# Generalization of the minimum covariance determinant algorithm for categorical and mixed data types[☆]

Derek Beaton[a,*], Kelly M. Sunderland[a], ADNI[b], Brian Levine[a,d,c], Jennifer Mandzia[e], Mario Masellis[d,f], Richard H. Swartz[d,f], Angela K. Troyer[c,g], ONDRI[h], Malcolm A. Binns[i,a], Hervé Abdi[j], Stephen C. Strother[k,a]

[a] *Rotman Research Institute at Baycrest Health Sciences*

[b] *Alzheimer's Disease Neuroimaging Initiative*

[c] *Department of Psychology, University of Toronto*

[d] *Department Medicine (Neurology), University of Toronto*

[e] *Department of Clinical Neurological Sciences, Western University*

[f] *Sunnybrook Health Sciences Centre*

[g] *Baycrest Health Sciences*

[h] *Ontario Neurodegenerative Disease Research Initiative*

[i] *Dalla Lana School of Public Health, University of Toronto*

[j] *School of Brain and Behavioral Sciences, The University of Texas at Dallas*

[k] *Department of Biophysics, University of Toronto*

## Abstract

The minimum covariance determinant (MCD) approach is one of the most common techniques to detect anomalous or outlying observations. The MCD approach depends on two features of multivariate statistics: the determinant of a matrix and Mahalanobis distances (MD). While the MCD algorithm is

[☆] 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 here.

[*] Corresponding author

*Email address:* dbeaton@research.baycrest.org (Derek Beaton)

commonly used, and has many extensions, the MCD is limited to analyses of quantitative—generally metric, or presumed continuous—data. The MCD does not extend to other data types such as categorical or ordinal data because MD is not strictly defined for data types other than continuous data. Here we present a generalization of the MCD: first on categorical data, and then we show how our generalized MCD (GMCD) extends beyond categorical data to other data types (e.g., ordinal), and even mixed data types (e.g., categorical, ordinal, and continuous). To do so, the GMCD relies on a multivariate technique called correspondence analysis (CA). Through CA we can define MD by way of the singular vectors and we can compute the determinant from CA's eigenvalues. We illustrate the GMCD 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 available `R` code and toy data in order to illustrate our generalized MCD (https://github.com/derekbeaton/ours).

*Keywords:* Correspondence analysis, categorical data, outliers, robust, neuroinformatics, neurodegenerative disorders

---

## Introduction

The minimum covariance determinant (MCD; Hubert & Debruyne, 2010) approach is a robust estimator for multivariate location (mean) and scatter (covariance). Given a matrix of observations (rows) and variables (columns), various MCD algorithms and derivatives 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. 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 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), including in the psychological sciences (Leys, Klein, Dominicy, & Ley, 2018). Since the introduction of the Fast-MCD approach (Rousseeuw & Van Driessen, 1999), there have been many improvements and variations on the technique, such as robust PCA (Hubert, Rousseeuw, & Branden, 2005) and deterministic MCD (Hubert, Rousseeuw, & Verdonck, 2012). 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 robust structures and 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 should we apply arbitrary metric values to categorical or ordinal data (Bürkner & Vuorre, 2018). 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 assigns values. In our work here we first show that the MCD can be generalized to categorical data by defining MD under the assumptions of $\chi^2$ (i.e., independence) metrics 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 types of variables (e.g., categorical, ordinal, and/or continuous).

Our paper is outlined as follows. In *Notation and software* we provide the notation set used in this paper and the software used for this paper. 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. Then, we show how these properties of the SVD generalize in order to compute MDs for categorical data through CA. In *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 achieve the MCD for categorical data. Next in *Applications and Extensions* we use a toy (simulated) genetics data from the supplemental material of Beaton, Dunlop, and Abdi (2016) in order to formalize a categorical version of the MCD. Following that, we illustrate our newly defined MCD to identify robust structures and outliers on several real data sets from the Ontario Neurodegenerative Disease Research Initiative (ONDRI) and the Alzheimer's Disease Neuroimaging Initiative (ADNI), including particular recoding schemes that allow for the use of other data types such as ordinal or continuous via our categorical MCD, and thus our approach is a generalized MCD (GMCD) to any data type or even mixed data types (i.e.,

a matrix of categorical, ordinal, and continuous variables). 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*.

## Software and notation

We used R (Version 3.5.1; R Core Team, 2016) and the R-packages *ExPosition* (Version 2.8.23; Beaton, Fatt, & Abdi, 2014), *MASS* (Version 7.3.50; Venables & Ripley, 2002), and *papaja* (Version 0.1.0.9842; Aust & Barth, 2018) for all our analyses and to write the manuscript itself (in RMarkdown with `papaja`). We make our software, code, and some examples of the GMCD available via the *Ou*tliers and *R*obust *S*tructures (OuRS) package at https://github.com/derekbeaton/ours.

Bold uppercase letters denote matrices (e.g., $\mathbf{X}$), bold lowercase letters denote vectors (e.g., $\mathbf{x}$), and italic lowercase letters denote specific elements (e.g., $x$). Upper case italic letters denote cardinality, size, or length (e.g., $I$) where a lower case italic denotes a specific index (e.g., $i$). A generic element of $\mathbf{X}$ would be denoted as $x_{i,j}$. Common letters of varying type faces for example $\mathbf{X}$, $\mathbf{x}$, $x_{i,j}$ come from the same data struture. Vectors are assumed to be column vectors unless otherwise specified. Two matrices side-by-side denotes standard matrix multiplication (e.g., $\mathbf{XY}$), where $\odot$ denotes element-wise (Hadamard) 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. Note that for the generalized inverse: $\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$, $(\mathbf{X}^+)^+ = \mathbf{X}$, and $(\mathbf{X}^T)^+ = (\mathbf{X}^+)^T$. The diagonal operation, diag{}, when given a vector will transform it into a diagonal matrix, or when given a matrix, will extract the diagonal elements as a vector. 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, $\mathbf{F}_I$ denotes the component scores associated with the $I$

items, (2) **O** and **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 of $\mathbf{X}$.

## Determinant and Mahalanobis distances via the SVD

### *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 on the columns. The MCD requires an initialization of subset of size $H$ which is a random subsample from $\mathbf{X}$ where $\lfloor (I + J + 1)/2 \rfloor \leqslant H \leqslant I$. In general, the MCD algorithm can be described as:

1. From $\mathbf{X}_H$ compute the column-wise row vector 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 $m_i = (\mathbf{x}_i - \boldsymbol{\mu}_H)\mathbf{S}_H^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_H)^T$.
4. Set $\mathbf{X}_H$ as the subset of $H$ observations with the smallest MDs computed from Step 3.

The size of $H$ is controlled by $\alpha$ to determine the subsample size; $\alpha$ belongs to the interval 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 $\Omega$. For the $\Omega$ subset, we can obtain robust mean vector $\boldsymbol{\mu}_\Omega$, robust covariance $\mathbf{S}_\Omega = (\mathbf{X}_\Omega^T \mathbf{X}_\Omega) \times (\Omega - 1)^{-1}$, and a set of robust square MDs for each $i$ observation from $\mathbf{X}$ as $m_i = (\mathbf{x}_i - \boldsymbol{\mu}_\Omega)\mathbf{S}_\Omega^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_\Omega)^T$. The two 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 identify outliers and aids in the search for 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 is the product of the eigenvalues; but because eigenvalues could be very large or small (in CA all eigenvalues are $\leqslant 1$), it is safer to compute the determinant 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), which we discuss in more detail in subsequent sections.

*Mahalanobis distance for continuous data*

Let $\mathbf{X}$ be a centered data matrix with $I$ rows and $J$ columns. Assume that $\mathbf{X}$ is full rank on the columns where rank is $J$. First we define the sample covariance matrix of $\mathbf{X}$ as $\mathbf{S} = (\mathbf{X}^T\mathbf{X}) \times (I-1)^{-1}$. Squared MD is defined as $\mathbf{m} = (\mathbf{X}\mathbf{S}^{-1} \odot \mathbf{X})\mathbf{1}$ 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 squared MD in two ways. The first reformulation is $\mathbf{M} = \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^T$ where $\mathbf{m} = \text{diag}\{\mathbf{M}\}$. If we relax the definition of squared MD to exclude the degrees of freedom scaling factor $I - 1$, the second reformulation is:

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

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

*Principal components analysis*

Principal components analysis (PCA) is performed on a matrix $\mathbf{X}$ with $I$ rows and $J$ columns. Say that $\mathbf{X}$ is column-wise centered ("covariance PCA") or column-wise scaled (e.g., z-scores or sums of squares equal to 1; "correlation PCA"). For the following formulation we assume only a column-wise centered $\mathbf{X}$. Again assume that $\mathbf{X}$ is full rank. We can perform the PCA of $\mathbf{X}$ through the SVD as:

$$\mathbf{X} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T \text{ where:} \tag{2}$$

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

In PCA there exist two sets of component scores, one set for the rows and one set for the columns defined as $\mathbf{F}_I = \mathbf{U}\mathbf{\Delta}$ and $\mathbf{F}_J = \mathbf{V}\mathbf{\Delta}$, respectively. We can define row and column scores through projection (or rotation) as:

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

In the regression and PCA literatures 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 $\text{diag}\{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}$. This definition of leverage is equivalent to our definition of a squared MD in Eq. (1).

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 $\mathbf{F}_I$ as $\mathbf{G}_I = \mathbf{F}_I(\mathbf{F}_I^T\mathbf{F}_I)^{-1}\mathbf{F}_I^T$ where each row item's leverage is an element in $\mathbf{g}_I = \text{diag}\{\mathbf{G}_I\}$. If we expand and substitute $\mathbf{U\Delta}$ for $\mathbf{F}_I$, we see that $\mathbf{G}_I = \mathbf{U\Delta}(\mathbf{\Delta U}^T\mathbf{U\Delta})^{-1}\mathbf{\Delta U}^T = \mathbf{U\Delta\Lambda}^{-1}\mathbf{\Delta U}^T = \mathbf{UU}^T$, thus $\mathbf{g}_I = \text{diag}\{\mathbf{UU}^T\}$. The connection between leverage and squared MD can also be shown through the SVD (*cf.* Eq. (2)):

$$\begin{aligned}
\mathbf{M}' = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \\
\mathbf{U\Delta V}^T(\mathbf{V\Delta U}^T\mathbf{U\Delta V}^T)^{-1}\mathbf{V\Delta U}^T = \\
\mathbf{U\Delta V}^T(\mathbf{V\Lambda V}^T)^{-1}\mathbf{V\Delta U}^T = \\
\mathbf{U\Delta V}^T\mathbf{V\Lambda}^{-1}\mathbf{V}^T\mathbf{V\Delta U}^T = \\
\mathbf{U\Delta}\Lambda^{-1}\mathbf{\Delta U}^T = \\
\mathbf{UU}^T = \mathbf{G}_I,
\end{aligned} \tag{4}$$

where $\mathbf{V}^T = \mathbf{V}^{-1}$ then $(\mathbf{V\Lambda V}^T)^{-1} = \mathbf{V\Lambda}^{-1}\mathbf{V}^T$ and thus $\mathbf{m}' = \text{diag}\{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\} = \text{diag}\{\mathbf{UU}^T\} = \mathbf{g}_I$. The relationship established in Eq. (4) is critical to our formulation of MD for categorical data and thus a generalized MCD.

*Representation of categorical data*

Say we have an $I \times D$ matrix called $\mathbf{D}$, 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 complete disjunctive coding—as in Table 1b—where each variable is represented by a vector of length $N_d$, where $N_d$ is the number of levels or nominal values for the $d^{\text{th}}$ variable. The disjunctive form of $\mathbf{D}$ is an $I \times J$ matrix $\mathbf{X}$. The multivariate analysis akin to a PCA of categorical data, via its complete disjunctive form, is usually performed with a technique called multiple correspondence analysis.

*Multiple correspondence analysis*

Multiple correspondence analysis (MCA) generalizes both: (1) PCA to categorical data (with observations on the rows) and (2) standard correspondence analysis (CA) from two-way contingency tables to multiple variables, that is: N-way contingency tables (Abdi & Valentin, 2007; Greenacre, 1984; Greenacre & Blasius, 2006; Lebart, Morineau, & Warwick, 1984; Saporta, 2006). MCA is a SVD-based technique with specific preprocessing and requires constraints (weights) for the rows and columns in order to decompose a matrix under the assumption of independence (i.e., $\chi^2$). 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:

$$\mathbf{r} = (\mathbf{1}^T\mathbf{X}\mathbf{1})^{-1} \times \mathbf{X}\mathbf{1} \text{ and } \mathbf{c} = (\mathbf{1}^T\mathbf{X}\mathbf{1})^{-1} \times \mathbf{X}^T\mathbf{1}, \tag{5}$$

where $\mathbf{r}$ and $\mathbf{c}$ are the row and column marginal sums, respectively, divided by the total sum. In the case of purely disjunctive data, each value in $\mathbf{r}$ is identical as each row contains the same number of 1s (see Table 1b): Each element in $\mathbf{r}$ is simply the total number of columns in $\mathbf{D}$ divided by the sum of all elements of $\mathbf{X}$. Similarly, $\mathbf{c}$ contains the column sums of $\mathbf{X}$ divided by the total sum. Next, just as in Pearson's $X^2$ analyses, we compute observed ($\mathbf{O}_\mathbf{X}$) and expected ($\mathbf{E}_\mathbf{X}$) matrices of $\mathbf{X}$ as $\mathbf{O}_\mathbf{X} = (\mathbf{1}^T\mathbf{X}\mathbf{1})^{-1} \times \mathbf{X}$ and $\mathbf{E}_\mathbf{X} = \mathbf{r}\mathbf{c}^T$, respectively, and deviations computed as $\mathbf{Z}_\mathbf{X} = \mathbf{O}_\mathbf{X} - \mathbf{E}_\mathbf{X}$. MCA is performed with the generalized

Table 1: Example categorical and disjunctive data. 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 imputed as barycentric (which is the mean of the levels) or any set of values that sum to one.

(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 |

SVD (GSVD) of $\mathbf{Z_X}$:

$$\begin{aligned} \mathbf{Z_X} &= \mathbf{P\Delta Q}^T, \text{where} \\ \mathbf{P}^T\mathbf{W}_I\mathbf{P} &= \mathbf{I} = \mathbf{Q}^T\mathbf{W}_J\mathbf{Q}, \end{aligned} \tag{6}$$

and where $\mathbf{W}_I = \text{diag}\{\mathbf{r}\}^{-1}$ and $\mathbf{W}_J = \text{diag}\{\mathbf{c}\}^{-1}$. We obtain the results of the GSVD of $\mathbf{Z_X}$ through the SVD of $\tilde{\mathbf{Z}}_\mathbf{X} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{Z_X}\mathbf{W}_J^{\frac{1}{2}} = \mathbf{U\Delta V}^T$ as in Eq. (2) 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 $\tilde{\mathbf{Z}}_\mathbf{X}$ via substitution:

$$\begin{aligned} \mathbf{Z_X} = \mathbf{W}_I^{-\frac{1}{2}}\tilde{\mathbf{Z}}_\mathbf{X}\mathbf{W}_J^{-\frac{1}{2}} = \\ \mathbf{W}_I^{-\frac{1}{2}}\mathbf{U\Delta V}^T\mathbf{W}_J^{-\frac{1}{2}} = \\ (\mathbf{W}_I^{-\frac{1}{2}}\mathbf{U})\mathbf{\Delta}(\mathbf{V}^T\mathbf{W}_J^{-\frac{1}{2}}) = \\ (\mathbf{W}_I^{-\frac{1}{2}}\mathbf{U})\mathbf{\Delta}(\mathbf{W}_J^{-\frac{1}{2}}\mathbf{V})^\mathbf{T} = \\ \mathbf{P\Delta Q}^T. \end{aligned} \tag{7}$$

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

We compute MCA component scores as $\mathbf{F}_I = \mathbf{W}_I\mathbf{P\Delta}$ and $\mathbf{F}_J = \mathbf{W}_J\mathbf{Q\Delta}$ or also compute MCA component scores through projections (or rotations; *cf.* Eq. (3)). To compute MCA component scores through projections, it is easier to work with what are called "profile" matrices where $\mathbf{\Phi}_I = \mathbf{W}_I\mathbf{O_X}$ are the row profiles and $\mathbf{\Phi}_J = \mathbf{W}_J\mathbf{O}_\mathbf{X}^T$ are the column profiles. For profile matrices, each element in $\mathbf{X}$ divided by its respective row or column sum (for the row and column profile

matrices, respectively). MCA row component scores via projection are:

$$\begin{aligned}
\mathbf{F}_I = \mathbf{\Phi}_I\mathbf{F}_J\mathbf{\Delta}^{-1} = \\
\mathbf{\Phi}_I\mathbf{W}_J\mathbf{Q}\mathbf{\Delta}\mathbf{\Delta}^{-1} = \\
\mathbf{\Phi}_I\mathbf{W}_J\mathbf{Q}\mathbf{I} = \\
\mathbf{\Phi}_I\mathbf{W}_J\mathbf{Q} = \\
\mathbf{\Phi}_I\mathbf{W}_J\mathbf{W}_J^{-\frac{1}{2}}\mathbf{V} = \\
\mathbf{\Phi}_I\mathbf{W}_J^{\frac{1}{2}}\mathbf{V} = \\
\mathbf{W}_I\mathbf{O}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}\mathbf{V},
\end{aligned} \tag{8}$$

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

*Mahalanobis distance 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., $\text{diag}\{\mathbf{U}\mathbf{U}^T\}$). 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 similarly we define squared MD for categorical data via MCA.

As noted in Eqs. (6) and (7) MCA is performed as GSVD($\mathbf{W}_I, \mathbf{Z}_\mathbf{X}, \mathbf{W}_J$) wherein we apply the SVD as $\tilde{\mathbf{Z}}_\mathbf{X} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T$. Recall the relationship between $\tilde{\mathbf{Z}}_\mathbf{X}$ and $\mathbf{Z}_\mathbf{X}$: $\tilde{\mathbf{Z}}_\mathbf{X} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{Z}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}} \Longleftrightarrow \mathbf{Z}_\mathbf{X} = \mathbf{W}_I^{-\frac{1}{2}}\tilde{\mathbf{Z}}_\mathbf{X}\mathbf{W}_J^{-\frac{1}{2}}$ where $\mathbf{Z}_\mathbf{X} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T$ (see Eq. (7)). In the case of MCA, $\tilde{\mathbf{Z}}_\mathbf{X}$ would be the analog of $\mathbf{X}$ in PCA (see Eq. (2)). Thus in MCA $\mathbf{U}\mathbf{U}^T = \tilde{\mathbf{Z}}_\mathbf{X}(\tilde{\mathbf{Z}}_\mathbf{X}^T\tilde{\mathbf{Z}}_\mathbf{X})^+\tilde{\mathbf{Z}}_\mathbf{X}^T$. Note that here we replace the standard inversion with the pseudo-inverse because $(\tilde{\mathbf{Z}}_\mathbf{X}^T\tilde{\mathbf{Z}}_\mathbf{X})$ is a "group" matrix, where all

columns that represent a single variable are collinear. We have shown previously for continuous data via PCA that leverage is squared MD (cf. Eq. 4), and applying the same principles to categorical data via MCA, we see that

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

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 we define squared MD for categorical data as $\mathbf{M'} = \tilde{\mathbf{Z}}_\mathbf{X}(\tilde{\mathbf{Z}}_\mathbf{X}^T\tilde{\mathbf{Z}}_\mathbf{X})^{+}\tilde{\mathbf{Z}}_\mathbf{X}^T = \mathbf{Z_X}(\mathbf{Z_X^T Z_X})^{+}\mathbf{Z_X^T} = \mathbf{UU}^T$ à la Eqs. (1) and (4). Finally and most importantly, because of this 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., $\mathbf{U}$).

## MCD algorithm for categorical data

Recall that the MCD algorithm requires two essential parts: the determinant (via the eigenvalues) and squared MD. 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 with respect to the mean and covariance of a sub-sample. We show how to do so via the SVD: first with continuous data in PCA to establish squared MD for excluded samples, and next how to do so with categorical data via MCA. However, categorical data pose a unique problem: nominal levels of variables (i.e., categories) are not guaranteed to exist in particular sub-samples, so we require a particular solution to obtain accurate distances—especially for squared MD and robust squared MD—of excluded or "supplementary" observations. In the CA and PCA literatures "supplementary" data are excluded or new observations we want to project onto existing components, where components were derived from an "active" set of data. For the MCD, we will consider "active" data those observations used to compute the mean and covariance matrix. The out of sample data will be referred to as supplementary.

*Active and supplementary Mahalanobis for continuous data*

In order to establish squared MD for sub- and out-of sample data, we must revisit squared MD through projections à la Eq. (3). 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: $\mathbf{X}\mathbf{V}\boldsymbol{\Delta}^{-1} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T\mathbf{V}\boldsymbol{\Delta}^{-1} = \mathbf{U}\boldsymbol{\Delta}\boldsymbol{\Delta}^{-1} = \mathbf{U}$. Let us refer to a $K \times J$ matrix $\mathbf{Y}$ as a supplementary 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 with the same center and scaling 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. (3) to compute supplementary component scores for $\mathbf{Y}$:

$$\mathbf{F}_K = \mathbf{Y}\mathbf{F}_J\boldsymbol{\Delta}^{-1} = \mathbf{Y}\mathbf{V}\boldsymbol{\Delta}\boldsymbol{\Delta}^{-1} = \mathbf{Y}\mathbf{V}\mathbf{I} = \mathbf{Y}\mathbf{V}. \tag{10}$$

We can compute the squared MDs for $\mathbf{Y}$ with only $\mathbf{V}$ and $\boldsymbol{\Delta}$ from the SVD

of $\mathbf{X}$ (*cf.* Eq. (2))and through projections (*cf.* Eq. (1)) instead of $\mathbf{M}'_Y = \mathbf{Y}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y}^T$ where $\mathbf{m}'_Y = \text{diag}\{\mathbf{M}'\}$, where the squared MDs of $\mathbf{Y}$ are $\mathbf{m}'_Y = \text{diag}\{\mathbf{F}_K\mathbf{\Lambda}^{-1}\mathbf{F}_K^T\}$ because

$$\begin{aligned}
\mathbf{M}'_Y = \mathbf{Y}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y}^T = \\
\mathbf{Y}(\mathbf{V\Delta U}^T\mathbf{U\Delta V}^T)^{-1}\mathbf{Y}^T = \\
\mathbf{Y}(\mathbf{V\Delta I\Delta V}^T)^{-1}\mathbf{Y}^T = \\
\mathbf{Y}(\mathbf{V\Lambda V}^T)^{-1}\mathbf{Y}^T = \\
\mathbf{Y}(\mathbf{V\Lambda}^{-1}\mathbf{V}^T)\mathbf{Y}^T = \\
\mathbf{YV\Lambda}^{-1}\mathbf{V}^T\mathbf{Y}^T = \\
\mathbf{F}_K\mathbf{\Lambda}^{-1}\mathbf{F}_K^T.
\end{aligned} \tag{11}$$

*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 $\mathbf{Y_X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y_X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{UU}^T$, 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 thus we can compute the squared MDs for active and supplementary sub-samples through the SVD and projections as $\mathbf{m'_Y} = \text{diag}\left\{\begin{bmatrix} \mathbf{UU}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{ZV\Lambda}^{-1}\mathbf{V}^T\mathbf{Z}^T \end{bmatrix}\right\}$.

*Active and supplementary Mahalanobis distances for categorical data*

We can compute Mahalanobis distances for active data through projections for categorical data. We can obtain the component scores for the active sub-sample

in MCA through Eq. (8) where $\mathbf{F}_I = \mathbf{\Phi}_I \mathbf{W}_J^{\frac{1}{2}} \mathbf{V} = \mathbf{W}_I \mathbf{O}_\mathbf{X} \mathbf{W}_J^{\frac{1}{2}} \mathbf{V} = \mathbf{W}_I \mathbf{P} \mathbf{\Delta} = \mathbf{W}_I \mathbf{W}_I^{-\frac{1}{2}} \mathbf{U} \mathbf{\Delta} = \mathbf{W}_I^{\frac{1}{2}} \mathbf{U} \mathbf{\Delta}$. To obtain the squared MD through projections:

$$\begin{aligned} \mathbf{U} = \mathbf{W}_I^{-\frac{1}{2}} \mathbf{W}_I^{\frac{1}{2}} \mathbf{U} \mathbf{\Delta} \mathbf{\Delta}^{-1} = \mathbf{I} \mathbf{U} \mathbf{I} \text{ and} \\ \mathbf{U} = \mathbf{W}_I^{-\frac{1}{2}} \mathbf{W}_I \mathbf{O}_\mathbf{X} \mathbf{W}_J^{\frac{1}{2}} \mathbf{V} \mathbf{\Delta}^{-1} = \mathbf{W}_I^{\frac{1}{2}} \mathbf{O}_\mathbf{X} \mathbf{W}_J^{\frac{1}{2}} \mathbf{V} \mathbf{\Delta}^{-1}. \end{aligned} \tag{12}$$

The projection described in Eq. (12) works because of the relationship between GSVD$(\mathbf{W}_I, \mathbf{Z}_\mathbf{X}, \mathbf{W}_J)$ and GSVD$(\mathbf{I}, (\mathbf{W}_I^{\frac{1}{2}} \mathbf{O}_\mathbf{X} \mathbf{W}_J^{\frac{1}{2}}), \mathbf{I})$. The standard GSVD for MCA, GSVD$(\mathbf{W}_I, \mathbf{Z}_\mathbf{X}, \mathbf{W}_J)$, is equivalent to GSVD$(\mathbf{I}, (\mathbf{W}_I^{\frac{1}{2}} \mathbf{O}_\mathbf{X} \mathbf{W}_J^{\frac{1}{2}}), \mathbf{I})$ with one slight exception: the first GSVD expression is already centered whereas the second GSVD expression is not, so the first singular value in GSVD$(\mathbf{I}, (\mathbf{W}_I^{\frac{1}{2}} \mathbf{O}_\mathbf{X} \mathbf{W}_J^{\frac{1}{2}}), \mathbf{I})$ is 1 and the first right and left singular vectors are trivial (equal to the marginal probabilities via $\mathbf{r}$ and $\mathbf{c}$). 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 mean. 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. (8) and Eq. (12) suggest ways to compute squared MD for *supplementary* (a.k.a. sub- or out-of sample) categorical data, there is a major barrier that prevents the computation of *correct* squared MD for supplementary categorical data: if complete disjunctive coding were applied to each subsample, then subsamples will not necessarily have the same columns (e.g., rare nominal levels may not exist in some subsamples). 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. To do so, we require a specific form of MCA called subset MCA (Greenacre, 2017; Greenacre & Pardo, 2006), which is a specific case of GCA (Escofier, 1983, 1984).

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

Hiding in the broader CA literature is "generalized correspondence analysis" (Escofier, 1983, 1984)—as referred to by Grassi and Visentin (1994). Escofier

established the idea of GCA as: "(DATA - MODEL) / MARGINS". GCA has many uses well beyond the standard applications in CA and MCA, 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).

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

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

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. 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 $\text{GSVD}(\mathbb{W}_I, \mathbb{Z}_{\mathbf{X}}, \mathbb{W}_J)$ (see Eqs. (6) and (7)). In this generalized MCA, the expected values and weights can be derived from any prespecified models or data that conform to the rows and columns of the given data.

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 $_{\overline{H}}$. First we compute all matrices required for MCA exactly as initially defined in Eqs. (6) and (7): $\mathbf{O}_{\mathbf{X}}$, $\mathbf{E}_{\mathbf{X}}$, $\mathbf{r}$ and $\mathbf{c}$. Let us frame $\mathbf{O}_{\mathbf{X}}$, $\mathbf{E}_{\mathbf{X}}$, $\mathbf{r}$ as constructed by the $H$ and $\bar{H}$ subsets as: $\mathbf{O}_{\mathbf{X}} = \begin{bmatrix} \mathbf{O}_{\mathbf{X}_H} \\ \mathbf{O}_{\mathbf{X}_{\overline{H}}} \end{bmatrix}$, $\mathbf{E}_{\mathbf{X}} = \begin{bmatrix} \mathbf{E}_{\mathbf{X}_H} \\ \mathbf{E}_{\mathbf{X}_{\overline{H}}} \end{bmatrix}$, and $\mathbf{r} = \begin{bmatrix} \mathbf{r}_H \\ \mathbf{r}_{\overline{H}} \end{bmatrix}$. To maintain the structure of the data and to be able to perform the necessary steps for MCD we require the generalized expected matrix of $\mathbb{E} = \begin{bmatrix} \mathbf{E}_{\mathbf{X}_H} \\ \mathbf{O}_{\mathbf{X}_{\overline{H}}} \end{bmatrix}$. That is, we replace part of the expected matrix with part of our observed matrix. Thus our assumption is that

the expected values are the observed and the deviation 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 subsample of $H$ and $\bar{H}$ guarantee that we maintain the same shape (rows and columns) as $\mathbf{X}$ when we use $\mathbb{E} = \begin{bmatrix} \mathbf{E}_{\mathbf{X}_H} \\ \mathbf{O}_{\mathbf{X}_{\bar{H}}} \end{bmatrix}$. 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. (6) and (7) we have $\begin{bmatrix} \mathbf{Z}_{\mathbf{X}_H} \\ \mathbf{0} \end{bmatrix} = \mathbb{Z}_\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T = \begin{bmatrix} \mathbf{P}_H \\ \mathbf{0} \end{bmatrix} \boldsymbol{\Delta}\mathbf{Q}^T$, where $\mathbf{P}^T\mathbf{W}_I\mathbf{P} = \mathbf{I} = \mathbf{Q}^T\mathbf{W}_J\mathbf{Q}$ and $\mathbf{P}^T\mathbf{W}_I\mathbf{P} = \begin{bmatrix} \mathbf{P}_H \\ \mathbf{0} \end{bmatrix}^T \mathbf{W}_I \begin{bmatrix} \mathbf{P}_H \\ \mathbf{0} \end{bmatrix} = \mathbf{I}$. Furthermore, recall that $\tilde{\mathbb{Z}}_\mathbf{X} = \mathbf{W}_I^{\frac{1}{2}}\mathbb{Z}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}}$ and thus $\mathbf{W}_I^{\frac{1}{2}}\mathbb{Z}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}} = \mathbf{W}_I^{\frac{1}{2}} \begin{bmatrix} \mathbf{Z}_{\mathbf{X}_H} \\ \mathbf{0} \end{bmatrix} \mathbf{W}_J^{\frac{1}{2}} = \begin{bmatrix} \tilde{\mathbf{Z}}_{\mathbf{X}_H} \\ \mathbf{0} \end{bmatrix} = \tilde{\mathbb{Z}}_\mathbf{X}$. Just as in Eq. (7) we would apply the SVD as:

$$\tilde{\mathbb{Z}}_\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T = \begin{bmatrix} \mathbf{U}_H \\ \mathbf{0} \end{bmatrix} \boldsymbol{\Delta}\mathbf{V}^T, \tag{13}$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{I} = \mathbf{V}^T\mathbf{V}$ and $\mathbf{U}^T\mathbf{U} = \begin{bmatrix} \mathbf{U}_H \\ \mathbf{0} \end{bmatrix}^T \begin{bmatrix} \mathbf{U}_H \\ \mathbf{0} \end{bmatrix} = \mathbf{U}_H^T\mathbf{U}_H = \mathbf{I}$. We can simplify this approach through a specific application of GCA called "subset MCA" (Greenacre, 2017; Greenacre & Pardo, 2006).

As seen in Eq. (13), the expected values of the excluded subset $\bar{H}$ are 0 so that we only analyze the active subset $H$. This case of GCA reduces to subset MCA. For subset MCA, when we compute the SVD of $\tilde{\mathbb{Z}}$ we only require all of the observations that are not 0. To compute subset MCA for use in the MCD we use the matrices defined in Eqs. (6) and (7). Specifically $\tilde{\mathbf{Z}}_\mathbf{X} = \mathbf{W}_I^{\frac{1}{2}}\mathbf{Z}_\mathbf{X}\mathbf{W}_J^{\frac{1}{2}} \iff \mathbf{Z}_\mathbf{X} = \mathbf{W}_I^{-\frac{1}{2}}\tilde{\mathbf{Z}}_\mathbf{X}\mathbf{W}_J^{-\frac{1}{2}}$. In subset MCA we require $\tilde{\mathbf{Z}}_{\mathbf{X}_H}$ which is the $H$ subsample

of $\tilde{\mathbf{Z}}_{\mathbf{X}}$ and perform the SVD as

$$\tilde{\mathbf{Z}}_{\mathbf{X}_H} = \mathbf{U}_H \boldsymbol{\Delta}_H \mathbf{V}_H^T, \tag{14}$$

which leaves us with the complement $\bar{H}$ where $\tilde{\mathbf{Z}}_{\mathbf{X}_{\overline{H}}} = \mathbf{U}_{\overline{H}} \boldsymbol{\Delta}_{\overline{H}} \mathbf{V}_{\overline{H}}^T$. In the CA literature the sum of the eigenvalues (i.e., diag$\{\boldsymbol{\Lambda}\}$) is called "inertia" and is denoted as $\mathcal{I}$. In effect, subset MCA partitions the full MCA space into two additive subspaces: $\mathcal{I} = \mathcal{I}_H + \mathcal{I}_{\overline{H}}$.

We can compute the determinant for $H$ subsamples via the geometric mean of the eigenvalues ($\boldsymbol{\Lambda}_H = \boldsymbol{\Delta}_H^2$) and we can compute robust squared MDs in subset MCA through the row profiles. Recall that we can compute component scores and squared MDs from the full MCA via projection as seen in Eqs. (8) and (12). We can also compute component scores for the full $I$ set from the subset MCA as

$$\begin{bmatrix} \mathbf{F}_H \\ \hat{\mathbf{F}}_{\overline{H}} \end{bmatrix} = \boldsymbol{\Phi}_I \mathbf{F}_{J_H} \boldsymbol{\Delta}_H^{-1} \tag{15}$$

which allows us to compute $\mathbf{U} = \mathbf{W}_I^{-\frac{1}{2}} \boldsymbol{\Phi}_I \mathbf{F}_J \boldsymbol{\Lambda}^{-1}$ because

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

We can compute *supplementary* squared MDs for the full set of observations

conditional to the subset MCA results as:

$$\begin{bmatrix} \mathbf{U}_H \\ \hat{\mathbf{U}}_{\overline{H}} \end{bmatrix} = \mathbf{W}_I^{-\frac{1}{2}} \mathbf{\Phi}_I \mathbf{F}_{J_H} \mathbf{\Lambda}_H^{-1} \tag{16}$$

where the $\mathbf{U}_H$ in Eq. (16) is exactly the $\mathbf{U}_H$ in Eqs. (13) and (14). Alternatively we can compute Eq. (16) from (15) as $\begin{bmatrix} \mathbf{U}_H \\ \hat{\mathbf{U}}_{\overline{H}} \end{bmatrix} = \mathbf{W}_I^{-\frac{1}{2}} \begin{bmatrix} \mathbf{F}_H \\ \hat{\mathbf{F}}_{\overline{H}} \end{bmatrix} \mathbf{\Delta}^{-1}$. Finally, the robust squared MDs are computed as $\text{diag}\{ \begin{bmatrix} \mathbf{U}_H \\ \hat{\mathbf{U}}_{\overline{H}} \end{bmatrix} \begin{bmatrix} \mathbf{U}_H \\ \hat{\mathbf{U}}_{\overline{H}} \end{bmatrix}^T \}$.

*MCD algorithm for categorical data*

Here we outline the core steps required for a MCD via subset MCA. First the MCD requires an initial random subsample of size $H$ where $\lfloor (I + J + 1)/2 \rfloor \leqslant H \leqslant I$. The $H$ subset of $\tilde{\mathbf{Z}}_{\mathbf{X}}$ is denoted as $\tilde{\mathbf{Z}}_{\mathbf{X}_H}$.

1. Apply the SVD: $\tilde{\mathbf{Z}}_{\mathbf{X}_H} = \mathbf{U}_H \mathbf{\Delta}_H \mathbf{V}_H^T$
2. Compute the determinant of $\tilde{\mathbf{Z}}_{\mathbf{X}_H}$ as the geometric mean of $\mathbf{\Lambda}_H$
3. Compute the squared MDs for the full $I$ sample from $\begin{bmatrix} \mathbf{U}_H \\ \hat{\mathbf{U}}_{\overline{H}} \end{bmatrix} = \mathbf{W}_I^{-\frac{1}{2}} \mathbf{\Phi}_I \mathbf{F}_{J_H} \mathbf{\Lambda}_H^{-1}$
4. Set $\tilde{\mathbf{Z}}_{\mathbf{X}_H}$ as the subset of $H$ observations with the smallest squared MDs computed from Step 3.

These steps are repeated until a we find a minimum determinant (Step 2) either through optimal search or limited to a given number of iterations. 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 a smaller or equal determinant in subsequent iterations.

*A note on $\chi^2$-distances*

It is important to note that a more commonly used distance exists in the CA and MCA literature: the $\chi^2$-distance (Guttman, 1941). Arguably the $\chi^2$-distance is the most well understood and defined metric for non-quantitative data and typically used for contingency and categorical data by way of CA and MCA (Greenacre, 2017; Greenacre & Hastie, 1987). Greenacre (2017) and Greenacre and Hastie (1987) say that the $\chi^2$-distance is a type of Mahalanobis distance. While the $\chi^2$-distance and Mahalanobis distance are both specific types of generalized Euclidean distances, we believe it is important to distinguish the two: we compute the $\chi^2$-distances from the component scores and we compute the MD from the singular vectors; these two distances are not equivalent in CA.

The $\chi^2$-distances for each observation from CA come from the component scores: $\text{diag}\{\mathbf{F}_I\mathbf{F}_I^T\}$ (*cf.* Eq. (15); see also Benzécri, 1973 and Escofier, 1965). The $\chi^2$-distances remain *unchanged* between subset and full MCA: $\text{diag}\left\{\begin{bmatrix}\mathbf{F}_H\\ \hat{\mathbf{F}}_{\overline{H}}\end{bmatrix}^T\begin{bmatrix}\mathbf{F}_H\\ \hat{\mathbf{F}}_{\overline{H}}\end{bmatrix}\right\} = \text{diag}\{\mathbf{F}_I^T\mathbf{F}_I\}$. Because the $\chi^2$-distances do not change they cannot be used in the MCD algorithm to identify optimal subsets. Therefore—as with the typical MCD algorithm—the GMCD relies on the MDs defined from the singular vectors for the observations: $\text{diag}\{\mathbf{U}^T\mathbf{U}\}$.

## Applications and Extensions

Now that we have established a MCD algorithm for categorical data we show examples of its use including how our approach generalizes the MCD to allow for virtually any data type. Because our approach is based 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 three data sets: (1) a small toy data set, (2) one from the Ontario Neurodegernative Disease Research Initiative (ONDRI), and (3) one from the Alzheimer's Disease Neuroimaging Initiative (ADNI). First we apply our approach to the toy genetic data—single nucleotide polymorphisms (SNPs)—which were simulated based on 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). Next we use ONDRI data, specifically the brief survey of autobiographical memory (BSAM; adapted from Palombo, Williams, Abdi, & Levine, 2013). The BSAM data are comprised of five ordinal responses to all questions, where some responses are relatively rare, and so we apply our approach to three versions of the BSAM: (1) where each question is treated as a categorical variable, (2) categorical but recoded to combine rare responses with their next comparable response (e.g., combine rare 5s with 4s), and (3) an approach that preserves the ordinality of the data. We compare and contrast the results from all three applications to the BSAM specifically to highlight how GMCD results change. Finally, we use a mixture of data types from ADNI data to demonstrate how our technique applies to data sets with mixed variables across multiple data modalities. The ADNI data include categorical data (SNPs), ordinal data (clinical dementia rating), and continuous data (volumetric estimates from brain regions). We explain each data set and applications in their own sections.

*Toy data*

The toy data are comprised of SNPs, which are categorical variables (toy data available as part of the software https://github.com/derekbeaton/ours). 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.

The toy data are of size $I = 60 \times C = 9$ (observations by variables). These data were transformed into a disjunctive coding (as in Table 1) that are of size $I = 60 \times J = 25$ (observations by columns); note that some SNPs have 3 genotypes present where as others only have 2 genotypes present. 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 where the intial determinant is quite high by comparison to later determinants and the GMCD stops (at $1,000$ iterations) on a much smaller determinant. Figure 1b shows the observed squared MD (horizontal) and $\chi^2$-distances (vertical). 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, we show the robust squared MD, which helps reveal "masking" effects: Figure 1d shows the observed vs. robust squared MD. Figure 1e is the same as Figure 1d except without the two most extreme individuals to provide a better view of the observed vs. robust squared MDs.

*ONDRI Data*

The Ontario Neurodegenerative Disease Research Initiative (ONDRI; http://ondri.ca/) is a longitudinal, multi-site, "deep-phenotyping" study across multiple neurodegenerative disease cohorts (Farhan et al., 2017): 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 strictly for illustrative purposes. As part of the quality control process, various subsets of data are subjected to outlier analyses. However,

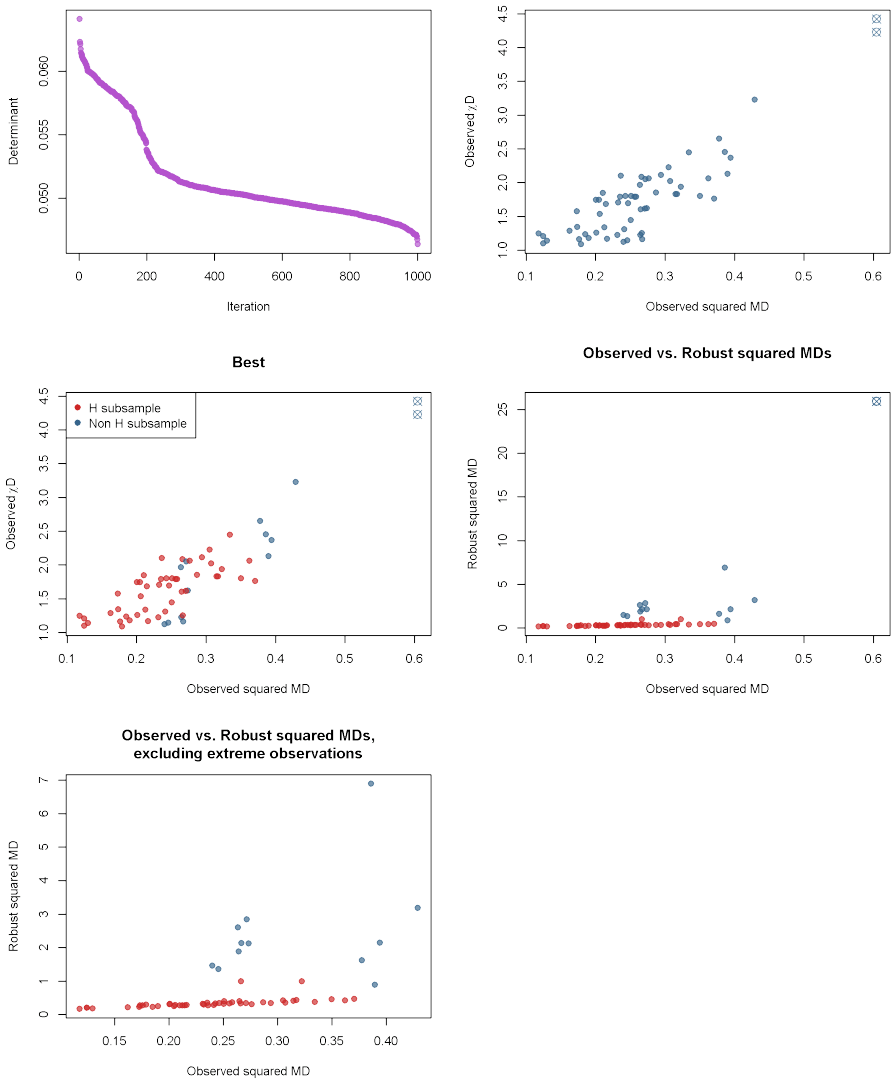


Figure 1: Application of generalized MCD to categorical data. We illustrate the generalized MCD here first on a small data set of single nucleotide polymorphisms (SNPs). (a) shows the search for the determinant. (b) shows the observed distances (via standard multiple correspondence analysis). (c) shows 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.

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). We illustrate our generalized MCD on the BSAM. We show that the GMCD of the BSAM on $N = 300$ particpants from two cohorts (VCI $n = 161$ and PD $n = 139$). This illustration compares and contrasts three ways of coding the data.

The BSAM is a short version of the survey of autobiographical memory (SAM) comprised of 10 questions where each question has five possible responses: two responses indicate disagreement with a question ("strongly disagree", "disagree somewhat"), two responses indicate agreement with a question ("agree somewhat", "strongly agree") and one neutral respone ("neither"). We analyze the BSAM 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 $\chi^2$-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 ($\lessapprox 5\%$) 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" (described below). 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 $\alpha$ levels (i.e., $\alpha = .5$,

$\alpha = .75$, $\alpha = .9$). Finally, we also illustrate how to use the bootstrap to define empirical thresholds for outliers given the robust squared MDs.

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 BSAM 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 like 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): 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.

We refer to the original BSAM data as "as is", the recoded for rare responses version "recode", and the ordinal version as "ordinal". We applied the GMCD to each with three alpha levels: $\alpha = .5$, $\alpha = .75$, $\alpha = .9$ for a total of nine applications. As in the Fast-MCD algorithm, $\alpha$ 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` (Todorov & Filzmoser, 2009) and `robustbase` (Maechler et al., 2018) packages in `R` via the `h.alpha.n` function). Because each data set has a different number of columns ($J$) the same $\alpha$ will produce a different size $H$. Each application of the analysis was run for $1,000$ iterations for the GMCD algorithm.

Table 2: Example of fuzzy coding. Example of data "doubling" via fuzzy for ordinal data with the brief survey of autobiographical memory (BSAM).

(a) CDR ordinal example

| | Q1: Specific events | Q2: Recall objects |
|---|---|---|
| *OND01_TWH_1100* | 2 | 4 |
| *OND01_SBH_1200* | 4 | 4 |
| ... | ... | ... |
| *OND01_HDH_5100* | 5 | 4 |
| *OND01_HDH_5201* | 4 | 5 |

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

| | Q1− | Q1+ | Q2− | Q2+ |
|---|---|---|---|---|
| *OND01_TWH_1100* | $\frac{2-1}{4} = \frac{1}{4}$ | $\frac{5-2}{4} = \frac{3}{4}$ | $\frac{4-1}{4} = \frac{3}{4}$ | $\frac{5-4}{4} = \frac{1}{4}$ |
| *OND01_SBH_1200* | $\frac{4-1}{4} = \frac{3}{4}$ | $\frac{5-4}{4} = \frac{1}{4}$ | $\frac{4-1}{4} = \frac{3}{4}$ | $\frac{5-4}{4} = \frac{1}{4}$ |
| ... | ... | ... | ... | ... |
| *OND01_HDH_5100* | $\frac{5-1}{4} = \frac{4}{4}$ | $\frac{5-5}{4} = \frac{0}{4}$ | $\frac{4-1}{4} = \frac{3}{4}$ | $\frac{5-4}{4} = \frac{1}{4}$ |
| *OND01_HDH_5201* | $\frac{4-1}{4} = \frac{3}{4}$ | $\frac{5-4}{4} = \frac{1}{4}$ | $\frac{5-1}{4} = \frac{4}{4}$ | $\frac{5-5}{4} = \frac{0}{4}$ |

Figure 2 shows all nine applications of the generalized MCD applied to the BSAM. To note in Figure 2 we show the MDs and robust MDs (not squared). Part of the reason we show these is because as 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 $\alpha$ 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 $\alpha$ increases. Finally we also see that the "ordinal" version of the data appears to behave quite differently from the other versions as $\alpha$ 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 $\alpha$. However the $H$ subsamples across each of these are highly similar.

The overlap of the $H$ subsample between all applications of the BSAM analyses is shown in Table 3. When $\alpha$ is small there is more variability between the different versions. As $\alpha$ 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.

Besides the identification of robust subsamples (and covariance structures), MCD algorithms reveal "masking" effects and identify outliers. In some of our BSAM 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 $\alpha$, it is clear that an outlier subset of individuals have unique response patterns that differ greatly from the rest of the sample.

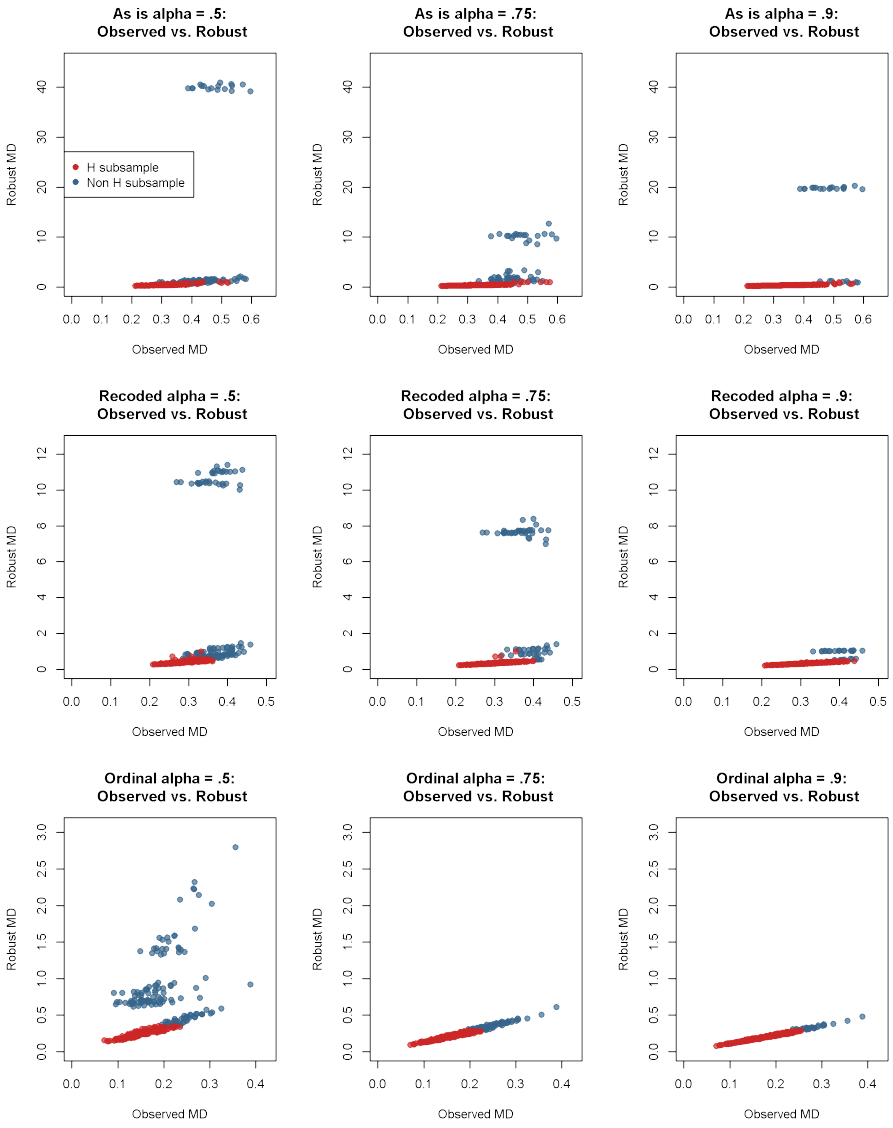


Figure 2: Generalized MCD applied to the brief survey of autobiographical memory (BSAM) data from ONDRI. Left to right is the alpha level and top to bottom is how the data were coded. All plots show observed vs. robust Mahalaniobis distances (MD). Observations in red are the H subsample, observations 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 = As is, R = Recode, O = Ordinal. Only the lower triangle and diagonal are shown. The diagonal here indicates the size of the $H$ subset (out of $N = 300$). The value in each cell is the number of overlapping subjects identified as the robust subsample for each version of the brief survey on autobiographical memory (BSAM). Generally as the $\alpha$ increases, overlap becomes more likely. As is and ordinal tend to overlap across all versions.

| | A.5 | A.75 | A.9 | R.5 | R.75 | R.9 | O.5 | O.75 | O.9 |
|---|---|---|---|---|---|---|---|---|---|
| As Is .5 | 58.33% | | | | | | | | |
| As Is .75 | 54.33% | 79.00% | | | | | | | |
| As Is .9 | 58.00% | 78.00% | 91.67% | | | | | | |
| Recode .5 | 44.67% | 49.00% | 54.00% | 57.00% | | | | | |
| Recode .75 | 52.33% | 63.33% | 73.33% | 56.00% | 78.33% | | | | |
| Recode .9 | 57.33% | 72.00% | 84.00% | 56.67% | 78.00% | 91.33% | | | |
| Ordinal .5 | 40.00% | 47.33% | 52.00% | 42.67% | 46.00% | 50.67% | 53.33% | | |
| Ordinal .75 | 54.33% | 69.00% | 74.67% | 51.00% | 63.00% | 70.00% | 52.33% | 76.67% | |
| Ordinal .9 | 56.67% | 76.33% | 87.00% | 55.33% | 72.67% | 82.67% | 53.33% | 76.67% | 90.67% |

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" BSAM at $\alpha = .5$. However other times, as in the ordinal BSAM at $\alpha = .9$, there is no clear division between the $H$ subsample and the rest of the observations. 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 outliers here as those above some threshold, and "inliers" as those individuals that are not identified as outliers.

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 BSAM 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 $\alpha$). 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 BSAM 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 $\alpha = .9$. Similar to the $H$ subsample identification, when $\alpha$ is small there is more variability between the different versions. In the fixed analyses (1) as $\alpha$ 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. For the identification of inliers and outliers we 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".

*ADNI Data*

Data used in the preparation of this article come from Phase 1 of the ADNI database (adni.loni.usc.edu). ADNI was launched in 2003 as a public-private funding partnership and includes public funding by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and the Food and Drug Administration. The primary goal of ADNI has been to test a wide variety of measures to assess the progression of mild cognitive impairment and early Alzheimer's disease. The ADNI project is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations. Michael W. Weiner (VA Medical Center, and University of California-San Francisco) is the ADNI Principal Investigator. Subjects have been recruited from over 50 sites across the United States and Canada (for up-to-date

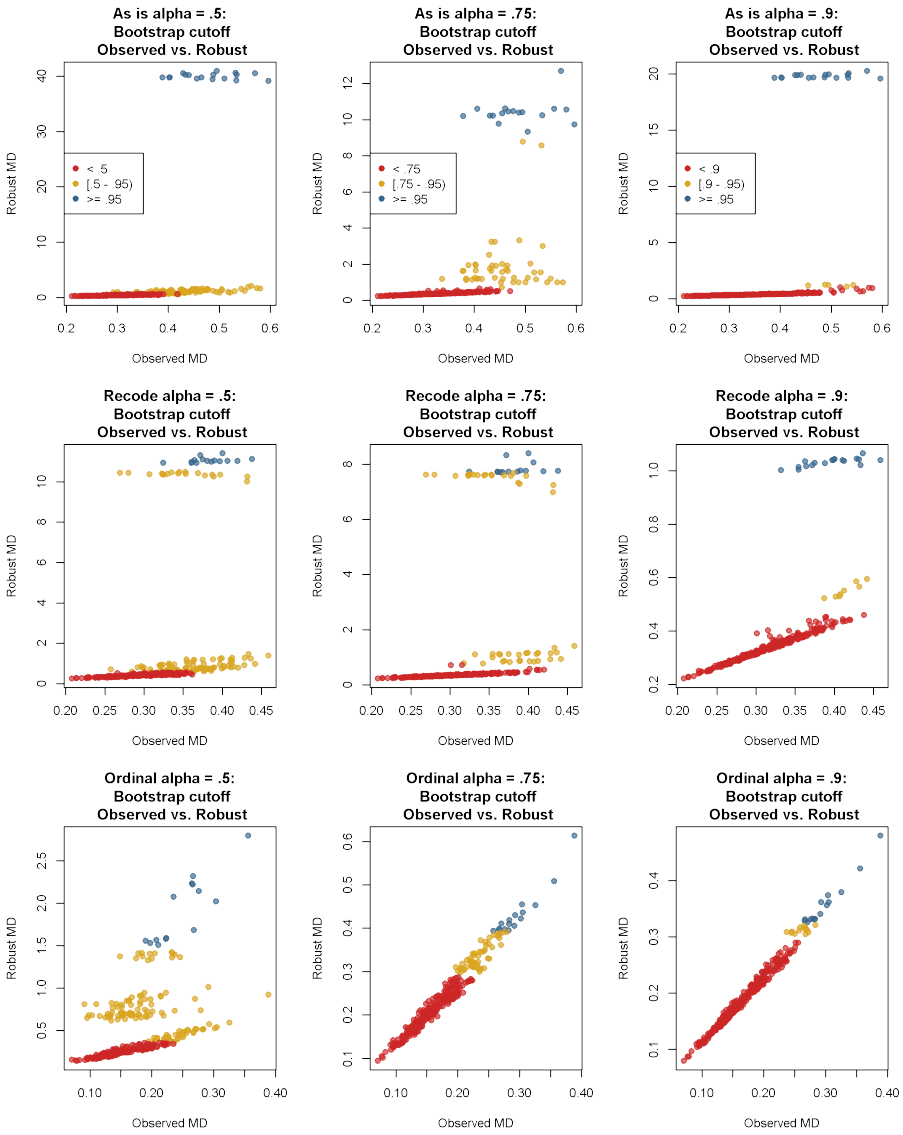


Figure 3: Bootstrapping procedure applied the generalized MCD of the brief survey of autobiographical memory (BSAM). All plots show observed vs. robust Mahalanobis distances (MD). Observations in red are the bootstrap 'inliers', observations in blue are outside the bootstrap threshold, and observations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

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. Just as with 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.

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"; a.k.a. the dominant inheritance model). We were thus left with 13 SNPs that spanned 35 columns: 4 of our SNPs were recoded to combine "Aa" and "aa" (i.e., "AA" vs. "Aa+aa") and the remaining 9 SNPs each had three levels (i.e., "AA", "Aa", and "aa").

The CDR is a structured interview that has six domains: memory, orientation, judgement, communication, home and hobbies, and self-care (Morris (1993); see also http://alzheimer.wustl.edu/cdr/cdr.htm). After the interview ratings are applied to each category with the following possible responses: 0 (normal), 0.5 (very mild), 1 (mild), 2 (moderate), and 3 (severe). The CDR has ordinal responses so 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, behave 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 introduced 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 4. 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. In Figure 4a we see the observed MD and $\chi^2$-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 4b and c, individuals in red are those within the subsample and blue are those excluded; Figure 4b shows the observed squared MD and $\chi^2$-distances for individuals (in red) that comprise the "best" determinant (the sample at convergence). Finally, as with the standard MCD algorithm we can also compute robust MD which helps revealing "masking" effects by exaggerating the more extreme individuals. Figure 4c shows the observed vs. robust MD; the robust MD was computed with respect to the individuals identified as the minimum determinant set in Figure 4b. Figure 4d shows outliers (in blue) as identified through our bootstrap procedure. As 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 4: Example of Escofier transform. Example of data "doubling" via the "Escofier transform" for continous data with the volumetric estimates for (structural) brain imaging as an example.

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

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

(b) Escofier-style doubling for continuous data

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

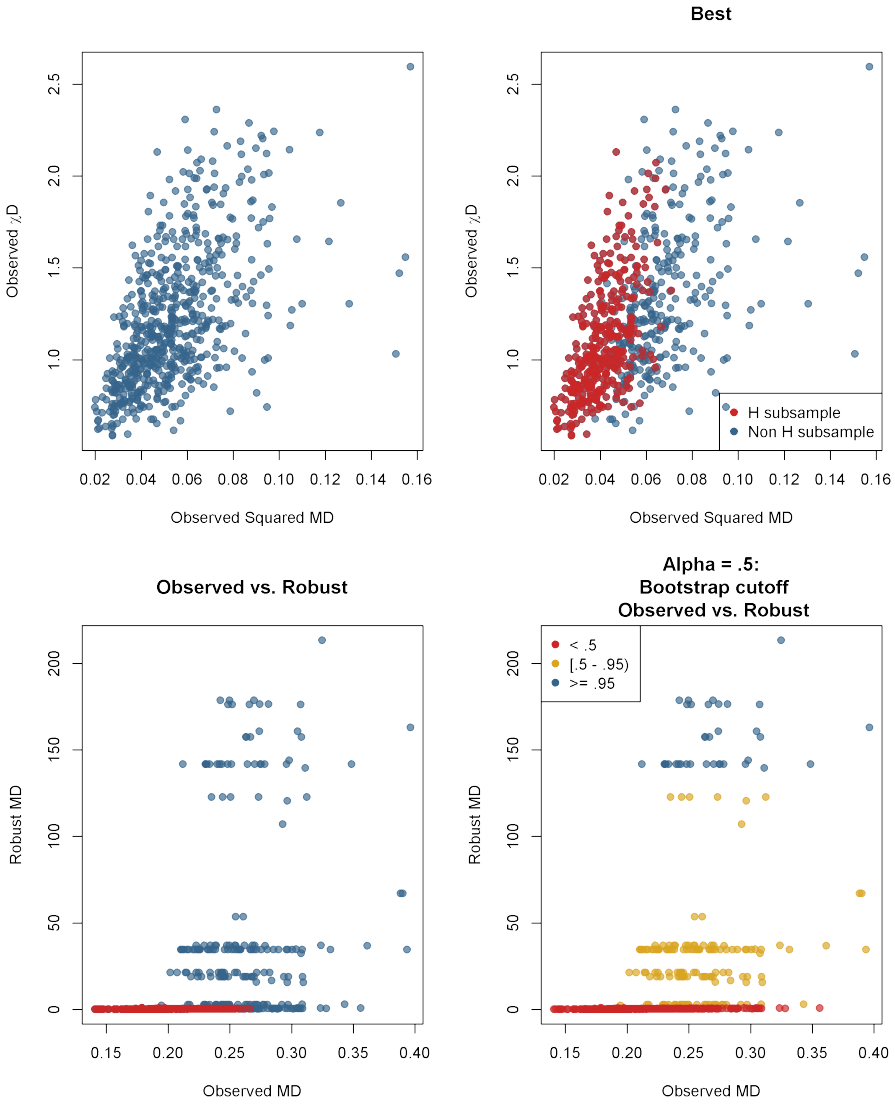


Figure 4: Application of generalized MCD to mixed data. Data are comprised of SNPs (categorical), clinical dementia rating (ordinal) and volumetric brain estimates (continuous) from ADNI. (a) shows the observed distances. (b) shows the observed distances with the $H$ subset colored in red. (c) shows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

## Discussion

We presented the first MCD-based approach that generalizes to virtually any data type (i.e., categorical, ordinal, continuous). Our approach relies on Mahalanobis distances derived from CA via the SVD and 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 like disjunctive tables (see Tables 2 and 4). 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 makes use of additional information to identify inliers vs. outliers, as opposed to 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 1,969 times at the time we wrote this paper), and available in languages routinely used for multivariate analyses (e.g., https://cran.r-project.org/web/packages/robustbase/index.html, https://cran.r-project.org/web/packages/rrcov/index.html, and https://wis.kuleuven.be/stat/robust/LIBRA). We have also made the GMCD available in an R package generally focused on outliers and robust structures (see https://github.com/derekbeaton/ours).

*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 in distinct data tables (e.g., SNPs, the CDR, and brain imaging). A family of techniques have been designed specifically to handle such data sets: Multiple Factor Analysis (MFA; Abdi, Williams, & Valentin, 2013; Escofier & Pagès, 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) and Escofier and Pagès (1994) 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. 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 outliers in the observed values.

Finally, in MCA it is typical to discard many of the low-variance components (Abdi & Valentin, 2007), though we do not do so for our GMCD. The removal of low-variance components stems from the fact that the dimensionality and variance in MCA comes from the complete disjunctive data that spans $J$ columns,

as opposed to the $C$ variables (where $J > C$). The "Benzécri correction" is the most common approach to identify which low-variance components to discard (Benzécri, 1979): components with eigenvalues smaller than $1/C$.

*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 set of high-variance components. However, this does not work well in practice: in most cases outliers typically express high values on low-variance components. If we were to discard low-variance components we might not identify a proper robust structure nor identify appropriate outliers. Furthermore, the dimensionality correction (Benzécri, 1979) of the space is only established for strictly categorical—and thus completely disjunctive—data. How to determine, and then use, the correct dimensionality of the space (especially for mixed variables) is an open question. For now we have opted for the more conservative approach: require full rank data $I > J$ 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. Rather, our robust center exists entirely within the $\chi^2$-space as defined by the factor scores because subset MCA "maintains the geometry [...] and $\chi^2$-distances of the complete MCA [...]" (Greenacre & Pardo, 2006). For example if we were to apply our technique with continuous data (recoded as in Table 4) and compare against the standard MCD we would obtain slightly different results.

*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

CA 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.

*Acknowledgements*

*Funding*

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

*ONDRI*

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

*ADNI*

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

## References

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

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

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

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

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

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

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. https://doi.org/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. https://doi.org/10.1111/j.0006-341X.2000.00394.x

Benzécri, J. P. (1973). L'analyse des données: L'analyse des correspondances.

*Dunod*.

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. https://doi.org/10.1002/cem.2692

Bürkner, P.-C., & Vuorre, M. (2017). Ordinal Regression Models in Psychology: A Tutorial. *PsyArxiv*. https://doi.org/10.31234/osf.io/x8swp

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-Cordier, B. (1965). L'Analyse des Correspondences. (Thèse, Faculté

des Sciences de Rennes). Université de Rennes.

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., & Pagès, 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. https://doi.org/10.2307/2528080

Grassi, M., & Visentin, S. (1994). Correspondence analysis applied to grouped

cohort data. *Statistics in Medicine*, *13*(23-24), 2407–2425.

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

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

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

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

Greenacre, M., & Hastie, T. (1987). The Geometric Interpretation of Correspondence Analysis. *Journal of the American Statistical Association*, *82*(398), 437–447. https://doi.org/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.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. *The Prediciton of Personal Adjustment.*

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. https:

//doi.org/10.1002/wics.61

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

Hubert, M., Rousseeuw, P., & Branden, K. V. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, *47*(1), 64–79. https://doi.org/10.1198/004017004000000563

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

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

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

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

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

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., ... & di Palma, M. A. (2018). robustbase: Basic robust statistics. R package version 0.93-2, http://CRAN.R-project.org/

package=robustbase.

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

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

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

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

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

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. https://doi.org/10.1080/00401706.1999.10485670

Saporta, G. (2006). Probabilités, analyse des données et statistique. *Paris:*

*Technip*.

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.

Todorov, V., & Filzmoser, P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. Journal of Statistical Software, 32(3), 1–47.

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. https://doi.org/10.1111/1755-0998.12579

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

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