage is Mahalanobis" definition there is no ambiguity nor arbitrary decisions in the definition of squared MD for categorical data (à la, Goodall, 1966); squared MD is defined by the singular vectors for the observations (i.e., $\mathbf{U}$).

3.6.1 From Mahalanobis to χ^2-distance via generalized Euclidean. While our primary focus in the previous section was and for the basis of this paper is to define squared MDs for categorical and mixed data, it is worth noting a much more common distance measure in the CA and MCA literature: the χ^2-distance. To define χ^2-distance we use the same matrices and notation as seen in the definition of MCA (*cf.* Eqs. (11), (12),

(13)). We start with a "row profile" matrix $\mathbf{\Phi}_I = \mathbf{W}_I\mathbf{O}_\mathbf{X}$, which is simply each element of $\mathbf{X}$ divided by its respective row sum. The average profile, $\boldsymbol{\phi}_{I\mu}$, is defined as the column means of $\mathbf{\Phi}_I$, where $\mathbf{\Phi}_{I\mu} = \text{diag}\{\boldsymbol{\phi}_{I\mu}\}$ is the diagonal matrix that contains the average profile on the diagonal. We can define χ^2-distance much in the same way as in Eq. (2) (which is the matrix algebra expression of the generalized squared Euclidean distance formula): $\mathbf{\Xi} = \mathbf{\Phi}_I\mathbf{\Phi}_{I\mu}^{-1}\mathbf{\Phi}_I^T$. We show that $\mathbf{\Xi}$ can be obtained directly through the component scores where, for ease, we use the projection version of the component scores $\mathbf{F}_I = \mathbf{W}_I\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V}$ as:

$$\begin{aligned}
\mathbf{\Xi} = \mathbf{F}_I\mathbf{F}_I^T &= \\
\mathbf{W}_I\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V}(\mathbf{W}_I\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V})^T &= \\
\mathbf{W}_I\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V}\mathbf{V}^T\mathbf{W}_J^{\frac{1}{2}}\mathbf{O}_\mathbf{X}^T\mathbf{W}_I &= \\
\mathbf{W}_I\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{I}\mathbf{W}_J^{\frac{1}{2}}\mathbf{O}_\mathbf{X}^T\mathbf{W}_I &= \\
\mathbf{W}_I\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{W}_J^{\frac{1}{2}}\mathbf{O}_\mathbf{X}^T\mathbf{W}_I &= \\
\mathbf{\Phi}_I\mathbf{W}_J^{\frac{1}{2}}\mathbf{W}_J^{\frac{1}{2}}\mathbf{\Phi}_I^T &= \\
\mathbf{\Phi}_I\mathbf{\Phi}_{I\mu}^{-1}\mathbf{\Phi}_I^T &
\end{aligned} \tag{17}$$

where $\mathbf{\Phi}_{I\mu}^{-1} = \mathbf{W}_J^{\frac{1}{2}}\mathbf{W}_J^{\frac{1}{2}} = \mathbf{W}_J$ because $\boldsymbol{\phi}_{I\mu} = \mathbf{w}_J$ as defined in Eq. (10). Greenacre (2017) and Greenacre and Hastie (1987) argue that the χ^2-distance is a type of Mahalanobis distance. However, we disagree and believe that while both are forms of generalized Euclidean distances they are distinct: squared MD is defined by the singular vectors where χ^2-distance is defined by the component scores. To note, the "χ^2-distance" for PCA would be defined as $\mathbf{F}_I\mathbf{F}_I^T = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}\mathbf{U}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$.

Here we provide shortened expressions of squared MD and χ^2-distance for categorical data both in their generalized Euclidean forms (à la Equation 2) and in their SVD forms. Squared MD is defined as $\mathbf{m}' = \text{diag}\{\mathbf{M}'\}$ where

$$\begin{aligned}
\mathbf{M}' = \widetilde{\mathbf{Z}}_\mathbf{X}(\widetilde{\mathbf{Z}}_\mathbf{X}^T\widetilde{\mathbf{Z}}_\mathbf{X})^+\widetilde{\mathbf{Z}}_\mathbf{X}^T = \\
\mathbf{Z}_\mathbf{X}(\mathbf{Z}_\mathbf{X}^T\mathbf{Z}_\mathbf{X})^+\mathbf{Z}_\mathbf{X}^T = \mathbf{U}\mathbf{U}^T,
\end{aligned} \tag{18}$$

and χ^2-distance is defined as $\boldsymbol{\xi} = \text{diag}\{\boldsymbol{\Xi}\}$ where

$$\begin{aligned}
\boldsymbol{\Xi} = \mathbf{F}_I\mathbf{F}_I^T &= \\
\mathbf{W}_I^{\frac{1}{2}}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T\mathbf{W}_I^{\frac{1}{2}} &= \\
\mathbf{W}_I\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{W}_J^{\frac{1}{2}}\mathbf{O}_\mathbf{X}^T\mathbf{W}_I &= \\
\boldsymbol{\Phi}_I\mathbf{W}_J^{\frac{1}{2}}\mathbf{W}_J^{\frac{1}{2}}\boldsymbol{\Phi}_I^T &= \\
\boldsymbol{\Phi}_I\mathbf{W}_J\boldsymbol{\Phi}_I^T &= \\
\boldsymbol{\Phi}_I\boldsymbol{\Phi}_\mu^{-1}\boldsymbol{\Phi}_I^T.
\end{aligned} \tag{19}$$

4 Establishing the MCD algorithm for categorical data

In this section we will refer to and use a particular toy data set. These toy data are single nucleotide polymorphisms (SNPs) which are categorical variables. The toy data were simulated from real genetic data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). For a description of the data see supplemental section from Beaton et al. (2016). The toy data available at https://github.com/derekbeaton/ExPosition-Family/tree/master/Publications/PsyMet_2015. Each SNP contains three possible genotypic categories: "AA", "Aa", and "aa" which are, respectively, the major homozygote, the heterozygote, and the minor homozygote. Both "A" and "a" represent alleles where "A" is the major (more frequent) allele and "a" is the minor allele. Each genotype is a level (category) within a SNP variable. We provide more details where necessary.

Recall that the MCD algorithm requires two essential parts: the determinant (via the eigenvalues) and squared MD. We now have a formal definition of both for continuous and categorical data. However, one key feature of the MCD algorithm is that it requires that we estimate squared MDs on *sub- or out-of* sample observations. The squared MDs for excluded samples are computed conditional to the mean and covariance of a sub-sample. While this can be done in the standard generalized Euclidean formula, we opt for a more

computationally efficient approach through the SVD and projections. We first show this with continuous data in PCA to establish squared MD for excluded samples. We then show how this can also be done with categorical data via MCA.

However, categorical data poses a unique problem: levels of variables (i.e., categories) are not guaranteed to exist in particular sub-samples. We show that the absence of categories leads to an underestimate of both squared MD and χ^2-distances. Following the illustration of these issues we provide a solution of how to obtain accurate distances (squared MD, robust squared MD, and χ^2-distance) of excluded observations.

In the CA and PCA literatures, "active" data refer to the data submitted to the SVD that form the components, where as "supplementary" data are excluded observations we want to project onto existing components. For squared MD, we will consider "active" data those observations used to compute the the mean and covariance matrix (e.g., $\mathbf{S}$ or $\mathbf{S}'$). The out of sample (or sub-sample) data will be referred to as supplementary.

4.1 Active Mahalanobis and χ^2-distances for continuous data

In order to establish the basis of squared MD for sub- and out-of sample data, we must revisit squared MD through projections à la Eq. (5). We show how to retrieve the singular vectors through the projection of data onto the space spanned by the components because only $\mathbf{U}$ is necessary to compute the squared MD. For quantitative data via PCA:

$$\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{\Sigma}\mathbf{I}\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{I}. \tag{20}$$

Similarly, we can also compute the χ^2-distances for quantitative data via PCA through just a simple rotation of $\mathbf{X}$, as already shown in Eq. (5) and recreated here for convenience: $\mathbf{F}_I = \mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V} = \mathbf{U}\mathbf{\Sigma}\mathbf{I} = \mathbf{U}\mathbf{\Sigma}$.

4.2 Supplementary Mahalanobis distance

Let us refer to a $K \times J$ matrix $\mathbf{Y}$ as a new or excluded set of observations with the same variables (columns) as $\mathbf{X}$. For $\mathbf{Y}$, we would obtain the squared MDs as $\mathbf{M}'_Y = \mathbf{Y}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y}^T$ where the columns of $\mathbf{Y}$ are preprocessed by the same center and scale as the columns in $\mathbf{X}$. In PCA we can project supplementary data onto previously defined components. We can use the same formula as in Eq. (5) to compute supplementary component scores for $\mathbf{Y}$:

$$\mathbf{F}_K = \mathbf{Y}\mathbf{F}_J\mathbf{\Sigma}^{-1} = \mathbf{Y}\mathbf{V}\mathbf{\Sigma}\mathbf{\Sigma}^{-1} = \mathbf{Y}\mathbf{V}\mathbf{I} = \mathbf{Y}\mathbf{V}, \tag{21}$$

where $\mathbf{V}$ are the right singular vectors (loadings) from Eq. (4). More simply we can interpret this as: $\mathbf{V}$ works as rotation of $\mathbf{Y}$ or that we project $\mathbf{Y}$ onto the components of $\mathbf{X}$ spanned by $\mathbf{V}$ to give supplementary component scores for the rows. We can also compute the squared MDs for $\mathbf{Y}$ through these projections instead of $\mathbf{M}'_Y = \mathbf{Y}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y}^T$ where $\mathbf{m}'_Y = \mathrm{diag}\{\mathbf{M}'\}$ (see Eq. (3)). As shown in Eq. (20), we can obtain the squared MDs of $\mathbf{X}$ either directly from the left singular vectors $\mathbf{U}$ or through projections. If we substitute $\mathbf{Y}$ for $\mathbf{X}$:

$$\begin{aligned}
\mathbf{M}'_Y = \mathbf{Y}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y}^T &= \\
\mathbf{Y}(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^{-1}\mathbf{Y}^T &= \\
\mathbf{Y}(\mathbf{V}\mathbf{\Sigma}\mathbf{I}\mathbf{\Sigma}\mathbf{V}^T)^{-1}\mathbf{Y}^T &= \\
\mathbf{Y}(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T)^{-1}\mathbf{Y}^T &= \\
\mathbf{Y}(\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T)\mathbf{Y}^T &= \\
\mathbf{Y}\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T\mathbf{Y}^T &= \\
\mathbf{F}_K\mathbf{\Lambda}^{-1}\mathbf{F}_K^T.
\end{aligned} \tag{22}$$

Thus the squared MDs of $\mathbf{Y}$ are $\mathbf{m}'_Y = \mathrm{diag}\{\mathbf{F}_K\mathbf{\Lambda}^{-1}\mathbf{F}_K^T\}$. Therefore to compute the squared MDs of $\mathbf{Y}$ we need only $\mathbf{V}$ and $\mathbf{\Sigma}$ from the SVD of $\mathbf{X}$ (see Eq. (4)).

4.3 Mahalanobis distances (via the SVD) for sub-samples.

Let us consider that the $K \times J$ matrix $\mathbf{Y} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Z} \end{bmatrix}$, where $\mathbf{X}$ is the same $I \times J$ data set as before and $\mathbf{Z}$ is a new data set of size $(I - K) \times J$; that is $\mathbf{Y}$ is comprised of both active and supplementary data where $\mathbf{Y}$ is (1) centered by the column mean of $\mathbf{X}$ and (2) scaled by the same scaling factor of $\mathbf{X}$. The hat (or projection) matrix of $\mathbf{Y_X} = \mathbf{X}$ is identical to that of $\mathbf{X}$ where $\mathbf{Y_X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y_X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{U}\mathbf{U}^T$, and the hat (or projection) matrix of $\mathbf{Y_Z} = \mathbf{Z}$ is computed as $\mathbf{Y_Z}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y_Z}^T = \mathbf{Z}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Z}^T$ and therefore the squared MDs are $\mathbf{m'_Y} = \text{diag}\{ \begin{bmatrix} \mathbf{U}\mathbf{U}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T\mathbf{Z}^T \end{bmatrix} \}$. Thus we can compute the squared MDs for both active and supplementary sub-samples through the SVD and projections.

4.4 Active Mahalanobis and χ^2-distances for categorical data

In the same way as with continuous data, we can compute both Mahalanobis and χ^2-distances for active data through projections for categorical data. As shown previously, we can obtain the component scores for active in MCA through Eq. (15) where $\mathbf{F}_I = \mathbf{W}_I\mathbf{O_X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V} = \mathbf{W}_I\mathbf{P}\mathbf{\Sigma} = \mathbf{W}_I\mathbf{W}_I^{-\frac{1}{2}}\mathbf{U}\mathbf{\Sigma} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{U}\mathbf{\Sigma}$.

Therefore, if we want to obtain the squared MD through projections and still use the component scores we can do so as:

$$\begin{aligned} \mathbf{U} &= \mathbf{W}_I^{-\frac{1}{2}}\mathbf{W}_I^{\frac{1}{2}}\mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^{-1} = \mathbf{I}\mathbf{U}\mathbf{I} \text{ and} \\ \mathbf{U} &= \mathbf{W}_I^{-\frac{1}{2}}\mathbf{W}_I\mathbf{O_X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V}\mathbf{\Sigma}^{-1} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{O_X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V}\mathbf{\Sigma}^{-1}. \end{aligned} \tag{23}$$

The projection described in Eq. (23) works because of the relationship between GSVD$(\mathbf{W}_I, \mathbf{Z_X}, \mathbf{W}_J)$ and GSVD$(\mathbf{I}, (\mathbf{W}_I^{\frac{1}{2}}\mathbf{O_X}\mathbf{W}_J^{\frac{1}{2}}), \mathbf{I})$. The standard GSVD for MCA, GSVD$(\mathbf{W}_I, \mathbf{Z_X}, \mathbf{W}_J)$, is equivalent to GSVD$(\mathbf{I}, (\mathbf{W}_I^{\frac{1}{2}}\mathbf{O_X}\mathbf{W}_J^{\frac{1}{2}}), \mathbf{I})$ with one slight exception: the first GSVD expression is already centered whereas the second GSVD expression is not. Thus the first singular value in GSVD$(\mathbf{I}, (\mathbf{W}_I^{\frac{1}{2}}\mathbf{O_X}\mathbf{W}_J^{\frac{1}{2}}), \mathbf{I})$ is 1 and the first right and left singular vectors are trivial (equal to the marginal probabilities). Because $\mathbf{V}$ is computed

with the centered data, the projection of $\mathbf{W}_I^{\frac{1}{2}}\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}$ onto $\mathbf{V}$ will remove the trivial mean vector (For more information on the relationship between GSVD$(\mathbf{W}_I, \mathbf{Z}_\mathbf{X}, \mathbf{W}_J)$ and GSVD$(\mathbf{I}, (\mathbf{W}_I^{\frac{1}{2}}\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}), \mathbf{I})$ see Lebart et al. (1984) and Greenacre (1984)).

While Eq. (15) and Eq. (23) suggest ways to compute squared MD or χ^2-distances for *supplementary* (a.k.a. sub- or out-of sample) categorical data, there is a major problem that prevents the computation of *correct* squared MD and χ^2-distances for supplementary categorical data: subsamples will not necessarily have the same columns. In the MCD algorithm, subsamples of size H are treated as active data and both the squared MDs and out of sample squared MDs are easy to compute from PCA (via the SVD). However with categorical data we cannot do so because subsamples are not guaranteed to have the same columns a problem likely to occur when some categories are relatively rare. Because these columns are not in the active data, but will appear in the supplementary data, we cannot compute out-of-sample squared MDs with what we have established so far. However, a specific form of correspondence analysis called generalized correspondence analysis addresses this issue and thus allows us to compute active and supplemental (and thus robust) squared MDs.

4.5 Generalized and subset correspondence analysis

That solution to compute accurate MDs is through a specific form of MCA called subset MCA (Greenacre, 2017; Greenacre & Pardo, 2006). Furthermore, subset MCA is a specific case of Generalized Correspondence Analysis (GCA; Escofier, 1983, 1984). Subset MCA via GCA preserves the shape (rows and columns) of the data and thus can identify the subset of observations within the multivariate space and with the smallest possible scatter.

4.5.1 Escofier's Generalized Correspondence Analysis. Hiding in the broader CA literature is "generalized correspondence analysis"—as referred to by Grassi and Visentin (1994)—established by (Escofier, 1983, 1984). GCA has many uses well beyond the standard applications in CA and MCA, including those for missing data and assumptions

that deviate from the typical assumptions of CA (e.g., quasi-CA with "quasi-independence" and "quasi-margins"; Leeuw & Heijden, 1988; Van der Heijden, De Falguerolles, & Leeuw, 1989).

Escofier established the idea of GCA as: "(DATA - MODEL) / MARGINS". Here we define the idea of "(DATA - MODEL) / MARGINS" more formally, and specifically for categorical data. With a disjunctive data set $\mathbf{X}$, the generalized MCA of $\mathbf{X}$ would be defined as follows. As noted in Grassi and Visentin (1994) we can characterize each cell of the matrix $\widetilde{\mathbf{Z}}_{\mathbf{X}}$ as derived from the generalized model of CA with the generic element

$$\widetilde{z}_{\mathbf{X}_{ij}} = \frac{(o_{\mathbf{X}_{ij}} - \mathbb{e}_{ij})}{\sqrt{\mathbb{W}_{I_{ii}}\mathbb{W}_{J_{jj}}}}. \tag{24}$$

The model ($\mathbb{E}$) and the margins ($\mathbb{W}_I$ and $\mathbb{W}_J$), could be defined in almost any way and it is not required to compute the expected values or margins from the data themselves. We can represent the generalized CA approach in a simpler form with the GSVD where $\mathbb{Z}_{\mathbf{X}} = \mathbf{O}_{\mathbf{X}} - \mathbb{E}$ and GSVD($\mathbb{W}_I, \mathbb{Z}_{\mathbf{X}}, \mathbb{W}_J$) (see Eqs. (13) and (14)). While this is a generalized form, not all choices of $\mathbb{E}$, $\mathbb{W}_I$, and $\mathbb{W}_J$ provide estimates that conform to Pearson's X^2 (i.e., independence; hence why Leeuw and Heijden (1988) refer to some variants as "quasi-independence").

In this generalized MCA, the expected values and weights can be derived from any prespecified models that conform to the rows and columns of the given data. For example let us refer to our toy SNPs data set transformed into disjunctive coding. We would typically use the standard MCA approach where we define $\mathbb{E} = \mathbf{E}_{\mathbf{X}}$, $\mathbb{W}_I = \mathbf{W}_I$, and $\mathbb{W}_J = \mathbf{W}_J$ all of which are derived from the row and column frequencies of $\mathbf{X}$. However for genetic data we could use *population* estimates of the allele frequencies (from, e.g., dbGaP, HapMap or 1000 Genomes) for the columns. Let us refer to the population allele probabilities as $\mathbf{w}_{\mathbb{A}}$ where $\mathbf{w}_{\mathbb{A}}$ are the same J columns as in $\mathbf{X}$. Our generalized MCA with population allele probabilities would be GSVD($\mathbf{W}_I, \mathbf{O}_{\mathbf{X}} - \mathbb{E}_{\mathbf{X}}, \mathbb{W}_J$) where $\mathbb{W}_J = \text{diag}\{\mathbf{w}_{\mathbb{A}}^{-1}\}$ and $\mathbb{E}_{\mathbf{X}} = \mathbf{w}_I\mathbf{w}_{\mathbb{A}}^T$.

GCA provides a way to maintain the row and column structure of the data and to

analyze only subsamples separately in order to make accurate squared MD estimates. To do so, we need only a simple change applied to the generic element $\widetilde{z}_{\mathbf{X}_{ij}}$ for GCA. Let us refer to the active subsample with the subscript $_H$ and the supplementary, or excluded, subsample as $_{\bar{H}}$. First we compute all matrices required for MCA exactly as initially defined in Eqs. (13) and (14): $\mathbf{O}_{\mathbf{X}}$, $\mathbf{E}_{\mathbf{X}}$, $\mathbf{w}_I$ and $\mathbf{w}_J$. Let us frame $\mathbf{O}_{\mathbf{X}}$, $\mathbf{E}_{\mathbf{X}}$, $\mathbf{w}_I$ as constructed 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 $\mathbf{w}_I = \begin{bmatrix} \mathbf{w}_{I_H} \mathbf{w}_{I_{\bar{H}}} \end{bmatrix}$. To maintain the structure of the data and to be able to perform the necessary steps for MCD we must use a generalized expected matrix $\mathbb{E}$ defined 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. Thus our assumption that expected values equal the observed is correct and therefore our assumed deviation (i.e., independence) is 0. The GSVD notation for this is: GSVD$(\mathbf{W}_I, \mathbb{Z}_{\mathbf{X}}, \mathbf{W}_J)$ where $\mathbb{Z}_{\mathbf{X}} = \begin{bmatrix} \mathbf{Z}_{\mathbf{X}_H} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{O}_{\mathbf{X}_H} \\ \mathbf{O}_{\mathbf{X}_{\bar{H}}} \end{bmatrix} - \begin{bmatrix} \mathbf{E}_{\mathbf{X}_H} \\ \mathbf{O}_{\mathbf{X}_{\bar{H}}} \end{bmatrix}$.

Any random subsample partitions of H and $\bar{H}$ guarantee that we maintain the same shape (rows and columns) as $\mathbf{X}$ when we use $\mathbb{E} = \begin{bmatrix} \mathbf{E}_{\mathbf{X}_H} \\ \mathbf{O}_{\mathbf{X}_{\bar{H}}} \end{bmatrix}$. Let us expand GSVD$(\mathbf{W}_I, \mathbb{Z}_{\mathbf{X}}, \mathbf{W}_J)$ where $\mathbb{Z}_{\mathbf{X}} = \mathbf{O}_{\mathbf{X}} - \mathbb{E}$. Like in Eqs. (13) and (14) we have:

$$\begin{bmatrix} \mathbf{Z}_{\mathbf{X}_H} \\ \mathbf{0} \end{bmatrix} = \mathbb{Z}_{\mathbf{X}} = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^T = \begin{bmatrix} \mathbf{P}_H \\ \mathbf{0} \end{bmatrix} \mathbf{\Sigma}\mathbf{Q}^T, \tag{25}$$

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 $\widetilde{\mathbb{Z}}_{\mathbf{X}} = \mathbf{W}_I^{\frac{1}{2}}\mathbb{Z}_{\mathbf{X}}\mathbf{W}_J^{\frac{1}{2}}$ and thus $\mathbf{W}_I^{\frac{1}{2}}\mathbb{Z}_{\mathbf{X}}\mathbf{W}_J^{\frac{1}{2}} = \mathbf{W}_I^{\frac{1}{2}} \begin{bmatrix} \mathbf{Z}_{\mathbf{X}_H} \\ \mathbf{0} \end{bmatrix} \mathbf{W}_J^{\frac{1}{2}} = \begin{bmatrix} \widetilde{\mathbf{Z}}_{\mathbf{X}_H} \\ \mathbf{0} \end{bmatrix} = \widetilde{\mathbb{Z}}_{\mathbf{X}}$. Just as in Eq. (14) we would apply the SVD as:

$$\widetilde{\mathbb{Z}}_{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \begin{bmatrix} \mathbf{U}_H \\ \mathbf{0} \end{bmatrix} \mathbf{\Sigma}\mathbf{V}^T, \tag{26}$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{I} = \mathbf{V}^T\mathbf{V}$ and

$$\mathbf{U}^T\mathbf{U} = \begin{bmatrix} \mathbf{U}_H \\ \mathbf{0} \end{bmatrix}^T \begin{bmatrix} \mathbf{U}_H \\ \mathbf{0} \end{bmatrix} = \mathbf{I}. \tag{27}$$

Through GCA we maintain the structure of the original data but now the data from the excluded sample are comprised of 0s through each step and thus do not contribute to the overall structure. Hence $\mathbf{V}$ and $\mathbf{Q}$ are the sets of singular vectors and generalized singular vectors, respectively, for only the H subsample with the same number of columns as $\mathbf{X}$. However, it is possible to simplify this approach through a derivation of this application of GCA called "subset MCA" (Greenacre, 2017; Greenacre & Pardo, 2006).

4.5.2 Subset Multiple Correspondence Analysis. Our specific use of GCA is designed for use with the MCD algorithm. As shown in Eqs. (25) and (26) we can set the exepcted values of the excluded subset $\bar{H}$ to 0 so that we only analyze the active subset H. Our specific form of GCA reduces to subset MCA via Eq. (27) where

$$\mathbf{U}^T\mathbf{U} = \begin{bmatrix} \mathbf{U}_H \\ \mathbf{0} \end{bmatrix}^T \begin{bmatrix} \mathbf{U}_H \\ \mathbf{0} \end{bmatrix} = \mathbf{U}_H^T\mathbf{U}_H = \mathbf{I} \tag{28}$$

thus when we compute the SVD of $\widetilde{\mathbb{Z}}$ we do not require all of the observations we only require those that are not 0. Our use of subset MCA can be characterized as follows.

To compute subset MCA for MCD we use the matrices defined in Eqs. (13) and (14). Specifically $\widetilde{\mathbf{Z}}_{\mathbf{X}} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{Z}_{\mathbf{X}}\mathbf{W}_J^{\frac{1}{2}} \iff \mathbf{Z}_{\mathbf{X}} = \mathbf{W}_I^{-\frac{1}{2}}\widetilde{\mathbf{Z}}_{\mathbf{X}}\mathbf{W}_J^{-\frac{1}{2}}$. 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

$$\widetilde{\mathbf{Z}}_{\mathbf{X}_H} = \mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^T. \tag{29}$$

Let us define the subset MCA of the complement $\bar{H}$ as

$$\widetilde{\mathbf{Z}}_{\mathbf{X}_{\bar{H}}} = \mathbf{U}_{\bar{H}}\mathbf{\Sigma}_{\bar{H}}\mathbf{V}_{\bar{H}}^T \tag{30}$$

where $\widetilde{\mathbf{Z}}_{\mathbf{X}} = \begin{bmatrix} \widetilde{\mathbf{Z}}_{\mathbf{X}_H} \\ \widetilde{\mathbf{Z}}_{\mathbf{X}_{\bar{H}}} \end{bmatrix}$. In the CA literature the sum of the eigenvalues (i.e., diag$\{\mathbf{\Lambda}\}$) is called "inertia" which we denote as $\mathcal{I}$. Note that

$$\mathcal{I} = \mathcal{I}_H + \mathcal{I}_{\bar{H}}, \tag{31}$$

In essence, subset MCA partitions the full MCA space into additive subspaces.

Our particular implementation of GCA and use of subset MCA produce identical results via the SVD because $\widetilde{\mathbf{Z}}_{\mathbf{X}_H} = \mathbf{W}_I^{\frac{1}{2}} \begin{bmatrix} \mathbf{Z}_{\mathbf{X}_H} \\ \mathbf{0} \end{bmatrix} \mathbf{W}_J^{\frac{1}{2}}$ and thus all other computations, such as the component scores and the generalized singular vectors, are also equal between our use of GCA and subset MCA.

4.5.3 Subset MCA for categorical MCD. For MCD algorithms we require (1) the singular values so we can compute the determinant and (2) robust squared MD estimates. We can obtain the singular values in subset MCA (see Eqs. (29) and (31)) and thus we can compute the determinant for H subsamples via the geometric mean of the eigenvalues: $\mathbf{\Lambda}_H = \mathbf{\Sigma}_H^2$. Robust squared MDs can be computed in subset MCA through the row profiles. Recall that we can compute component scores from the full MCA via projection as $\mathbf{F}_I = \mathbf{\Phi}_I \mathbf{F}_J \mathbf{\Sigma}^{-1} = \mathbf{W}_I \mathbf{O}_{\mathbf{X}} \mathbf{F}_J \mathbf{\Sigma}^{-1}$. We can also compute component scores for the full I set from the subset MCA as

$$\begin{bmatrix} \mathbf{F}_H \\ \widehat{\mathbf{F}}_{\bar{H}} \end{bmatrix} = \mathbf{\Phi}_I \mathbf{F}_{J_H} \mathbf{\Sigma}_H^{-1} \tag{32}$$

where the χ^2-distances are equivalent between subset and full MCA as diag$\{ \begin{bmatrix} \mathbf{F}_H \\ \widehat{\mathbf{F}}_{\bar{H}} \end{bmatrix}^T \begin{bmatrix} \mathbf{F}_H \\ \widehat{\mathbf{F}}_{\bar{H}} \end{bmatrix} \} = \text{diag}\{\mathbf{F}_I^T \mathbf{F}_I\}$. Recall that we can compute the squared MDs in full MCA as $\mathbf{U} = \mathbf{W}_I^{\frac{1}{2}} \mathbf{O}_{\mathbf{X}} \mathbf{W}_J^{\frac{1}{2}} \mathbf{V} \mathbf{\Sigma}^{-1}$ (as seen in Eq. (23)) or alternatively as $\mathbf{U} = \mathbf{W}_I^{-\frac{1}{2}} \mathbf{\Phi}_I \mathbf{F}_J \mathbf{\Lambda}^{-1}$ because $\mathbf{W}_I^{-\frac{1}{2}} \mathbf{\Phi}_I \mathbf{F}_J \mathbf{\Lambda}^{-1} = \mathbf{W}_I^{-\frac{1}{2}} \mathbf{\Phi}_I \mathbf{W}_J \mathbf{Q} \mathbf{\Sigma} \mathbf{\Lambda}^{-1} = \mathbf{W}_I^{-\frac{1}{2}} \mathbf{\Phi}_I \mathbf{W}_J \mathbf{W}_J^{-\frac{1}{2}} \mathbf{V} \mathbf{\Sigma} \mathbf{\Lambda}^{-1} = \mathbf{W}_I^{-\frac{1}{2}} \mathbf{\Phi}_I \mathbf{W}_J \mathbf{W}_J^{-\frac{1}{2}} \mathbf{V} \mathbf{\Sigma}^{-1} = \mathbf{W}_I^{-\frac{1}{2}} \mathbf{W}_I \mathbf{O}_{\mathbf{X}} \mathbf{W}_J \mathbf{W}_J^{-\frac{1}{2}} \mathbf{V} \mathbf{\Sigma}^{-1} = \mathbf{W}_I^{\frac{1}{2}} \mathbf{O}_{\mathbf{X}} \mathbf{W}_J^{\frac{1}{2}} \mathbf{V} \mathbf{\Sigma}^{-1}$.

With this formulation, we can compute *supplementary* squared MDs for the full set of observations conditional to the subset MCA results as:

$$\begin{bmatrix} \mathbf{U}_H \\ \widehat{\mathbf{U}}_{\bar{H}} \end{bmatrix} = \mathbf{W}_I^{-\frac{1}{2}} \mathbf{\Phi}_I \mathbf{F}_{J_H} \mathbf{\Lambda}_H^{-1} \tag{33}$$

where the $\mathbf{U}_H$ here in Eq. (33) is exactly the $\mathbf{U}_H$ in Eqs. (26) and (29). Alternatively we can compute Eq. (33) from (32) as $\begin{bmatrix} \mathbf{U}_H \\ \widehat{\mathbf{U}}_{\bar{H}} \end{bmatrix} = \mathbf{W}_I^{-\frac{1}{2}} \begin{bmatrix} \mathbf{F}_H \\ \widehat{\mathbf{F}}_{\bar{H}} \end{bmatrix} \mathbf{\Sigma}^{-1}$. Finally, the robust squared MDs are computed as diag$\{ \begin{bmatrix} \mathbf{U}_H \\ \widehat{\mathbf{U}}_{\bar{H}} \end{bmatrix} \begin{bmatrix} \mathbf{U}_H \\ \widehat{\mathbf{U}}_{\bar{H}} \end{bmatrix}^T \}$.

Thus subset MCA can be used to generalized the MCD algorithm to categorical data: via subset MCA for H subsamples we obtain (1) the determinant and (2) the robust squared MD estimates via projection onto the singular vectors.

4.5.4 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: $\widetilde{\mathbf{Z}}_{\mathbf{X}_H} = \mathbf{U}_H \mathbf{\Sigma}_H \mathbf{V}_H^T$
2. Compute the determinant of $\widetilde{\mathbf{Z}}_{\mathbf{X}_H}$ as the geometric mean of $\mathbf{\Lambda}_H$
3. Compute the squared MDs for the full I sample as $\begin{bmatrix} \mathbf{U}_H \\ \widehat{\mathbf{U}}_{\bar{H}} \end{bmatrix} = \mathbf{W}_I^{-\frac{1}{2}} \mathbf{\Phi}_I \mathbf{F}_{J_H} \mathbf{\Lambda}_H^{-1}$
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 either through optimal search or limited to a given number of iterations. A minimum determinant (Step 2) reflects the minimum dispersion of individuals within the space. Our approach employs steps like the Fast-MCD algorithm (Rousseeuw & Van Driessen, 1999); See also Hubert and

Debruyne (2010) for overview. Step 3 helps find a minimum determinant faster because the subset of H observations with the smallest squared MDs are likely to produce the smallest determinant. Furthermore, these steps can start with many H subsets and perform the algorithm from each random H set. See *Supplemental Material* for the complete steps for our generalized MCD algorithm.

We illustrate our formalization of a categorical MCD with our toy data set. The toy data are of size $I = 60 \times C = 9$ (observations by variables). These data were transformed into a disjunctive coding (see Table 1) that are of size $I = 60 \times J = 25$ (observations by columns). 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; note that the intial determinant is quite high by comparison to later determinants. Eventually the algorithm stops at $1,000$ (the number of iterations we set) on a much smaller determinant. Figure 1b shows the observed squared MD and χ^2-distances. The observed distances alone (Fig, 1b) suggest the presence of possible outliers. We have denoted two individuals with "X" circumscribed by a circle. These two individuals are the only two with "aa" for one particular SNP. In Figure 1c we show the observed distances again but now individuals in red are those within the H subsample and individuals in blue are those excluded from the "best" subsample (i.e., the subsample with the smallest determinant at convergence). Finally, like with the standard MCD algorithm we can also compute robust squared MD which helps reveal "masking" effects by exaggerating the more extreme individuals. Figure 1d shows the observed vs. robust squared MD. Figure 1e is the same as Figure 1d except the two most extreme individuals to provide a better view of the observed vs. robust squared MDs.

5 Applications and Extensions

Now that we have established a MCD algorithm for categorical data we show in this section examples of its use including how our approach generalizes the MCD to allow for virtually any data type (i.e., categorical, ordinal, continuous). Because our approach is based

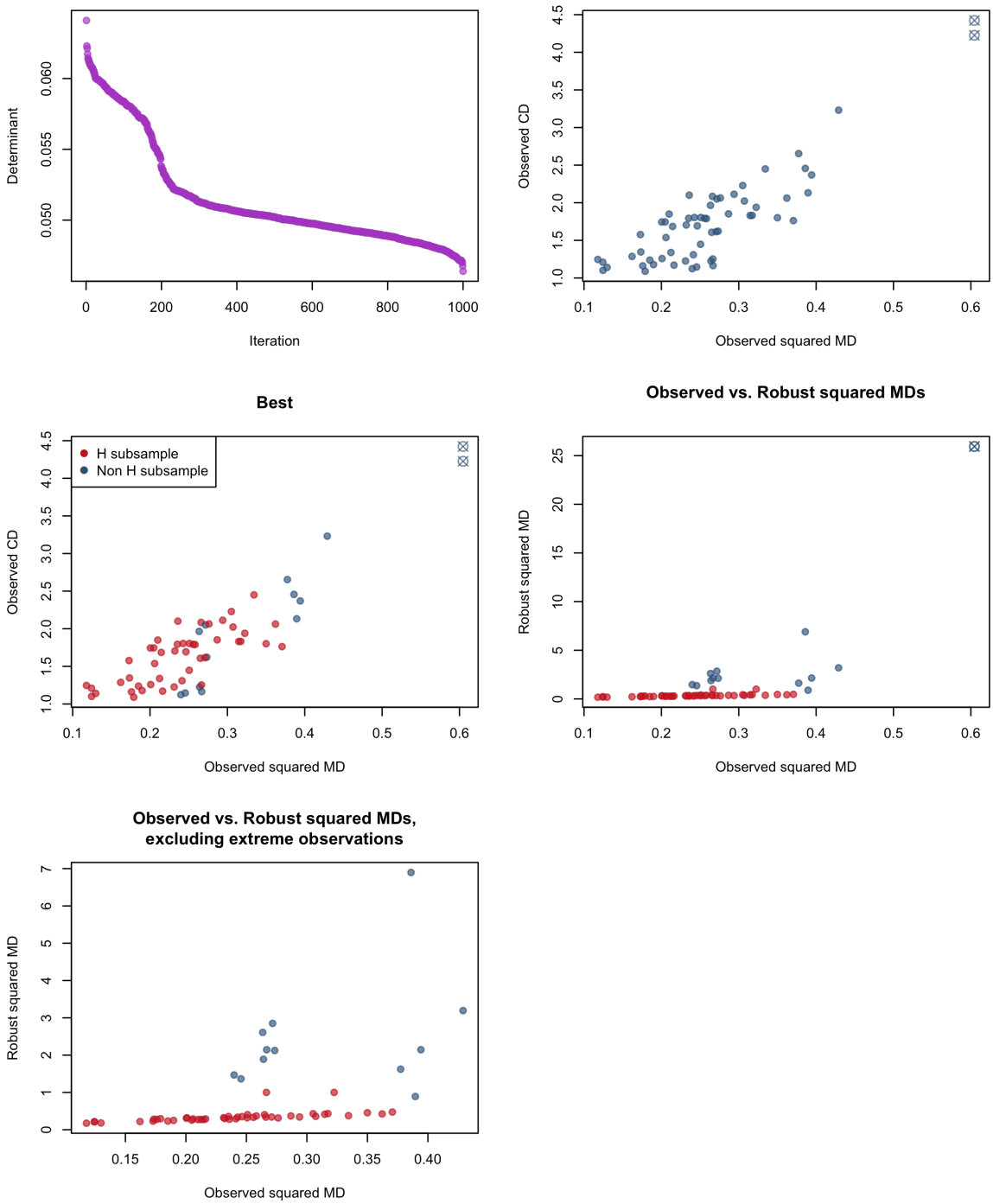

Figure 1. Application of generalized MCD to categorical data. We illustrate the generalized MCD here first on a small data set of single nucleotide polymorphisms (SNPs). (a) shows the search for the determinant. (b) shows the observed distances (via standard multiple correspondence analysis). (c) shows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

on CA, we can capitalize on particular properties of CA and X^2 to allow for the analyses of non-categorical data such as ordinal, continuous, or mixed-data. We show how this is possible through a data recoding technique called data "doubling" (a.k.a. fuzzy coding, bipolar coding, or "Escofier transform").

In this section we illustrate our generalized MCD with two data sets: one from the Ontario Neurodegernative Disease Research Initiative (ONDRI) and one from the Alzheimer's Disease Neuroimaging Initiative (ADNI). First we apply our approach to the miniature survey of autobiographical memory (MSAM; adapted from Palombo, Williams, Abdi, & Levine, 2013) and treat the MSAM categorical data. We then apply our approach to two recoded versions of the MSAM: one recoded to eliminate rare categories and another that treats the data as ordinal responses. We compare and contrast the results from all three applications to the MSAM. Finally, we use a mixture of data types from ADNI data to demonstrate how our technique applies to data sets with mixed variables. The ADNI data include categorical data (genotypic markers from single nucleotide polymorphisms), ordinal data (clinical dementia rating), and continuous data (volumetric estimates from brain regions). We explain each data set and applications in their own sections.

5.1 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): Alzheimer's disease (AD) and amnestic mild cognitive impairment (MCI), amyotrophic lateral sclerosis (ALS), frontotemporal lobar degeneration (FTD), Parkinson's disease (PD), and vascular cognitive impairment (VCI). Currently, the ONDRI study is still in the data collection, curation, and quality control stages. Thus these are preliminary data used for illustrative purposes.

As part of the quality control process, various subsets of data are subjected to outlier analyses []. However, one of the major challenges is that many of the variables and

instruments are not quantitative and we therefore could not use existing techniques such as the MCD, robust PCA (Candès, Li, Ma, & Wright, 2011) or principal orthogonal complement threshold (Fan, Liao, & Mincheva, 2013).

Here, we illustrate our generalized MCD on one such data set: the MSAM. We show the MSAM on $N = 300$ particpants from two cohorts (vascular cognitive impairment $n = 161$ and parkinson's disease $n = 139$). We focus our illustration on identification of outliers and a robust subset of individuals in the MSAM regardless of disease, age, sex, and other factors.

5.1.1 MSAM data. The MSAM is a short version of the survey of autobiographical memory (SAM). The MSAM is comprised of 10 questions and each question has five possible responses: two indicate disagreement with a question ("strongly disagree", "disagree somewhat"), two indicate agreement with a question ("agree somewhat", "strongly agree") and one neutral respones ("neither"). Here, we analyze the MSAM in three ways so that we can compare and contrast the results of each. In our first example, we treat the data as categorical and apply our technique as previously derived and illustrated (see Figure 1). As noted in Figure 1 very rare responses can greatly exaggerate the squared MD, robust squared MD, and even χ^2-distances. Furthermore, in many practical applications rare responses would likely be recoded so that they are combined with similar responses. For example if a response of "strongly agree" occurred in less than approximately 5% of responses for a particular question, it would be recoded with "agree somewhat" to make a single "agreement" response. Therefore for our second example we recode rare ($\approx 5\%$) responses where we combine a rare response (e.g., "strongly agree") with its most similar response (e.g., "agree"). We continue to treat the data as categorical and apply our technique again. Finally, because the responses could be treated as ordinal data we use a specific recoding scheme that has many names such as "bipolar coding" (Greenacre, 1984), fuzzy coding, or "doubling" (Greenacre, 2014; Lebart et al., 1984). We refer to this coding henceforth as "fuzzy coding". First we explain fuzzy coding. Next we compare and contrast the results of each analyses with respect to which observations were identified as the robust

subsample across each technique and several α levels (i.e., $\alpha = .5$, $\alpha = .75$, $\alpha = .9$). Finally, we also illustrate how to use the bootstrap procedure to define thresholds for outliers given the robust squared MDs.

5.1.1.1 Fuzzy coding. With fuzzy coding each original variable was represented by two columns where the "$-$" column is the observed value minus the minimum of the variable for that question, and the "$+$" column is the maximum of the variable for that question minus the observed value. For example, the MSAM has 5 possible levels (1, 2, 3, 4, 5) so each "$-$" column will use 1 as the minimum and each "$+$" column will use 5 as the maximum, even if there are no *observed* values of 1 or 5 in the data themselves. The data are then normed so that each pair of columns (i.e., $+$ and $-$) sum to 1. Because the data behave as disjunctive data, we can apply MCA and thus our GMCD approach also works to find a robust subset and help identify outliers. The two columns are then normalized so that the row sums equal the total number of original variables. Fuzzy coding is illustrated in Table 2.

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) even though the data are not disjunctive: the sums of the rows equal to the number of (original) variables, the sum of variables (each pair of columns) equal the number of rows, and the sum of the table is equal to the number of rows $\times$ the number of (original) variables.

5.1.2 ONDRI Examples. In the following examples we refer to the original MSAM data as "as is", the recoded for rare responses version "recode", and the ordinal version as "ordinal". We applied the generalized MCD to each with three alpha levels: $\alpha = .5$, $\alpha = .75$, $\alpha = .9$ for a total of nine applications. Like 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` and `robustbase` packages in `R` via the `h.alpha.n` function). Because each data set has a different number of columns (J) the same α will produce a different size H. Each application of the analysis was run for $1,000$

Table 2

Example of fuzzy coding

(a) *CDR ordinal example*

	Q1: Specific events	Q2: Recall objects
OND01_BYC_1006	2	4
OND01_CAM_1020	4	4
...	...	...
OND01_HGH_5001	5	4
OND01_HGH_5003	4	5

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

	Q1−	Q1+	Q2−	Q2+
OND01_BYC_1006	$\frac{2-1}{4} = \frac{1}{4}$	$\frac{5-2}{4} = \frac{3}{4}$	$\frac{4-1}{4} = \frac{3}{4}$	$\frac{5-4}{4} = \frac{1}{4}$
OND01_CAM_1020	$\frac{4-1}{4} = \frac{3}{4}$	$\frac{5-4}{4} = \frac{1}{4}$	$\frac{4-1}{4} = \frac{3}{4}$	$\frac{5-4}{4} = \frac{1}{4}$
...	...	...	...	...
OND01_HGH_5001	$\frac{5-1}{4} = \frac{4}{4}$	$\frac{5-5}{4} = \frac{0}{4}$	$\frac{4-1}{4} = \frac{3}{4}$	$\frac{5-4}{4} = \frac{1}{4}$
OND01_HGH_5003	$\frac{4-1}{4} = \frac{3}{4}$	$\frac{5-4}{4} = \frac{1}{4}$	$\frac{5-1}{4} = \frac{4}{4}$	$\frac{5-5}{4} = \frac{0}{4}$

Note. Example of data "doubling" via fuzzy for ordinal data with the miniature survey of autobiographical memory (MSAM).

iterations for the GMCD algorithm.

Figure 2 shows all nine applications of the generalized MCD applied to the MSAM. To note in Figure 2 we show the MDs and robust MDs (not squared). Part of the reason we show these is because like with the toy data in the previous section some observations excluded from the H subsample (denoted in blue) have very high robust squared MDs by comparison to the H subsample (denoted in red). Though these are the same data, we see interesting behaviors when we treat the data differently. First, when α is low, we see a greater exaggeration of outlying individuals with respect to the "inliers" and the H subset. However this behavior essentially disappears as α increases. Finally we also see that the "ordinal" version of the data appears to behave quite differently from the other versions as α increases. In general, the MD and robust MD estimation are highly similar, with only a small gap between the H subsample and the observations excluded from H. The robust MD estimates across techniques tends to vary as a consequence of both data representation (as is, recoded, ordinal) and α. However the H subsamples across each of these are actually highly similar.

The overlap of the H subsample between all applications of the MSAM analyses is shown in Table 3. When α is small there is more variability between the different versions. As α increases the H subset becomes much more similar across all versions. However it is important to note that of all of these analyses, the "as is" and the "ordinal" generally find the same observations as part of the H subset, even though the data are represented in two completely different ways. The mostly common subsamples between "as is" and ordinal analyses show that both retain essentially the same information.

5.1.2.1 Identifying inliers and outliers. Besides the identification of robust subsamples (and covariance structures), MCD algorithms reveal "masking" effects and identify outliers. In some of our MSAM examples in Figure 2 it is fairly easy to identify "inliers" and the extreme outliers that were previously subject to the "masking" effects. For example, in the "as is" data regardless of the α, it is clear that an outlier subset of individuals

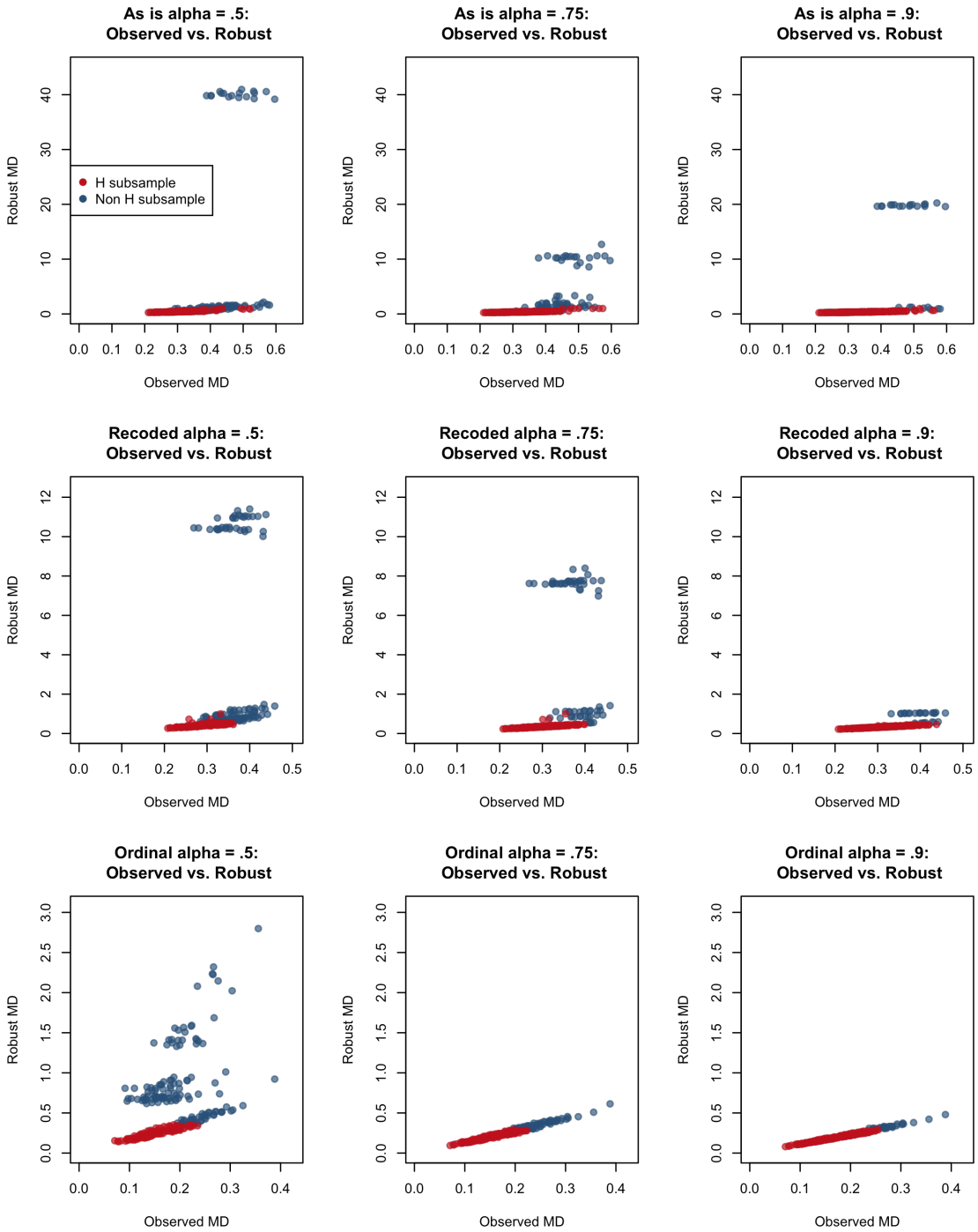

Figure 2. Generalized MCD applied to the miniature survey of autobiographical memory (MSAM) data from ONDRI. Left to right is the alpha level and top to bottom is how the data were coded. All plots show observed vs. robust Mahalaniobis distances (MD). Observations in red are the H subsample, observations 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.

	A.5	A.75	A.9	R.5	R.75	R.9	O.5	O.75	O.9
As Is .5	175								
As Is .75	163	237							
As Is .9	174	234	275						
Recode .5	134	147	162	171					
Recode .75	157	190	220	168	235				
Recode .9	172	216	252	170	234	274			
Ordinal .5	120	142	157	128	138	152	160		
Ordinal .75	163	207	224	153	189	210	157	230	
Ordinal .9	170	229	261	166	218	248	160	230	272

Note. A = As is, R = Recode, O = Ordinal. Only the lower triangle and diagonal are shown. The diagonal here indicates the size of the H subset. The value in each cell is the number of overlapping subjects identified as the robust subsample for each version of the miniature survey on autobiographical memory (MSAM). Generally as the α increases, overlap becomes more likely. As is and ordinal overlap the most across all versions.

have a highly unique response pattern and differ greatly from the rest of the sample.

As previously described MCD algorithms aim to find the most robust H subsample wherein the scatter (covariance) is minimized. Sometimes, observations in the H subsample have higher robust squared MDs than those outside of the H subsample, like in the "as is" MSAM at $\alpha = .5$. However other times, like in the oridinal MSAM at $\alpha = .9$, there is no clear division between the H subsample and the rest of the observations. However Therefore to help identify inliers and outliers we recommend using the bootstrap (Efron, 1982, 1992) to generate a distribution of robust squared MDs. We refer to "inliers" here as those individuals that are not identified as outliers. We refer to outliers here as those above some threshold.

The bootstrap derived distribution of robust squared MDs can be used to identify observations in the sample that exist inside or outside of particular thresholds. The bootstrap procedure we used is as follows. First we bootstrapped the data (in disjunctive or fuzzy format) to create a new MSAM data set of the same size. Next we preprocessed the data just as in standard MCA and projected the preprocessed data onto the robust singular vectors to compute robust squared MDs. We repeat this process many (e.g., 100) times. Each bootstrapped version of the data generates a set of robust squared MDs. All of the robust squared MDs across all bootstrapped sets are then used to create a distribution of robust squared MDs. We use the distribution of bootstrapped robust squared MDs to identify any of the originally observed robust squared MDs that are outside of a given threshold. For this example we used two thresholds: (1) a fixed threshold at 95% of the bootstrapped robust squared MDs distribution, and (2) proportional to H subsample size for each analysis (as determined by α). In Figure 3 we highlight individuals below both thresholds (red), between the thresholds (gold), and above the 95% threshold (blue). For the bootstrap procedure we refer to "inliers" as those individuals that are always under both thresholds. We show two types of outliers in this example: the less extreme outliers between the two thresholds and the extreme outliers beyond the 95% threshold. We use both outlier thresholds to highlight the behavior of the bootstrap procedure across the different

applications of the MSAM analyses.

With the bootstrap procedure, especially in the $\alpha = .5$ analyses, we can see that the "inliers" constitute a more homogeneous set than those identified as the H subsample (Fig. 3). This is in part because the MCD algorithm identifies a subset with the minimum determinant (scatter) but not necessarily the most homogeneous H subsample. Furthermore the bootstrap procedure helps better identify inliers and outliers when the robust MDs have a relatively linear relationship the MDs like in the ordinal examples with $\alpha = .75$ and $\alpha = .9$. The overlap of the inliers between all versions of the MSAM analyses are shown in Figure 4. Similar to the H subsample identification, when α is small there is more variability between the different versions. Again like in the fixed analyses (1) as α increases the inliers become much more similar across all versions, and (2) the "as is" and the "ordinal" generally find the same observations as part of the inliers, even though the data are represented in two completely different ways.

Finally we highlight the overlap between the H subsamples and the inliers identified via bootstrap in Table 5. The diagonal in Table 5 shows the overlap between the H subsample and bootstrap inliers for the same analysis. Off diagonal elements show the relationships between the analyses (i.e., different coding approaches and α levels). While this table reflects the same general story as the previous tables it does provide a clearer sense of similarity between the H subsamples and the bootstrap-based inliers. For the identification of inliners and outliers we do strongly recommend using the bootstrap procedure instead of just the H subsample (e.g., red individuals in Fig. 2) or, for example, quantiles of the robust squared MDs. The bootstrap procedure provides a more stable assessment of individuals and their likelihood of classification as "inlier" or "outlier".

5.2 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

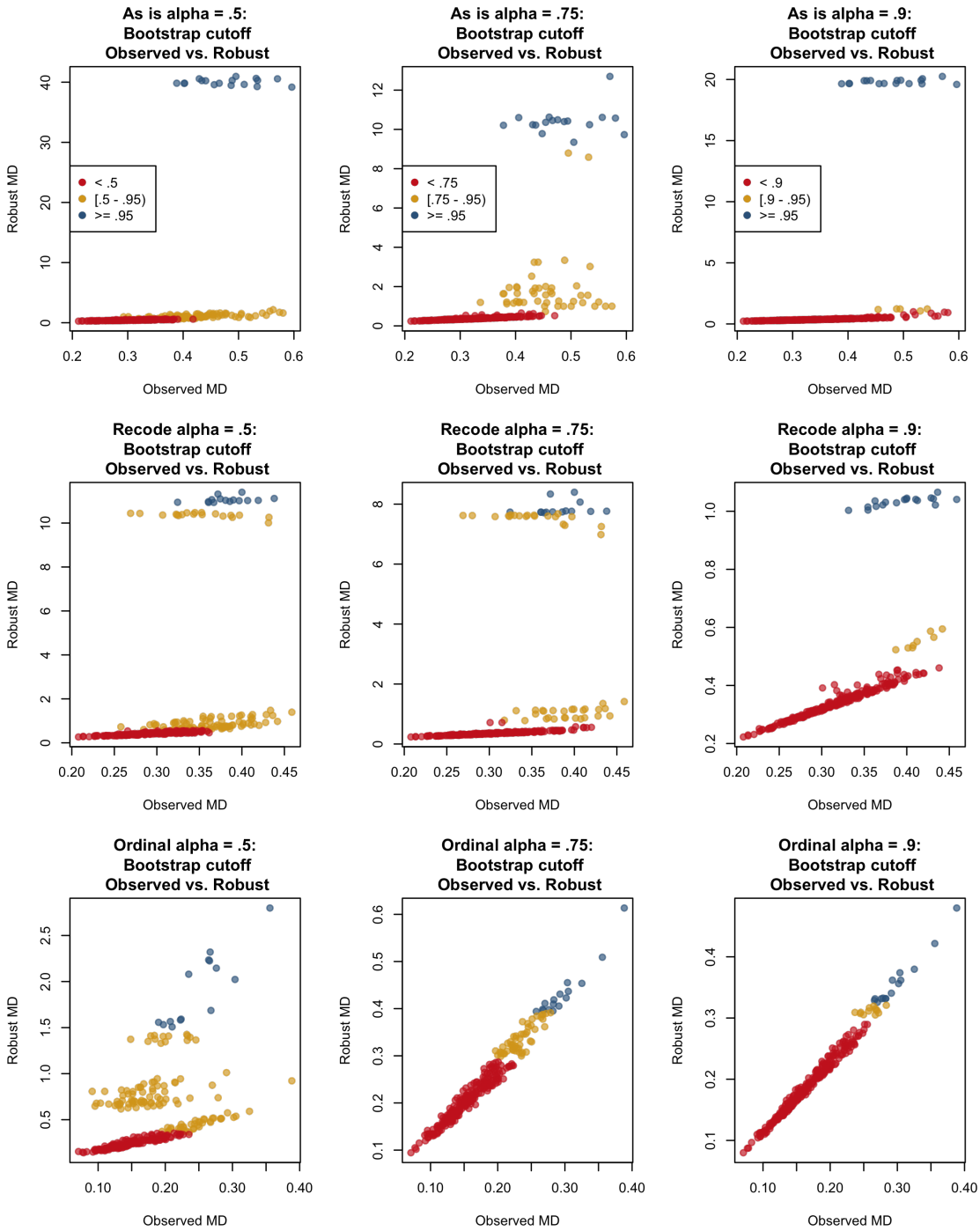

Figure 3. Bootstrapping procedure applied the generalized MCD of the miniature survey of autobiographical memory (MSAM). All plots show observed vs. robust Mahalaniobis distances (MD). Observations in red are the bootstrap 'inliers', observations in blue are outside the bootstrap threshold, and observations in gold are in between. Axes are scaled to the values within each plot to better highlight the cutoffs via the bootstrap distributions.

Table 4

Inliers identified via bootstrap.

	A.5	A.75	A.9	R.5	R.75	R.9	O.5	O.75	O.9
As Is .5	178								
As Is .75	178	236							
As Is .9	178	236	278						
Recode .5	140	151	168	176					
Recode .75	167	192	225	175	239				
Recode .9	178	216	254	176	239	274			
Ordinal .5	122	144	157	126	137	152	161		
Ordinal .75	171	209	224	156	189	210	158	230	
Ordinal .9	177	232	262	171	221	248	161	230	272

Note. A = As is, R = Recode, O = Ordinal. Only the lower triangle and diagonal are shown. The diagonal here indicates the size of the H subset. The value in each cell is the number of overlapping subjects not identified as outliers (i.e., inliers) via bootstrap with reference to the robust structure for each version of the miniature survey on autobiographical memory (MSAM). The threshold for each was the lower value (i.e., α). Generally as the α increases, overlap becomes more likely. As is and ordinal overlap the most across all versions.

Table 5

Overlap between the robust subsamples and bootstrap inliers.

	B_A.5	B_A.75	B_A.9	B_R.5	B_R.75	B_R.9	B_O.5	B_O.75	B_O.9
As Is .5	159	165	174	136	159	172	120	163	170
As Is .75	175	230	235	151	192	216	143	207	229
As Is .9	178	236	275	167	224	252	157	224	261
Recode .5	136	148	163	167	168	170	128	153	166
Recode .75	166	190	221	175	234	234	138	189	218
Recode .9	178	216	254	176	239	274	152	210	248
Ordinal .5	122	144	157	126	137	152	160	157	160
Ordinal .75	171	209	224	156	189	210	158	230	230
Ordinal .9	177	232	262	171	221	248	161	230	272

Note. B_ = Bootstrap. A = As is, R = Recode, O = Ordinal. Overlap between the robust subsamples identified via generalized MCD and the inliers identified via bootstrap with respect to the robust structures. Generally, the inliers overlap with the robust subsample, especially as α increases.

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).

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. We used several data sets of different data types from ADNI: single nucleotide polymorphisms (SNPs) as categorical data, the clinical dementia rating (CDR) as ordinal data, and volumetric estimates from five brain regions as continuous data. Like ordinal data, continuous data can also be transformed in a way where it behaves like disjunctive data. These data can be acquired in ADNI via the genome-wide data and the `ADNIMERGE R` package.

5.2.1 ADNI Mixed-data example. We used SNPs from the ADNI cohort that were associated with APOE and TOMM40 which are strong genetic contributions to AD (Roses et al., 2010). SNP data were preprocessed for participant and SNP missingness (5%, i.e., required 95% completeness), and minor allele frequency $< 5\%$. Infrequent genotypes (e.g., $< 5\%$) were recoded; in this particular case all "aa" levels that were $< 5\%$ were combined with "Aa" – essentially dichotomizing the presence and absence of "a" (i.e., "AA" vs. presence of "a"). 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. Similar to ordinal data, we code the data as "doubled" continuous data so that they, 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) because it was proposed by Escofier (1979). The transformation is similar to "fuzzy coding" in that it too is "bipolar". The "doubling" of continuous data is simpler than with ordinal and only requires that we add or subtract 1 from the centered and scaled data, and divide each value by 2. An example of "doubling" for continuous data can be seen in Table 6. With "doubled" continous data we can apply GMCD to continuous data because it now behaves like disjunctive data. Our mixed data was thus a matrix of $I = 613 \times J = 57$ (13 SNPs categorical that spanned 35 columns, 6 CDR domains that spanned 12 columns, and 5 brain regions that spanned 10 columns).

Figure 4 shows our GMCD applied to the mixed data set. Figure 4a, shows the search for a minimum determinant. In Figure 4b we see the observed MD and χ^2-distances which suggests the presence of some possible outliers as well as a possibly homogeneous group of individuals near that extend outward from 0. In Figures 4c and d, individuals in red are those within the subsample and blue are those excluded; Figure 4c shows the observed squared MD and χ^2-distances for individuals (in red) that comprise the "best" determinant (the sample at convergence). Finally, like with the standard MCD algorithm we can also compute robust MD which helps revealing "masking" effects by exaggerating the more extreme individuals. Figure 4d shows the observed vs. robust MD; the robust MD was computed with respect to the individuals identified as the minimum determinant set in Figure 4c. Figure 4e shows outliers (in blue) as identified through our bootstrap procedure. Like before we used two cutoffs: the lower bound was proportional to the H subsample and the upper was the 95%-ile of the bootstrap distribution. In the mixed data case 335 observations were identified as H subsample and 432 were identified as inliers of the total 613 observations. In this example, 335 of the H subsample overlapped with the inliers.

Table 6

Example of Escofier transform

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

	Hippocampus	Fusiform
002_S_0295	0.34	-0.46
016_S_0590	-0.39	1.80
...	...	...
027_S_0850	-1.02	0.92
041_S_0898	-0.10	0.26

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

	Hippocampus−	Hippocampus+	Fusiform−	Fusiform+
002_S_0295	$\frac{1-0.34}{2} = 0.33$	$\frac{1+0.34}{2} = 0.67$	$\frac{1-(-0.46)}{2} = 0.73$	$\frac{1+(-0.46)}{2} = 0.27$
016_S_0590	$\frac{1-(-0.39)}{2} = 0.70$	$\frac{1+(-0.39)}{2} = 0.30$	$\frac{1-1.80}{2} = -0.40$	$\frac{1+1.80}{2} = -1.40$
...	...	...	...	...
027_S_0850	$\frac{1-(-1.02)}{2} = 1.01$	$\frac{1+(-1.02)}{2} = -0.01$	$\frac{1-0.92}{2} = 0.04$	$\frac{1+0.92}{2} = 0.96$
041_S_0898	$\frac{1-(-0.1)}{2} = 0.55$	$\frac{1+(-0.1)}{2} = 0.45$	$\frac{1-0.26}{2} = 0.37$	$\frac{1+0.26}{2} = 0.63$

Note. Example of data "doubling" via the "Escofier transform" for continous data with the volumetric estimates for (structural) brain imaging as an example.

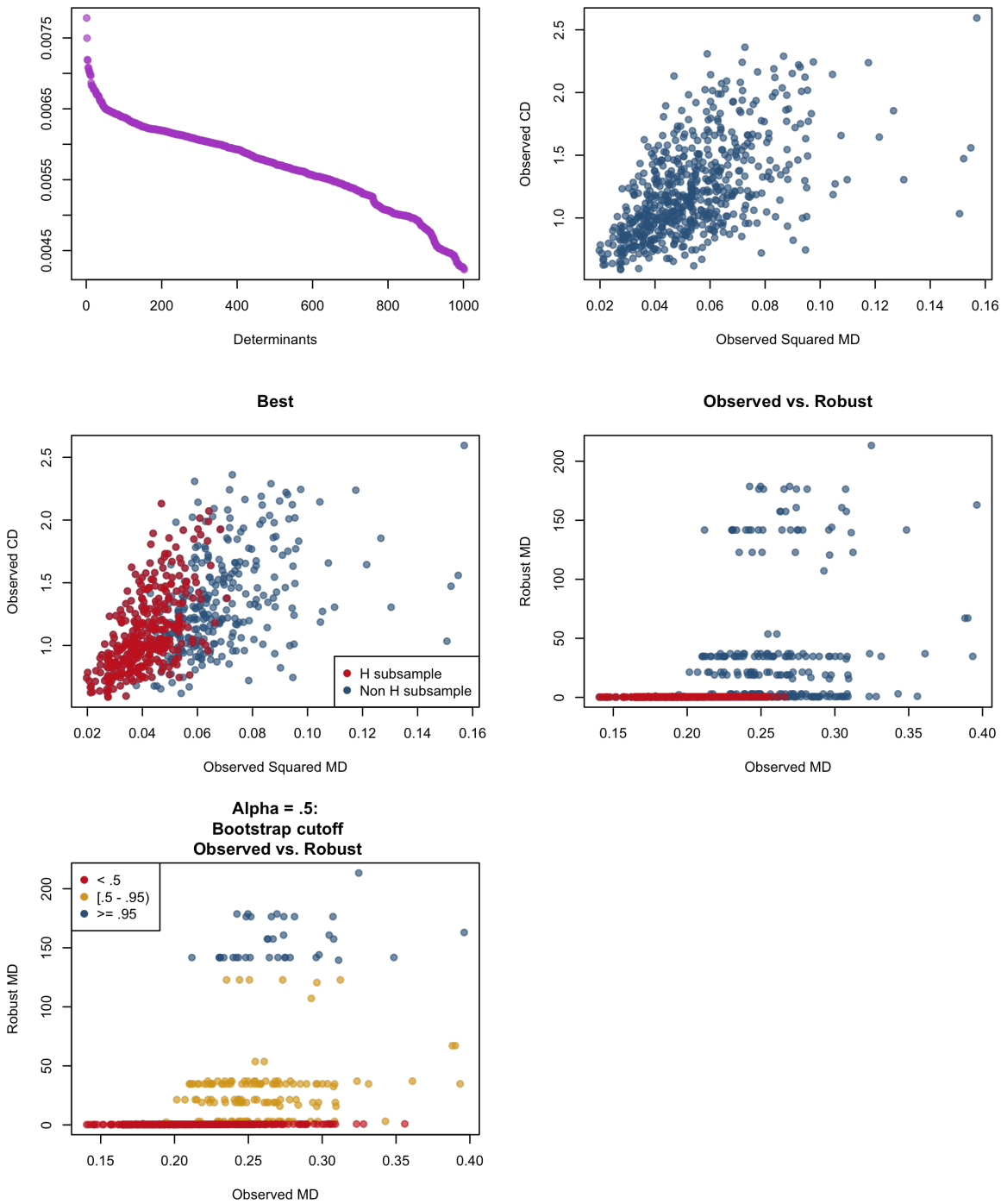

Figure 4. Application of generalized MCD to mixed data. Data are comprised of SNPs (categorical), clinical dementia rating (ordinal) and volumetric brain estimates (continuous) from ADNI. (a) shows the search for the determinant. (b) shows the observed distances. (c) shows the observed distances with the H subset colored in red. (d) shows the observed vs. robust squared Mahalanobis distances (MDs). (e) shows the same as (d) but colored according to outlier/inlier thresholds from our bootstrap procedure.

6 Discussion

We presented the first MCD-based approach that generalizes to virtually any data type (i.e., categorical, ordinal, continuous). Our approach relies entirely on Mahalnobis distances derived from CA via the SVD and quite importantly avoids many of the issues of Mahalanobis-like distances that Goodall (1966) noted (i.e., excessive weight or arbitrary metrics). Furthermore, because our generalized MCD relies on CA we can use some simple recoding approaches that allow data tables to behave as disjunctive tables (see Tables 2 and 6). Our approach 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) identifies a robust substructure. Also like many other approaches, our generalized MCD requires that we identify outliers in a way that is more reliable than simply using the H subset. Here we have opted to use a bootstrap procedure; this bootstrap procedure is easily adapted, can accomodate any threshold, and provides frequency estimates of outlierness (or inlierness) per observation.

Our generalized MCD was designed around the necessity to identify robust structures and outliers in complex non-quantitative data. Within the ONDRI study we had realized that many if not all existing MCD-like and related techniques could not accomodate our needs. We required a multivariate outlier technique that could work on categorical, ordinal, or mixed data with ease. While there are other possible algorithms (e.g., Candès et al., 2011; Fan et al., 2013) we chose to extend the MCD because it is a reliable and widely-used technique (Rousseeuw & Van Driessen, 1999 has been cited $1,740$ times at the time we wrote this paper) 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 software available in R as part of a package focused on outliers and robust structures (see https://github.com/derekbeaton/ours).

6.1 Generalized MCD beyond our examples

We established a generalized across data types MCD. In most practical and simple cases, data matrices typically consist of one *type* of data. However with studies such as ONDRI and ADNI, the boundaries of what constitutes a "data set" are blurry: researchers analyze multiple variables from a variety of sources. Often these variables are of different types. Our example with the ADNI data highlights a typical data set that stems from these types of studies. However in the ADNI example we treated each variable as an independent measure which was somewhat unrealistic: each individual variable was part of a larger related set of variables (e.g., SNPs, the CDR, and brain imaging). A family of techniques have been designed specifically to handle such data sets: Multiple Factor Analysis (Abdi, Williams, & Valentin, 2013; MFA; Escofier & Pages, 1994). The MFA technique was designed for many variables that stem from specific data partitions measured on the same observations. Furthermore, the MFA technique has been extended based on the work by Escofier (1979) to accomodate mixed data types (Bécue-Bertaut & Pagès, 2008). Our GMCD approach does not only apply for single data tables but also for MFA-type problems.

Our approach also inherently allows for *a priori* defined covariance structures à la Eq. (24). GCA, as defined by Escofier (1983) and Escofier (1984), allows for some known sets of weights (e.g., $\mathbb{W}_I$) or expected values (e.g., $\mathbb{E}_{\mathbf{X}}$). Because our generalized MCD is based on GCA, we do not depend on the $\bar{H}$ sample set to 0s in the iterative process. Rather, if some known weights and/or some known structure exists then we can use those to identify the most robust set and thus outliers from the observed values.

Furthmore our approach stems broadly from CA but was presented here as MCA. In MCA the typical assumptions are that the columns comprise all levels of all variables, the rows are observations, and that generally all observations are assigned an equal weight. However in standard CA rows do not have equal weights. In standard CA the row weights much like the column weights are proportional to all elements in that row (*cf.* Eq. (10)). While our approach would still work for data with various weights assigned to the rows we

have not addressed this. Because if there are different weights per observation, then there is no simple scaling factor between $\mathbf{U}$ and $\mathbf{P}$. Rather, $\mathbf{P}$ is a weighted version of $\mathbf{U}$ and each provides a different estimate of distance. Whether $\mathbf{U}$ or $\mathbf{P}$ are more appropriate to compute Mahalanobis-distances (when row weights differ) is an open question.

Finally our approach is based on MCA. However in MCA it is standard to discard many of the low-variance components (Abdi & Valentin, 2007). The removal of low-variance components stems from the fact that the number of components is actually an overestimate of the subspace and variance because the number of columns is greater than the actual number of variables. The "Benzécri correction" is the most common approach to identify which low-variance components to discard (Benzécri, 1979). Note that in MCA it is easy to find the rank of the data and the actual components associated with the data: With C nominal variables the subspace is defined by components with eigenvalues greater than $1/C$. In our GMCD approach we use the full set of components and thus rely on data that are full rank and where $I > J$, not just $I > C$. However the notion that many low-variance components can be discarded suggests a simple approach for a high-dimensional generalized MCD.

6.2 Limitations

While our approach opens many avenues to extend the MCD there are some limitations.

First, the "Benzécri correction" (Benzécri, 1979) discards low-variance components and thus we can compute a Mahalanobis-like distance from some arbitrary number of high-variance components this does not work well in practice. In most cases outliers typically express high values on low-variance components. If we were to discard components we might not identify a proper robust structure and thus outliers. However the dimensionality correction (Benzécri, 1979) of the space is only established for strictly categorical data because the "doubling" approaches alone provide correction dimensionality estimates

(Escofier, 1979). How to determine, and then use, the correct dimensionality of the space is also an open question. For now we have opted for the more conservative approach (require full rank data and retain all components).

Finally our algorithm does not include one particular feature of the standard MCD algorithm: we are not able to compute a robust center on the data. Rather, our robust center exists entirely within the χ^2-space as defined by the factor scores 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 with continuous data (recoded as in Table 6) and compare against the standard MCD we would obtain slightly different results.

6.3 Conclusions

The MCD is a very reliable and widely-used approach to identify robust multivariate structures and likely outliers. However until now the MCD could only be used for data that are (or were assumed to be) continuous. Our approach uses the correspondence analysis to define Mahalanobis distances (via the singular vectors) and allows us to do so for 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 complex studies that contain many mixed-variables, techniques like our generalized MCD will be essential to maintain data quality and integrity.

6.4 Acknowledgements

6.4.1 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 partly supported by a Canadian Institutes of Health Research grant (MOP 201403).

6.4.2 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.

6.4.3 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.

7 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. (2017). *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. doi: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.

Beaton, D., Fatt, C. R. C., & Abdi, H. (2014). An ExPosition of multivariate analysis with the singular value decomposition in R. *Computational Statistics & Data Analysis*, *72*(0), 176–189. doi:http://dx.doi.org/10.1016/j.csda.2013.11.006

Bedrick, E. J., Lapidus, J., & Powell, J. F. (2000). Estimating the Mahalanobis Distance from Mixed Continuous and Discrete Data. *Biometrics*, *56*(2), 394–401. doi: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. Retrieved from http://epubs.siam.org/doi/abs/10.1137/1.9781611972788.22

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. doi: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.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM.

Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In *Breakthroughs in statistics* (pp. 569–593). Springer.

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.

Goodall, D. W. (1966). A New Similarity Index Based on Probability. *Biometrics*, *22*(4), 882–907. doi: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., & Hastie, T. (1987). The Geometric Interpretation of Correspondence Analysis. *Journal of the American Statistical Association*, *82*(398), 437–447. doi:10.2307/2289445

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.

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. doi: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. doi: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. doi: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. doi:10.1016/j.jmva.2003.08.006

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

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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. doi: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.

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.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

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. doi: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.