Journal of Statistics and Computer Science
Vol. 2, Issue 1, 2023, pp. 29-45
https://doi.org/10.47509/JSCS.2023.v02i01.03

An alternative geometric interpretation of sample-based Mahalanobis distances useful for interpreting outliers

Jorge Cadima

Centro de Estatística e Aplicacões da Universidade de Lisboa (CEAUL); Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, 1349-017, Lisboa, Portugal. Email: jcadima@isa.ulisboa.pt

Abstract

Mahalanobis distances are almost nine decades old and are extensively used in many areas of multivariate statistics. But there are surprisingly recent results for classical sample-based Mahalanobis distances to the centre or between individuals, such as data-independent sharp upper bounds and the fact that they become uninformative when the sample size is less than, or equal to, the number of variables plus one. It is argued here that an alternative representation of an $n \times p$ data set in the 'space of variables' that associates an axis with each of the n individuals and each variable with a vector in $\mathbb{R}^n$, provides an alternative setting where Mahalanobis distances have a precise geometric interpretation and where these recent results become obvious. It is shown that this setting also suggests a natural scaled Mahalanobis distance, in the interval $[0, 1]$, which does not depend on distributional assumptions and can be used to measure the severity of an outlier. Furthermore, a direct connection is demonstrated between Mahalanobis distances and linear regressions of certain dummy vectors in $\mathbb{R}^n$ on the p variables in the dataset, implying that standard linear regression subset selection algorithms will identify variables that are most responsible for large Mahalanobis distances, thereby assisting in the interpretation of outliers. Examples are discussed. Results are extended to Mahalanobis distances of the mean of a group of individuals to the centre or between means of two groups.

Keywords: Variable selection, space of variables, orthogonal projections, scaled Mahalanobis distance
MSC: 62H99, 15A99

1 Introduction

Ever since Mahalanobis (1936) introduced his "generalized distances in Statistics", they have been used extensively in Multivariate statistics. Mardia (1977) wrote that "Mahalanobis distance has become one of the most fundamental concepts in Multivariate Analysis". The emergence of robust alternatives, driven by concerns that include the so-called 'masking effect' by severe outliers (Barnett & Lewis, 1994) does not deprive classical Mahalanobis distances (MDs) of their importance. Population or sample-based variants of classical Mahalanobis distances are used (Anderson, 1958) in the definition of the multivariate Normal density function, in multivariate inference (Hotelling's T^2 statistic), in discriminant analyses, as well as in outlier detection (Barnett & Lewis, 1994), among other important areas. Despite their long history, some fundamental properties of classical sample-based Mahalanobis distances have only recently been discovered. Following preliminary results by Olkin (1992) for univariate and bivariate data, Gath and Hayes (2006) obtained sharp bounds for the largest sample Mahalanobis distance of an individual to the centre in a multivariate data set, that depend only on sample size and not on the data as such. Branco and Pires (2011) determined a similar data-independent sharp upper bound for sample Mahalanobis distances between pairs of individuals and proved that in the increasingly common situation where the number n of observations in the data set is less than or equal to $p + 1$, where p is the number of variables, sample-based Mahalanobis distances are, in the absence of additional multicollinearities, always equal to their upper bounds, regardless of the data values – and are therefore uninformative.

It is shown in Sections 3 and 4 that these results have a natural geometric explanation when the traditional representation of the data set defined by an $n \times p$ data matrix $\mathbf{X}$ and its column-centred counterpart $\mathbf{X}_c$, as

Article History
Received : 27 March 2023; Revised : 18 May 2023; Accepted : 18 May 2023; Published : 30 June 2023

To cite this paper

Jorge Cadima (2023). An alternative geometric interpretation of sample-based Mahalanobis distances useful for interpreting outliers. *Journal of Statistics and Computer Science.* 2(1), 29-45.

n points in $\mathbb{R}^p$, is replaced by the alternative representation in $\mathbb{R}^n$, often called the space of variables. In this latter representation, each observed individual is associated with an axis of n-dimensional Euclidean space and each observed variable defines a vector in that space, as discussed in Section 2. Mahalanobis distances of an individual to the centre depend only on the sample size and on the inclination of the subspace spanned by the columns of $\mathbf{X}_c$, $\mathcal{C}(\mathbf{X}_c)$, in relation to the axes of $\mathbb{R}^n$. This fact, which is extensively explored here, lies at the geometric heart of the results by Gath and Hayes (2006) and Branco and Pires (2011). It also provides a direct link between MDs to the centre and the coefficient of determination of the multiple linear regression of the canonical vector for each axis of $\mathbb{R}^n$ (i.e., for each individual) and the p variables in the dataset (columns of $\mathbf{X}$). This coefficient of determination is the MD scaled to the interval $[0, 1]$, thus providing a natural indicator for the severity of outliers which does not depend on any probabilistic assumptions about the population from which the sample was selected. Similar results hold for MDs between two individuals and the values of the coefficient of determination R^2 in the linear regression of the difference between their canonical axis vectors on the p variables in the dataset.

Standard variable selection methods may be applied to identify the best subsets of predictors in the above linear regressions. In this way, it is possible to identify subsets of variables that are most responsible for the value of any given MD, and thus to interpret outliers in terms of smaller subsets of the original variables. These procedures are discussed in Section 5 and three reproducible examples are provided.

Section 6 extends these ideas to Mahalanobis distances involving the vectors of means for groups of individuals. Finally, Section 7 discusses some further implications of these results.

2 Notation and preliminaries

Let $\mathbf{X}$ represent an $n \times p$ data matrix, with variables defining columns and observations (individuals) defining rows. Denote the i-th row of matrix $\mathbf{X}$ by $\vec{\mathbf{x}}_{[i]}^t$. Let $\vec{\mathbf{m}}$ be the p-dimensional vector of variable sample means and $\mathbf{X}_c$ the $n \times p$ column-centred data matrix, whose i-th row is the vector $(\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{m}})^t$. The sample covariance matrix is then $\mathbf{S} = \frac{1}{n-1}\mathbf{X}_c^t\mathbf{X}_c$. Two classical sample-based Mahalanobis distances (MDs) are the *Mahalanobis distance of individual i to the centre*:

$$d_i^2 \;=\; (\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{m}})^t \mathbf{S}^{-1} (\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{m}}) \tag{1}$$

and the *Mahalanobis distance between individuals i and j*:

$$d_{ij}^2 \;=\; (\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{x}}_{[j]})^t \mathbf{S}^{-1} (\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{x}}_{[j]}) \tag{2}$$

In keeping with standard usage we call them distances although they are, in fact, squared distances.

The above definitions can be re-written using the matrix of orthogonal projections on the subspace $\mathcal{C}(\mathbf{X}_c) \subseteq \mathbb{R}^n$ spanned by the columns of $\mathbf{X}_c$, *i.e.*, matrix $\mathbf{P_c} = \mathbf{X}_c(\mathbf{X}_c^t\mathbf{X}_c)^{-1}\mathbf{X}_c^t$. In fact, denote by $\vec{\mathbf{e}}_i$ the i-th canonical vector of $\mathbb{R}^n$, that is the vector with a single non-zero element, a 1 in position i. Pre-multiplying $\mathbf{X}_c$ by $\vec{\mathbf{e}}_i^t$ extracts the i-th row of matrix $\mathbf{X}_c$: $\vec{\mathbf{e}}_i^t\mathbf{X}_c = (\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{m}})^t$. Hence, $d_i^2 = \vec{\mathbf{e}}_i^t\mathbf{X}_c \left(\frac{1}{n-1}\mathbf{X}_c^t\mathbf{X}_c\right)^{-1} \mathbf{X}_c^t\vec{\mathbf{e}}_i = {}_{(n-1)}\,\vec{\mathbf{e}}_i^t\mathbf{P_c}\vec{\mathbf{e}}_i$. Likewise, $(\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{x}}_{[j]})^t = (\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j)^t\mathbf{X}_c$, where $\vec{\mathbf{e}}_j$ is the j-th canonical vector of $\mathbb{R}^n$. Hence, $d_{ij}^2 = {}_{(n-1)}\,(\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j)^t\mathbf{P_c}(\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j)$. Denoting the (i, j)-th element of the matrix of orthogonal projections $\mathbf{P_c}$ by p_{ij} we can further write (Puntanen, Styan, & Isotalo, 2011, eq.8.16) the MD to the centre as:

$$d_i^2 \;=\; {}_{(n-1)}\,\vec{\mathbf{e}}_i^t\mathbf{P_c}\vec{\mathbf{e}}_i \;=\; {}_{(n-1)}\,p_{ii} \tag{3}$$

and the Mahalanobis distance between individuals i and j as:

$$d_{ij}^2 \;=\; {}_{(n-1)}\,(\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j)^t\mathbf{P_c}(\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j) \;=\; {}_{(n-1)}\,(p_{ii} + p_{jj} - 2p_{ij}) \;=\; d_i^2 + d_j^2 - 2(n-1)\,p_{ij} \tag{4}$$

If the sample covariance matrix $\mathbf{S}$ is not invertible (which always happens when $n < p$), the matrix inverse $\mathbf{S}^{-1}$ in equations (1) and (2) can be replaced by the Moore-Penrose generalized inverse $\mathbf{S}^-$ and the orthogonal projection matrix by $\mathbf{P_c} = \mathbf{X}_c(\mathbf{X}_c^t\mathbf{X}_c)^-\mathbf{X}_c^t$. Expressions (3) and (4) remain valid.

Also considered is the $n \times (p+1)$ matrix $\mathbf{X}_m$ that results from binding to matrix $\mathbf{X}$ a first column of n ones, the vector $\vec{\mathbf{1}}_n$, so that $\mathbf{X}_m = \left[\vec{\mathbf{1}}_n \,:\, \mathbf{X}\right]$. This augmented matrix is the model matrix for a multiple linear regression of any response variable on the p predictors given by the columns of $\mathbf{X}$. Its associated matrix of orthogonal projections is the linear regression "hat matrix", $\mathbf{H} = \mathbf{X}_m(\mathbf{X}_m^t\mathbf{X}_m)^{-1}\mathbf{X}_m^t$.

As is well-known, the matrices $\mathbf{P}$ of orthogonal projections onto subspaces of $\mathbb{R}^n$ are the $n \times n$ symmetric ($\mathbf{P}^t = \mathbf{P}$) and idempotent ($\mathbf{PP} = \mathbf{P}$) matrices. An orthogonal projection matrix $\mathbf{P}$ projects onto its column space, $\mathcal{C}(\mathbf{P})$, which is a $trace(\mathbf{P})$-dimensional subspace of $\mathbb{R}^n$, whereas matrix $\mathbf{I}_n - \mathbf{P}$, with $\mathbf{I}_n$ the $n \times n$ identity matrix, orthogonally projects onto the orthogonal complement $\mathcal{C}(\mathbf{P})^\perp$. The subspace $\mathcal{C}(\mathbf{P})$ is the subspace of vectors that remain invariant under the application $\mathbf{P}$, i.e., such that $\mathbf{P}\vec{\mathbf{y}} = \vec{\mathbf{y}}$.

The orthogonal projection of any vector $\vec{\mathbf{y}} \notin \mathcal{C}(\mathbf{P})^\perp$ onto $\mathcal{C}(\mathbf{P})$ defines a right triangle and the cosine of the angle between $\vec{\mathbf{y}}$ and its orthogonal projection $\mathbf{P}\vec{\mathbf{y}}$ is the ratio of the norms of $\mathbf{P}\vec{\mathbf{y}}$ and $\vec{\mathbf{y}}$: $\cos(\vec{\mathbf{y}}, \mathbf{P}\vec{\mathbf{y}}) = \frac{\|\mathbf{P}\vec{\mathbf{y}}\|}{\|\vec{\mathbf{y}}\|} = \sqrt{\frac{\vec{\mathbf{y}}^t\mathbf{P}\vec{\mathbf{y}}}{\vec{\mathbf{y}}^t\vec{\mathbf{y}}}}$. Extending this relation to any non-zero vector $\vec{\mathbf{y}} \in \mathbb{R}^n$, the squared cosine of the angle between $\vec{\mathbf{y}}$ and its orthogonal projection $\mathbf{P}\vec{\mathbf{y}}$ is the Rayleigh-Ritz ratio (Horn & Johnson, 1985) of matrix $\mathbf{P}$, with vector $\vec{\mathbf{y}}$:

$$\cos^2(\vec{\mathbf{y}}, \mathbf{P}\vec{\mathbf{y}}) = \frac{\|\mathbf{P}\vec{\mathbf{y}}\|^2}{\|\vec{\mathbf{y}}\|^2} = \frac{\vec{\mathbf{y}}^t\mathbf{P}\vec{\mathbf{y}}}{\vec{\mathbf{y}}^t\vec{\mathbf{y}}} \tag{5}$$

Two particularly important subspaces of $\mathbb{R}^n$ are $\mathcal{C}(\vec{\mathbf{1}}_n)$, the one-dimensional subspace spanned by the vector of n ones, $\vec{\mathbf{1}}_n$, and its orthogonal complement, $\mathcal{C}(\vec{\mathbf{1}}_n)^\perp$, that is, the subspace of all vectors of $\mathbb{R}^n$ that are orthogonal to $\vec{\mathbf{1}}_n$, *i.e.*, whose elements add to zero. The matrix of orthogonal projections on $\mathcal{C}(\vec{\mathbf{1}}_n)$ is matrix $\mathbf{P}_{\vec{\mathbf{1}}_n} = \vec{\mathbf{1}}_n(\vec{\mathbf{1}}_n^t\vec{\mathbf{1}}_n)^{-1}\vec{\mathbf{1}}_n^t = \frac{1}{n}\vec{\mathbf{1}}_n\vec{\mathbf{1}}_n^t$, all of whose elements are $\frac{1}{n}$. The matrix of orthogonal projections on $\mathcal{C}(\vec{\mathbf{1}}_n)^\perp$, $\mathbf{I}_n - \mathbf{P}_{\vec{\mathbf{1}}_n}$, centres any vector $\vec{\mathbf{y}} \in \mathbb{R}^n$, in the sense that $\vec{\mathbf{y}}^* = (\mathbf{I}_n - \mathbf{P}_{\vec{\mathbf{1}}_n})\vec{\mathbf{y}}$ has as its i-th element $y_i - \overline{y}$, where y_i is the generic element of vector $\vec{\mathbf{y}}$ and $\overline{y}$ is the arithmetic mean of those elements.

Orthogonal projections onto nested subspaces of $\mathbb{R}^n$ have interesting properties. If $\mathcal{M} \subseteq \mathcal{N} \subseteq \mathbb{R}^n$ are two nested subspaces and $\mathbf{P}_M$ and $\mathbf{P}_N$ their respective matrices of orthogonal projections, then $\mathbf{P}_M\mathbf{P}_N = \mathbf{P}_N\mathbf{P}_M = \mathbf{P}_M$ and $\mathbf{P}_N - \mathbf{P}_M$ is the matrix of orthogonal projections onto the subspace $\mathcal{N} \cap \mathcal{M}^\perp \subset \mathbb{R}^n$ (Puntanen et al., 2011, Prop.7.1). This last result, together with equation (5), directly imply that

$$\cos^2(\vec{\mathbf{y}}, \mathbf{P}_M\vec{\mathbf{y}}) \;=\; \cos^2(\vec{\mathbf{y}}, \mathbf{P}_N\vec{\mathbf{y}}) - \cos^2(\vec{\mathbf{y}}, (\mathbf{P}_N - \mathbf{P}_M)\vec{\mathbf{y}}) \;. \tag{6}$$

Furthermore, and since $\cos^2(\vec{\mathbf{y}}, \mathbf{P}_M\vec{\mathbf{y}}) = \frac{\|\mathbf{P}_M\vec{\mathbf{y}}\|^2}{\|\vec{\mathbf{y}}\|^2} = \frac{\|(\mathbf{P}_M\mathbf{P}_N)\vec{\mathbf{y}}\|^2}{\|\mathbf{P}_N\vec{\mathbf{y}}\|^2} \cdot \frac{\|\mathbf{P}_N\vec{\mathbf{y}}\|^2}{\|\vec{\mathbf{y}}\|^2} = \cos^2\left(\mathbf{P}_N\vec{\mathbf{y}}, \mathbf{P}_M(\mathbf{P}_N\vec{\mathbf{y}})\right) \cdot \cos^2(\vec{\mathbf{y}}, \mathbf{P}_N\vec{\mathbf{y}}) = \cos^2(\mathbf{P}_N\vec{\mathbf{y}}, \mathbf{P}_M\vec{\mathbf{y}}) \cdot \cos^2(\vec{\mathbf{y}}, \mathbf{P}_N\vec{\mathbf{y}})$, we have:

$$\cos^2(\vec{\mathbf{y}}, \mathbf{P}_M\vec{\mathbf{y}}) \;=\; \cos^2(\vec{\mathbf{y}}, \mathbf{P}_N\vec{\mathbf{y}}) \cdot \cos^2(\mathbf{P}_N\vec{\mathbf{y}}, \mathbf{P}_M\vec{\mathbf{y}}) \quad \leq \quad \cos^2(\vec{\mathbf{y}}, \mathbf{P}_N\vec{\mathbf{y}}) \;. \tag{7}$$

Linear regressions are, in linear algebraic terms, orthogonal projections of a response variable $\vec{\mathbf{y}} \in \mathbb{R}^n$ onto a subspace spanned by the predictor variables and the vector $\vec{\mathbf{1}}_n$, i.e., onto $\mathcal{C}(\mathbf{X}_m)$. The orthogonal projection matrix onto $\mathcal{C}(\mathbf{X}_m)$ is the "hat matrix" $\mathbf{H} = \mathbf{P}_\mathbf{c} + \mathbf{P}_{\vec{\mathbf{1}}_n}$, where $\mathbf{P}_\mathbf{c}$ and $\mathbf{P}_{\vec{\mathbf{1}}_n}$ are the matrices of orthogonal projections onto, respectively, $\mathcal{C}(\mathbf{X}_c) = \mathcal{C}(\mathbf{X}_m) \cap \mathcal{C}(\vec{\mathbf{1}}_n)^\perp$ and $\mathcal{C}(\vec{\mathbf{1}}_n)$ (Puntanen et al., 2011, p.157). It is therefore straightforward to show that the coefficient of determination of the multiple linear regression of a response variable $\vec{\mathbf{y}} \in \mathbb{R}^n$ on the p variables is the squared cosine of the angle between the centred vector $\vec{\mathbf{y}}^* = (\mathbf{I}_n - \mathbf{P}_{\vec{\mathbf{1}}_n})\vec{\mathbf{y}}$ and either subspace $\mathcal{C}(\mathbf{X}_m)$ or $\mathcal{C}(\mathbf{X}_c)$:

$$R^2 = \cos^2(\vec{\mathbf{y}}^*, \mathbf{H}\vec{\mathbf{y}}^*) = \cos^2(\vec{\mathbf{y}}^*, \mathbf{P}_\mathbf{c}\vec{\mathbf{y}}^*) \;. \tag{8}$$

In fact, the coefficient of determination in a linear regression is defined as the ratio of regression and total sum of squares, where $SSR = \|\mathbf{H}\vec{\mathbf{y}}^*\|^2$ and $SST = \|\vec{\mathbf{y}}^*\|^2$. Hence, $R^2 = \frac{\vec{\mathbf{y}}^{*t}\mathbf{H}\vec{\mathbf{y}}^*}{\vec{\mathbf{y}}^{*t}\vec{\mathbf{y}}^*} = \cos^2(\vec{\mathbf{y}}^*, \mathbf{H}\vec{\mathbf{y}}^*)$. However, $\vec{\mathbf{y}}^{*t}\mathbf{H}\vec{\mathbf{y}}^* = \vec{\mathbf{y}}^{*t}(\mathbf{P}_\mathbf{c} + \mathbf{P}_{\vec{\mathbf{1}}_n})\vec{\mathbf{y}}^* = \vec{\mathbf{y}}^{*t}\mathbf{P}_\mathbf{c}\vec{\mathbf{y}}^* + \vec{\mathbf{y}}^{*t}\mathbf{P}_{\vec{\mathbf{1}}_n}\vec{\mathbf{y}}^*$. Since the latter term is zero, $R^2 = \frac{\vec{\mathbf{y}}^{*t}\mathbf{P}_\mathbf{c}\vec{\mathbf{y}}^*}{\vec{\mathbf{y}}^{*t}\vec{\mathbf{y}}^*} = \cos^2(\vec{\mathbf{y}}^*, \mathbf{P}_\mathbf{c}\vec{\mathbf{y}}^*)$.

3 Mahalanobis distances to the centre

Proposition 3.1 *Let* $\mathbf{X}$ *be an* $n \times p$ *data matrix,* $\mathbf{X}_c = (\mathbf{I}_n - \mathbf{P}_{\vec{\mathbf{1}}_n})\mathbf{X}$ *its column-centred counterpart and* $\mathcal{C}(\mathbf{X}_c)$ *the column-space of* $\mathbf{X}_c$*. Let* $\vec{\mathbf{e}}_i$ *be the* i*-th canonical basis vector of* $\mathbb{R}^n$ *and* $\vec{\mathbf{e}}_i^* = (\mathbf{I}_n - \mathbf{P}_{\vec{\mathbf{1}}_n})\vec{\mathbf{e}}_i$ *its column-centred counterpart, whose* i*-th element is* $1 - \frac{1}{n}$ *and all other elements are* $-\frac{1}{n}$*. The classical Mahalanobis distance of individual* i *to the centre,* d_i^2*, is given by:*

1. $d_i^2 = {}_{(n-1)} \cos^2 \theta_i$*, where* θ_i *is the angle between* $\vec{\mathbf{e}}_i$ *and the subspace* $\mathcal{C}(\mathbf{X}_c)$*.*
2. $d_i^2 = \frac{(n-1)^2}{n} \cos^2 \theta_i^*$*, where* θ_i^* *is the angle between* $\vec{\mathbf{e}}_i^*$ *and the subspace* $\mathcal{C}(\mathbf{X}_c)$*.*
3. $d_i^2 = \frac{(n-1)^2}{n} R_i^2$*, where* R_i^2 *is the coefficient of determination of the multiple linear regression of canonical vector* $\vec{\mathbf{e}}_i$ *on the* p *columns of* $\mathbf{X}$*.*

Proof

1. The result follows directly from equations (3) and (5), since $\vec{\mathbf{e}}_i^t \vec{\mathbf{e}}_i = 1$.
2. Any linear combination of the columns of matrix $\mathbf{X}_c$ must be centred (that is, its elements must add to zero), hence $\mathcal{C}(\mathbf{X}_c) \subseteq \mathcal{C}(\vec{\mathbf{1}}_n)^\perp$. Vector $\vec{\mathbf{e}}_i^*$ is the orthogonal projection of $\vec{\mathbf{e}}_i$ onto $\mathcal{C}(\vec{\mathbf{1}}_n)^\perp$. By equation (7), with $\mathcal{M} = \mathcal{C}(\mathbf{X}_c)$ and $\mathcal{N} = \mathcal{C}(\vec{\mathbf{1}}_n)^\perp$, we have $\cos^2 \theta_i = \cos^2(\vec{\mathbf{e}}_i, \mathbf{P_c}\vec{\mathbf{e}}_i) = \cos^2(\vec{\mathbf{e}}_i, \vec{\mathbf{e}}_i^*) \cdot \cos^2(\vec{\mathbf{e}}_i^*, \mathbf{P_c}\vec{\mathbf{e}}_i)$. The first factor is $\frac{\|\vec{\mathbf{e}}_i^*\|^2}{\|\vec{\mathbf{e}}_i\|^2} = \|\vec{\mathbf{e}}_i^*\|^2 = \frac{n-1}{n}$. Thus, $d_i^2 = (n-1)\cos^2 \theta_i = \frac{(n-1)^2}{n} \cos^2(\vec{\mathbf{e}}_i^*, \mathbf{P_c}\vec{\mathbf{e}}_i)$. Since $\mathcal{C}(\mathbf{X}_c) \subseteq \mathcal{C}(\vec{\mathbf{1}}_n)^\perp$, we have $\mathbf{P_c} = \mathbf{P_c}(\mathbf{I}_n - \mathbf{P}_{\vec{\mathbf{1}}_n})$. Hence, $\mathbf{P_c}\vec{\mathbf{e}}_i = \mathbf{P_c}(\mathbf{I}_n - \mathbf{P}_{\vec{\mathbf{1}}_n})\vec{\mathbf{e}}_i = \mathbf{P_c}\vec{\mathbf{e}}_i^*$ and therefore $\cos^2(\vec{\mathbf{e}}_i^*, \mathbf{P_c}\vec{\mathbf{e}}_i) = \cos^2(\vec{\mathbf{e}}_i^*, \mathbf{P_c}\vec{\mathbf{e}}_i^*) = \cos^2 \theta_i^*$.
3. From equation (8) the coefficient of determination in the multiple linear regression of $\vec{\mathbf{e}}_i$ on the columns of $\mathbf{X}$ is $R_i^2 = \cos^2(\vec{\mathbf{e}}_i^*, \mathbf{P_c}\vec{\mathbf{e}}_i^*)$. Hence, $d_i^2 = \frac{(n-1)^2}{n} R_i^2$, as was to be shown.

These characterizations show that Mahalanobis distances to the centre are angular measurements in $\mathbb{R}^n$, reflecting the inclination of the subspace $\mathcal{C}(\mathbf{X}_c)$ in relation to the coordinate axes of $\mathbb{R}^n$ or to the orthogonal projection of those axes onto $\mathcal{C}(\vec{\mathbf{1}}_n)^\perp$. The smaller the angle that any given axis forms with the subspace $\mathcal{C}(\mathbf{X}_c)$, the larger the MD to the centre of the individual associated with that axis.

Gath & Hayes' sharp upper bound for any Mahalanobis distance to the centre, $d_i^2 \leq \frac{(n-1)^2}{n}$ (Gath & Hayes, 2006, Theorem 2.1), follows directly from point 2. This bound does not depend on the data, but only on the sample size n. The upper bound can be re-written as $\frac{(n-1)^2}{n} = n - 2 + \frac{1}{n}$ which, for large n, is approximately $n - 2$. The characterization in point 1, $d_i^2 = (n-1)\cos^2 \theta_i$, does not provide a sharp bound, because angle θ_i is the angle between $\vec{\mathbf{e}}_i \notin \mathcal{C}(\vec{\mathbf{1}}_n)^\perp$ and its orthogonal projection onto a subspace to which it does not belong, the subspace $\mathcal{C}(\mathbf{X}_c) \subseteq \mathcal{C}(\vec{\mathbf{1}}_n)^\perp$. Thus, $\cos^2 \theta_i$ must always be strictly less than 1.

It is useful to define a scaled version of the Mahalanobis distance to the centre, necessarily in the interval $[0, 1]$, which can be used as an indicator of the severity of any observation as an outlier, regardless of probability distributions.

Definition 3.1 *Let* $\mathbf{X}$ *be an* $n \times p$ *data matrix and* $\mathbf{X}_c$ *its column-centred counterpart. Let* d_i^2 *be the Mahalanobis distance to the centre of observation* i*. Define the scaled Mahalanobis distance to the centre of observation* i *as:*

$$s_i^2 = \frac{d_i^2}{M} = \cos^2 \theta_i^* = R_i^2 \,, \tag{9}$$

where $M = \frac{(n-1)^2}{n}$ *is the largest possible MD to the centre for this dataset;* θ_i^* *is the angle between the* i*-th centred canonical vector* $\vec{\mathbf{e}}_i^*$ *and* $\mathcal{C}(\mathbf{X}_c)$*; and* R_i^2 *is the coefficient of determination of the multiple linear regression of the* i*-th canonical vector* $\vec{\mathbf{e}}_i$ *on the* p *variables in the dataset.*

For a dataset from a Multivariate Normal distribution, Johnson and Wichern (2007, p.184) state that sample-based Mahalanobis distances d^2 to the centre approximately follow a χ^2_p distribution, for large n and $n-p$. Furthermore, Morrison (1990, p.177, eq.3) indicates that the following transformation follows an $F_{p,n-p-1}$ distribution:

$$F = \frac{(n-p-1)n\,d^2}{p(n-1)^2 - np\,d^2} \ . \tag{10}$$

Simple algebra shows that F in equation (10) is the standard linear regression goodness-of-fit F statistic $\frac{n-p-1}{p}\,\frac{R^2}{1-R^2}$, for the regression of the canonical vectors in $\mathbb{R}^n$ on the p variables in the dataset, since by point 3 of Proposition 3.1, we have $R^2 = \frac{d^2}{\frac{(n-1)^2}{n}}$, the scaled MD to the centre of observation i. Morrison (1990, p.178) provides an asymptotic distribution for scaled MDs to the centre, stating that if the value of the F statistic in equation (10) is beyond the range of the available F tables, the statistic $u = \frac{n}{(n-1)^2}\,d^2$ can be referred to tables of the incomplete beta function, with $a = \frac{p}{2}$ and $b = \frac{n-p-1}{2}$.

Proposition 3.1 shows that an MD to the center attains its maximum value $M = \frac{(n-1)^2}{n}$ if and only if $\vec{\mathbf{e}}_i^*$ belongs to $\mathcal{C}(\mathbf{X}_c)$ or, equivalently, iff $R_i^2 = 1$, in which case $\vec{\mathbf{e}}_i$ is a linear combination of the columns of matrix $\mathbf{X}_m$. Thus,

$$d_i^2 = M \quad \Leftrightarrow \quad \vec{\mathbf{e}}_i \in \mathcal{C}(\mathbf{X}_m) \quad \Leftrightarrow \quad \vec{\mathbf{e}}_i^* \in \mathcal{C}(\mathbf{X}_c). \tag{11}$$

The first equivalence can be re-written as:

$$d_i^2 \ = \ M \quad \Leftrightarrow \quad \vec{\mathbf{e}}_i \ = \ b_0\,\vec{\mathbf{1}}_n + b_1\,\vec{\mathbf{x}}_1 + b_2\,\vec{\mathbf{x}}_2 + ... + b_p\,\vec{\mathbf{x}}_p \ , \tag{12}$$

for some set of real coefficients $b_0, b_1, ..., b_p$. This means that there is a linear combination $b_1\,\vec{\mathbf{x}}_1 + b_2\,\vec{\mathbf{x}}_2 + ... + b_p\,\vec{\mathbf{x}}_p$ of the p columns of matrix $\mathbf{X}$ that perfectly discriminates individual i (with coordinate $1-b_0$) from the remaining $n-1$ individuals (with common coordinate $-b_0$) – see also Pires and Branco (2018). This result can be slightly relaxed, taking into consideration point 3 of Proposition 3.1. An individual with $d_i^2 < M$ (*i.e.*, $R_i^2 < 1$) will, on the discriminant axis $b_1\,\vec{\mathbf{x}}_1 + b_2\,\vec{\mathbf{x}}_2 + ... + b_p\,\vec{\mathbf{x}}_p$ defined by the linear regression of $\vec{\mathbf{e}}_i$ on the p predictors, have coordinate $(1-b_0)-r_i$ where r_i is the (usual) regression residual of individual i. Any other individual $j \neq i$ will, on the same axis, have coordinate $-b_0-r_j$ where r_j is the corresponding regression residual. We have $R_i^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{j=1}^n r_j^2}{\frac{n-1}{n}}$, since the total sum of squares in this regression is $SST = \|\vec{\mathbf{e}}_i^*\|^2 = \frac{n-1}{n}$. Thus, the MD to the centre of individual i can also be written as $d_i^2 = \frac{(n-1)^2}{n} - (n-1)\sum_{j=1}^{n} r_j^2$ and the sum of squared residuals is $\sum_{j=1}^{n} r_j^2 = \frac{n-1}{n}\left[1 - \frac{d_i^2}{M}\right]$. An individual with d_i^2 close to the upper bound M must therefore have all regression residuals r_j close to zero, and will therefore be well separated from the remaining individuals on the discriminant axis.

Alternatively, equation (12) can be interpreted in $\mathbb{R}^p$, since it implies that individual i is at the maximum Mahalanobis distance from the center if and only if, for all $j \neq i$, we have $b_0 + b_1\,x_{1_{(j)}} + b_2\,x_{2_{(j)}} + ... + b_p\,x_{p_{(j)}} = 0$, i.e., the $n-1$ points $\vec{\mathbf{x}}_{[j]} \in \mathbb{R}^p$ $(j \neq i)$ belong to the hyperplane in $\mathbb{R}^p$ of equation $b_0 + b_1\,x_1 + b_2\,x_2 + ... + b_p\,x_p = 0$. This is the result in Corollaries 2.2 and 2.3 of Gath & Hayes (Gath & Hayes, 2006).

The mean of all MDs to the centre is, from equation (3),

$$\overline{d^2} \ = \ \frac{1}{n}\sum_{i=1}^{n} d_i^2 \ = \ \frac{1}{n}\,(n-1)\,trace(\mathbf{P_c}) \ = \ \frac{n-1}{n}\,r \ , \tag{13}$$

where r is the dimension of the subspace $\mathcal{C}(\mathbf{X}_c)$. Since the maximum of a set of n numbers is no smaller than their mean, and equality exists iff all n numbers are equal, the largest MD to the centre verifies the inequality $\max_i d_i^2 \geq \frac{n-1}{n}\,r$ (for large n, approximately r), with equality iff all n individuals share this smallest possible maximum MD. This is the result in Gath & Hayes's Theorem 3.1 and Corollary 3.2 (which assumed $r = p$). Geometrically, this is equivalent to stating that subspace $\mathcal{C}(\mathbf{X}_c)$ forms equal angles with all n axes in $\mathbb{R}^n$. Gath & Hayes point out that this is only possible for certain combinations of n and r.

The results by Pires and Branco (2018) for $n \leq p+1$ follow directly from point 3 in Proposition 3.1. In fact, $rank(\mathbf{X}_m) \leq min\{n, p+1\}$. If $n \leq p+1$, and assuming that there are no further linear dependencies, then $dim(\mathcal{C}(\mathbf{X}_m)) = n$. But since $\mathcal{C}(\mathbf{X}_m) \subseteq \mathbb{R}^n$, this implies that $\mathcal{C}(\mathbf{X}_m) = \mathbb{R}^n$. Hence, all canonical vectors of $\mathbb{R}^n$ belong to $\mathcal{C}(\mathbf{X}_m)$, implying that all individuals are at the maximum Mahalanobis distance to the centre, regardless of the data, thus making sample MDs to the centre uninformative in this context.

The following Proposition implies that the Mahalanobis distance to the centre of a given individual cannot increase if one or more of the p variables is dropped.

Proposition 3.2 *Let $\mathbf{X}$ be an $n \times p$ data matrix and $\mathbf{X}_c$ its column-centred counterpart. Let $\mathbf{X_s}$ be a data matrix on the same individuals, such that the column-space of its column-centred counterpart is contained in the column-space of $\mathbf{X}_c$, i.e., $\mathcal{C}(\mathbf{X}_{\mathbf{s}_c}) \subseteq \mathcal{C}(\mathbf{X}_c)$. Let d_i^2 and $d_{i\,(s)}^2$ be the MDs to the centre of individual i, based on matrices $\mathbf{X}$ and $\mathbf{X_s}$, respectively. Then,*

1. $d_{i\,(s)}^2 \leq d_i^2$.
2. $d_i^2 \geq \max\limits_{j=1,\ldots,p} z_{ij}^2$, *where* $z_{ij}^2 = \frac{(x_{ij}-\bar{x}_j)^2}{s_j^2}$ *is individual i's squared value on the j-th standardized variable.*
3. $d_i^2 - d_{i\,(s)}^2 = (n-1)\cos^2\phi_i = \frac{(n-1)^2}{n}\cos^2\phi_i^*$, *where ϕ_i and ϕ_i^* are the angles between, respectively, $\vec{\mathbf{e}}_i$ and $\vec{\mathbf{e}}_i^*$, and their orthogonal projections onto $\mathcal{C}(\mathbf{X}_c) \cap \mathcal{C}(\mathbf{X}_{\mathbf{s}_c})^\perp$.*

Proof

1. Since $\mathcal{C}(\mathbf{X}_{\mathbf{s}_c}) \subseteq \mathcal{C}(\mathbf{X}_c)$, the inequality in (7) implies that $\cos^2(\vec{\mathbf{e}}_i^*, \mathbf{P}_{\mathbf{s}_c}\vec{\mathbf{e}}_i^*) \leq \cos^2(\vec{\mathbf{e}}_i^*, \mathbf{P_c}\vec{\mathbf{e}}_i^*)$, where $\mathbf{P}_{\mathbf{s}_c}$ is the matrix of orthogonal projections onto the column-space of $\mathbf{X}_{\mathbf{s}_c}$. It follows directly from point 2 of Proposition 3.1 that $d_{i\,(s)}^2 \leq d_i^2$.
2. For a subset with the single variable j, the MD of individual i to the centre is, by definition, the squared value, on the scaled variable j, for individual i: $\frac{(x_{ij}-\bar{x}_j)^2}{s_j^2}$ (Gath & Hayes, 2006). By the previous point, d_i^2 is no smaller than all such univariate MDs.
3. From equations (3) and (5), d_i^2–$d_{i\,(s)}^2 = (n$–$1)[\cos^2(\vec{\mathbf{e}}_i, \mathbf{P_c}\vec{\mathbf{e}}_i)$–$\cos^2(\vec{\mathbf{e}}_i, \mathbf{P}_{\mathbf{s}_c}\vec{\mathbf{e}}_i)]$. But since $\mathcal{C}(\mathbf{X}_{\mathbf{s}_c}) \subseteq \mathcal{C}(\mathbf{X}_c)$, equation (6) implies that $d_i^2 - d_{i\,(s)}^2 = (n-1)\cos^2(\vec{\mathbf{e}}_i, (\mathbf{P_c} - \mathbf{P}_{\mathbf{s}_c})\vec{\mathbf{e}}_i)$, where matrix $\mathbf{P_c} - \mathbf{P}_{\mathbf{s}_c}$ is the matrix of orthogonal projections onto $\mathcal{C}(\mathbf{X}_c) \cap \mathcal{C}(\mathbf{X}_{\mathbf{s}_c})^\perp$. This proves the first equality. The second equality is proved analogously to point 2 in Proposition 3.1.

Proposition 3.2 implies that the a given individual's MD to the centre can never increase when subsets of the original variables are discarded. It also implies that $d_i^2 = d_{i\,(s)}^2$ if and only if $\mathbf{P_c}\vec{\mathbf{e}}_i = \mathbf{P}_{\mathbf{s}_c}\vec{\mathbf{e}}_i$, i.e., iff $\vec{\mathbf{e}}_i$ belongs to the orthogonal complement of $\mathcal{C}(\mathbf{X}_c) \cap \mathcal{C}(\mathbf{X}_{\mathbf{s}_c})^\perp$.

4 Mahalanobis distances between individuals

Similar results can be obtained for sample-based MDs between individuals.

Proposition 4.1 *Let $\mathbf{X}$, $\mathbf{X}_c$, $\mathbf{P_c}$ and $\mathbf{X}_m$ be defined as in Section 3, and $\vec{\mathbf{e}}_i$ ($\vec{\mathbf{e}}_j$) represent the i-th (j-th) canonical basis vector of $\mathbb{R}^n$. The classical Mahalanobis distance between individuals i and j, defined in equations (2) and (4) is given by:*

1. $d_{ij}^2 = 2(n-1)\cos^2\theta_{ij}$, *where θ_{ij} is the angle between $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j$ and the subspace $\mathcal{C}(\mathbf{X}_c)$;*
2. $d_{ij}^2 = 2(n-1)R_{ij}^2$, *where R_{ij}^2 is the coefficient of determination of the multiple linear regression of vector $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j$ on the p predictors defined by matrix $\mathbf{X}$.*
3. $d_{ij}^2 = 2(n-1) \quad \Leftrightarrow \quad \vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j \in \mathcal{C}(\mathbf{X}_c) \quad \Leftrightarrow \quad \vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j \in \mathcal{C}(\mathbf{X}_m)$.

Proof

1. The result follows directly from equations (4) and (5), since the norm of vector $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j$ is $\sqrt{2}$.

2. Since vector $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j \in \mathcal{C}(\vec{\mathbf{1}}_n)^\perp$ (its elements add to zero), equation (8) implies that the coefficient of determination in the multiple linear regression of $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j$ on the variables defining matrix $\mathbf{X}$ is given by $R^2_{ij} = \cos^2(\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j, \mathbf{P_c}(\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j))$.

3. From point 1 above, the maximum value $d^2_{ij} = 2(n-1)$ is achieved iff $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j \in \mathcal{C}(\mathbf{X}_c)$. From point 2, we also have the maximum d^2_{ij} iff $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j$ is an exact linear combination of the columns of matrix $\mathbf{X}_m$.

In these characterizations, vector $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j$ plays the role of both vector $\vec{\mathbf{e}}_i$ and vector $\vec{\mathbf{e}}^*_i$ in Section 3. Since vector $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j$ belongs to $\mathcal{C}(\vec{\mathbf{1}}_n)^\perp$, it may also belong to the target subspace $\mathcal{C}(\mathbf{X}_c)$. The bound $d^2_{ij} \leq 2(n-1)$, given in Branco and Pires (2011) results directly from point 1 above. It is a sharp bound, since point 2 implies that the maximum value for d^2_{ij} is achieved when $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j \in \mathcal{C}(\mathbf{X}_m)$. A consequence of the equivalences in point 3 is that if $d^2_{ij} = d^2_{jk} = 2(n-1)$ (for $k \neq i$) then it must also be the case that $d^2_{ik} = 2(n-1)$, since $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j \in \mathcal{C}(\mathbf{X}_m)$ and $\vec{\mathbf{e}}_j - \vec{\mathbf{e}}_k \in \mathcal{C}(\mathbf{X}_m)$ implies that $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_k \in \mathcal{C}(\mathbf{X}_m)$. The results by Branco and Pires (2011) regarding the situation $rank(\mathbf{X}_m) = n \leq p+1$, are again a direct result of the fact that, in this case, $\mathcal{C}(\mathbf{X}_m) = \mathbb{R}^n$, so that all differences $\vec{\mathbf{e}}_i - \vec{\mathbf{e}}_j$ $(i \neq j)$ are vectors in $\mathcal{C}(\mathbf{X}_m)$. Hence the maximum MD $d^2_{ij} = 2(n-1)$ is attained, regardless of the data, for any pair (i, j) of individuals. MDs between individuals are also uninformative in this case.

The results in Proposition 3.2 can be adapted to the context of this Section. In particular, MDs between a given pair (i, j) of individuals form a non-increasing sequence along a set of ever-smaller nested subspaces.

It is also possible to define a scaled Mahalanobis distance between individuals, as in Definition 3.1: $s^2_{ij} = \cos^2\theta_{ij} = R^2_{ij}$, with the notation of Proposition 4.1. This scaled MD must belong to the interval $[0, 1]$.

5 Variable selection for Mahalanobis distances

The results in Sections 3 and 4 suggest an algorithm for identifying subsets of variables than are mostly responsible for the value of a given MD, thereby highlighting the causes for large MD values. This procedure is illustrated in this Section, for MDs to the centre.

1. select an individual i whose MD to the centre d^2_i is of interest (usually one of the largest);

2. perform the multiple linear regression of the i-th canonical basis vector $\vec{\mathbf{e}}_i \in \mathbb{R}^n$ on the p predictor variables defined by $\mathbf{X}$;

3. use any standard method of selecting a k-variable subset $\mathcal{S}$ of predictors that preserves an acceptable proportion π of the original coefficient of determination R^2.

Point 3 in Proposition 3.1 guarantees that the k variables in subset $\mathcal{S}$ define an $n \times k$ data matrix $\mathbf{X}_\mathcal{S}$ in which individual i has MD to the centre $\pi\, d^2_i$. In other words, the k variables in $\mathcal{S}$ account for $\pi \times 100\%$ of the original MD to the centre of individual i. This reduction in dimensionality makes it easier to assess, interpret and possibly visualize the causes of large values of d^2_i.

The above procedure is illustrated with three examples, using `R` (R Core Team, 2021). The three datasets are from the University of California's UCI Machine Learning Repository (Dheeru & Karra Taniskidou, 2017, http://archive.ics.uci.edu/ml). Mahalanobis distances were calculated with the `mahalanobis` function in the `stats` package, which is part of the standard distributions of `R`. Subset selection of predictors in the linear regressions was carried out using the `subselect R` package (Orestes Cerdeira, Duarte Silva, Cadima, & Minhoto, 2023), details of which are given in the package vignette. Package `subselect` provides function `eleaps` that guarantees the identification of the best subsets for each cardinality, based on Furnival and Wilson's efficient leaps-and-bounds algorithm (Furnival & Wilson, 1974; Duarte Silva, 2001, 2002). This function is computationally feasible for datasets with up to approximately $p = 30$ variables. For larger datasets, `subselect` also provides functions with heuristic algorithms (functions `anneal`, `genetic`

and `improve`), that can be used to seek the optimal subsets of predictors. See Cadima, Cerdeira, and Minhoto (2004) for further details.

The graphs shown below and the results for the optimal subset selections were obtained with a simple `R` function, `xploremaha`, which is shown in Appendix A to ensure reproducible examples. Function `xploremaha` requires as input arguments the `data` frame with the dataset and the `rank` of the Mahalanobis distance to the centre that is to be explored. If the number of variables in the dataset is less than or equal to an argument `switch` (by default 30), function `eleaps` is invoked to carry out a full search of optimal subsets, otherwise function `anneal` will provide a heuristic solution based on a simulated annealing algorithm. The output argument is a list with the best subsets (for all cardinalities) in component `bestsets`, as well as their corresponding scaled MD (`bestvalues`), which represents the proportion of the MD's upper bound $(n-1)^2/n$ that is associated with each subset. By default, `xploremaha` also produces a plot of the selected individual's MDs to the centre, for each of the identified best subsets, for all cardinalities from 1 to $p-1$. This option may be turned off by setting the input argument `plot` to the logical value `FALSE`.

5.1 Example: the `abalone` data

The `abalone` dataset was contributed to the UCI Repository by Nash, Sellers, Talbot, Cawthorn, and Ford (1994), from the Marine Research Laboratories in Taroona, Australia. The data set has a moderately large number of individuals, with observations on $n=4177$ abalone sea snails. Of the nine original variables, one (sex) was categorical and has been excluded, with the remaining $p=8$ numerical physical measurements labelled $V2$ to $V9$. The largest MD to the centre is $d^2_{2052}=2120.615$, which is approximately one half of the upper bound $\frac{(n-1)^2}{n}=4175$, and much larger than the mean MD to the centre, $\frac{n-1}{n}p=7.998$ (equation 13). The 0.9999 quantile in a χ^2_8 distribution is 31.83, providing another benchmark to state that observation 2052 is a very severe outlier. Figure 1, which was created with the command `xploremaha(data=abalone, rank=1)`, shows the Mahalanobis distances of observation $i=2052$ to the centre (vertical axis) for the best variable subsets of all cardinalities from 1 to 7 (horizontal axis). Beneath each point is the proportion of the original MD retained by these best subsets. As can be seen, three variables are sufficient to capture the main reasons for such a severe outlier, and even two variables will display the essence of the outlying nature of this observation, retaining 90% of its MD to the centre.

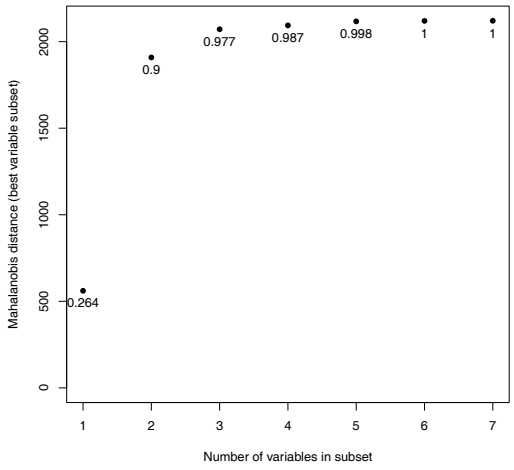

Figure 1: Mahalanobis distances of observation $i=2052$ to the centre for the best variable subsets of all cardinalities from 1 to 7.

With the help of the `eleaps` function in package `subselect` (which is invoked by function `xploremaha`),

the two optimal variables were identified as the variables $V3$ (diameter) and $V4$ (height). The scatterplot of all $n=4177$ observations on these two variables is shown in Figure 2. Observation $i=2052$, which has the largest MD to the centre using all $p=8$ variables, is the triangle at the top of the plot. As can be seen, individual 2052 is a univariate outlier on variable $V4$ alone, but Figure 1 tells us that we would lose over two-thirds of its MD to the centre by focusing only on variable $V4$ and ignoring the pattern of the scatterplot associated with both $V3$ and $V4$. The values for variable $V4$ suggest that a decimal point may have been mis-placed when recording the height of individual 2052 (which is given as 1.130, whereas 99.95% of the values are no larger than 0.250).

The other outlying point in Figure 2 is individual $i=1418$, represented by the square at the centre-right of the plot. This individual's MD to the centre with all $p=8$ variables is the third largest in the dataset, at $d^2_{1418}=222.201$ (5.3% of the upper bound). A plot similar to Figure 1 for this individual (not shown) would reveal that almost 88% of this MD is preserved by only two variables, again $V3$ and $V4$, and in this optimal bivariate data set the MD to the centre of individual 1418 is 194.760.

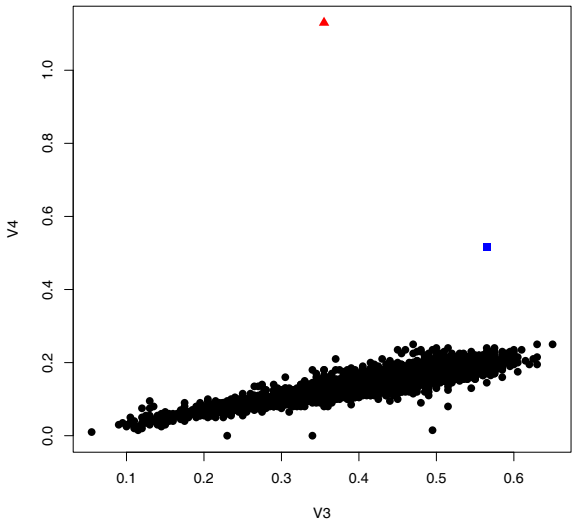

Figure 2: The scatterplot of the `abalone` observations on variables $V3$ (diameter, in mm) and $V4$ (height, in mm).

The second largest MD to the centre in the `abalone` dataset is for individual 1211, with $d^2_{1211}=250.739$, which is 6.0% of the upper bound, but still much larger than the 0.9999 quantile on a χ^2_8 distribution (31.83). Once more, two variables suffice to capture most of this Mahalanobis distance to the centre, but the optimal variables in this case are $V2$ and $V3$. The plot on the left in Figure 3, which was created with the command `xploremaha(data=abalone, rank=2)`, shows the MDs to the centre for individual $i=1211$, with the best variable subsets of all sizes from 1 to 7, as well as the proportion of the original MD that is retained. The plot on the right in Figure 3 is the best two-variable scatterplot, with variables $V2$ (length, in mm) and $V3$ (diameter, in mm), which preserves almost 95% of this individual's MD to the centre in the full dataset. In this bivariate plot, individual 1211 is the triangle at the centre-left, with MD to the centre 237.830. Individuals 2052 and 1418, with the largest and third largest MDs in the full dataset, are not identifiable in this scatterplot.

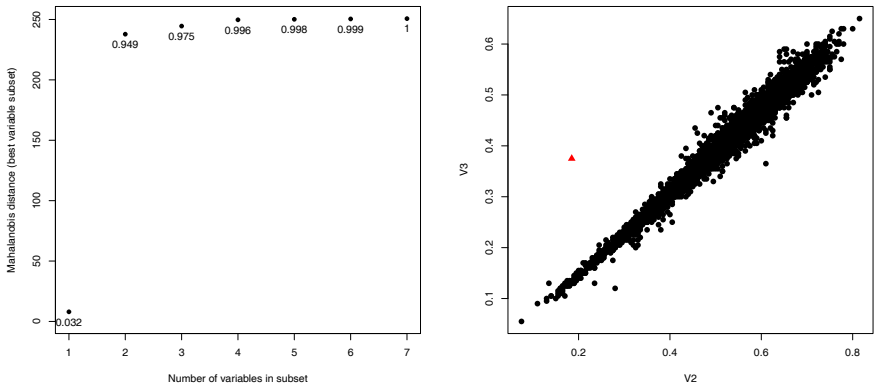

Figure 3: On the left, Mahalanobis distances to the centre of individual 1211 for the best smaller-cardinality subsets of the `abalone` dataset. On the right, the bivariate scatterplot which best highlights this individual's MD to the centre.

5.2 Example: the `SPECTF` heart data

A second dataset from the UCI Repository is the `SPECTF` dataset, with features from the Single Proton Emission Computed Tomography (SPECT) cardiac images of $n = 267$ patients (the training and the test subsets were merged for our purposes). This data set was contributed to the UCI Repository by L.A. Kurgan and K.J. Cios of the University of Colorado at Denver (Kurgan, Cios, Tadeusiewicz, Ogiela, & Goodenday, 2001). The first, binary, variable was excluded and the Mahalanobis distances were computed using the remaining $p = 44$ continuous variables. The upper bound for MDs to the centre is $\frac{(n-1)^2}{n} = 265.0037$, and the mean MD is $\frac{(n-1)\,p}{n} = 43.835$, a relatively large value because p is large relative to $n-1$. The 0.9999 quantile in the χ^2_{44} distribution is 87.68.

The largest MD to the centre is for individual 238, with $d^2_{238} = 206.93$. As the number of variables is large ($p > 30$), the `xploremaha` function invokes the `anneal` heuristic search function of `R` package `subselect` to search for subsets of variables that preserve most of this MD to the centre. The number of iterations of the algorithm was set to `niter=50000` and computation times (as measured with `R`'s `system.time` command) were of the order of 75 seconds on a HPZ2 Mini G3 workstation with 16GiB memory. The left plot in Figure 4 gives the results for the selected subsets of all cardinalities k from 1 to $p-1 = 43$. This plot was produced with the command `xploremaha(SPECTF, rank=1, niter=50000)`. Although with a substantial loss, a three-variable subset will still give, for individual 238, an MD to the centre of 119.638 (about 58% of the original MD). The best three-dimensional scatterplot, as indicated by function `xploremaha`, uses variables $V28$, $V32$ and $V33$. With the help of the `R` function `plot3d`, in package `rgl`, a three-dimensional plot of these variables can be rotated to highlight individual 238 (Figure 4, right plot). Individual 238 is the isolated point near the bottom-left of the plot. Its MD to the centre in this 3D plot is 119.638, approximately 58% of the original MD. It should be stressed that, due to the random nature of the simulated annealing search algorithm, these results can change at each run of the `anneal` function.

5.3 Example: the `ISOLET` speech data

A third dataset is `ISOLET` (Isolated Letter Speech Recognition), contributed to the UCI Repository by Tom Dietterich of the Department of Computer Science at Oregon State University and discussed by Ron Cole and Mark Fanty at the Oregon Graduate Institute (Fanty & Cole, 1991). This example shows how the proposed methodology may be useful even for datasets with a fairly large number of variables. The `ISOLET` dataset has $n = 7797$ speech records and, originally, 618 measurements. Four binary variables were excluded

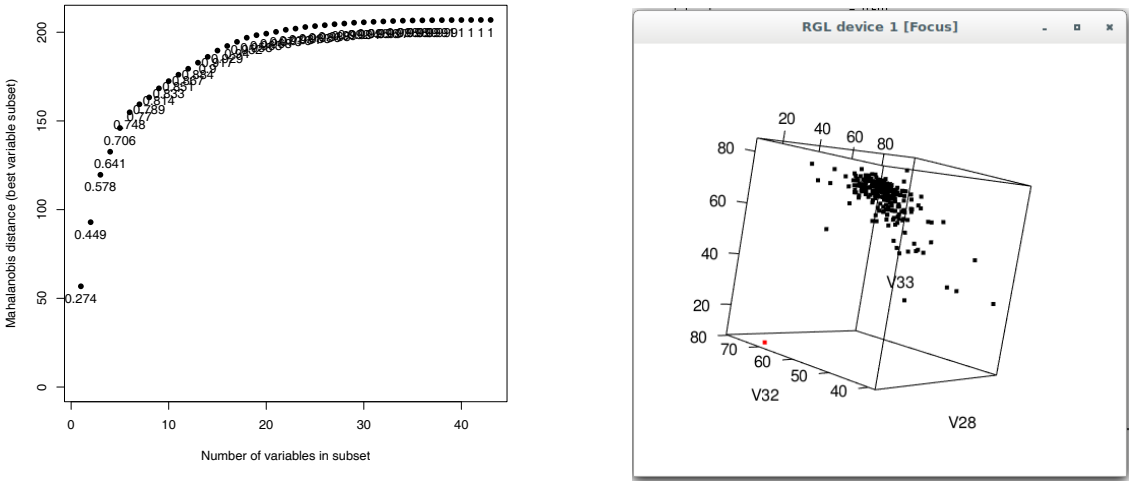

Figure 4: The left plot gives the MDs to the centre of individual 238 in the `SPECTF` dataset, for the selected subsets of $k<44$ variables. The right plot shows the best 3D scatterplot, using variables $V28$, $V32$ and $V33$.

(variables $V578$, $V579$, $V580$ and $V585$) and $p=614$ numerical variables were used. Mahalanobis distances to the centre were computed for each speech record (a computationally light task). The upper bound for MDs to the centre is $\frac{(n-1)^2}{n}=7795$, and the mean MD is $\frac{(n-1)\,p}{n}=613.92$ (which is also a lower bound for the largest MD to the centre). The 0.9999 quantile in the χ^2_{614} distribution is 728.01.

Individual 219 has the largest MD to the centre, with $d^2_{219}=3092.1352$, corresponding to a scaled MD of 0.3967. The large number of variables is computationally demanding even for heuristic algorithms, if all possible cardinalities of subsets from 1 to $p-1$ are searched. It is, however, feasible to look for subsets of only a few smaller cardinalities, by invoking the `xploremaha` function (hence the `anneal` function of package `subselect`) with the optional arguments `kmin` and `kmax`, specifying a range of subset cardinalities that is to be explored. An example of the use of the command to select subsets of all cardinalities from 1 to 20 is shown below. Given the large number of variables ($p=614$), the optional `force=TRUE` argument is used to force the `anneal` function to run. The command ran in less than 40 seconds of system time.

```
> xploremaha(ISOLET, rank=1, kmin=1, kmax=20, force=TRUE)
```

Given the random nature of the simulated annealing algorithm, the selection of the optimal subsets for each cardinality is not guaranteed. This is evident in the left plot of Figure 5, where the selected subset with 7 variables has a lower MD than the selected subset of cardinality 6, a situation which cannot happen with the optimal subsets (Proposition 3.2). For this reason, it is advisable to run the search algorithm more than once. A subset of size $k=20$ which retains over half the original MD has been found, with scaled MD 0.2030. This subset has less than 3.5% of the original variables but retains 51.2% of the MD value. Since a reasonable subset with 20 variables has been identified, the `xploremaha` command can now be run on the resulting 7797×20 sub-matrix of `ISOLET`, which will invoke the `eleaps` command for a guaranteed complete search of subsets of lower cardinality. The result is shown on the right-hand plot of Figure 5. The proportions indicated beneath each point on this right plot are relative to the MD to the centre of individual 219 on the selected 20-variable data subset, $d^2_{219[20]}=1582.961$.

Alternatively, a standard stepwise selection method, such as backward elimination, for the linear regression of the i-th canonical vector of $\mathbb{R}^n$ on the p variables provides a computationally lighter algorithm to search for good subsets of the predictors that retain a large proportion of the original MD to the centre of individual i.

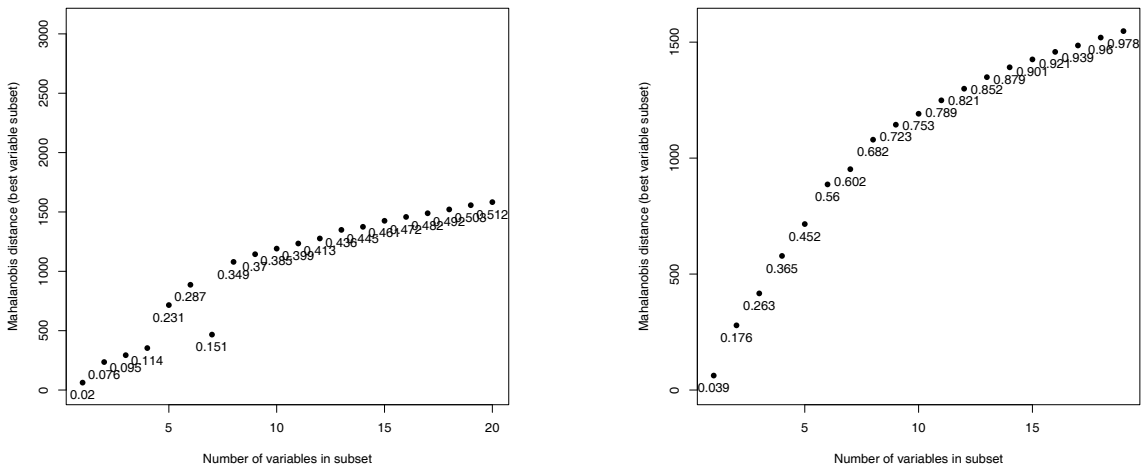

Figure 5: On the left, selected Mahalanobis distances to the centre of individual 219 in the `ISOLET` dataset, for variable subsets of cardinalities 1 to 20. These MD values are non-monotone because the heuristic search algorithm does not guarantee optimal solutions. The sub-matrix of the 20 variable-subset that was selected in this first step was then subjected to a full search of subsets of smaller cardinality, with the resulting plot shown on the right.

6 Mahalanobis distances involving means of groups of individuals

Previous results can be generalised to MDs involving the mean vectors of subgroups of individuals. Thus, if $\vec{\mathbf{m}}_I$ is the vector of p means of the n_I observations of a subgroup I, we can define the sample *Mahalanobis distance of group I to the centre* as:

$$d_I^2 \;=\; (\vec{\mathbf{m}}_I - \vec{\mathbf{m}})^t \mathbf{S}^{-1} (\vec{\mathbf{m}}_I - \vec{\mathbf{m}}) \;, \tag{14}$$

and the sample *Mahalanobis distance between two disjoint groups of individuals, I and J*:

$$d_{IJ}^2 \;=\; (\vec{\mathbf{m}}_I - \vec{\mathbf{m}}_J)^t \mathbf{S}^{-1} (\vec{\mathbf{m}}_I - \vec{\mathbf{m}}_J) \tag{15}$$

where $\vec{\mathbf{m}}_J$ is the mean vector of the n_J individuals in group J. It should be stressed that matrix $\mathbf{S}$ in equation (15) is the variance matrix of the entire set of n individuals, and not the pooled variance matrix which is often used to define MDs between two different samples.

The equations above are essentially analogues of equations (1) and (2). If $\vec{\mathbf{i}}_I$ is the dummy vector identifying the n_I individuals that belong to group I and $\vec{\mathbf{g}}_I = \frac{1}{n_I}\vec{\mathbf{i}}_I$, then $(\vec{\mathbf{m}}_I - \vec{\mathbf{m}})^t = \vec{\mathbf{g}}_I^t \mathbf{X}_c$. Thus, vector $\vec{\mathbf{g}}_I$ plays the role of vector $\vec{\mathbf{e}}_i$ in the MD of individual i to the centre. With similar notation for group j, we have $(\vec{\mathbf{m}}_I - \vec{\mathbf{m}}_J)^t = (\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J)^t \mathbf{X}_c$.

Results similar to those of Sections 3 and 4 follow, since $\vec{\mathbf{g}}_I \notin \mathcal{C}(\vec{\mathbf{1}}_n)^\perp$ (its elements add to one) but $\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J \in \mathcal{C}(\vec{\mathbf{1}}_n)^\perp$ (its elements add to zero). These results are stated in the following Proposition.

Proposition 6.1 *Let $\mathbf{X}$, $\mathbf{X}_m$, $\mathbf{X}_c$ and $\mathbf{P_c}$ be defined as in Section 3. Let I be the set of indices identifying the n_I individuals in a given group, $\vec{\mathbf{i}}_I$ the corresponding indicator (dummy) vector and $\vec{\mathbf{g}}_I = \frac{1}{n_I}\vec{\mathbf{i}}_I$. Then,*

1. *For the MD of group I to the centre, d_I^2, we have:*

   (a) $d_I^2 = {}_{(n-1)}\, \vec{\mathbf{g}}_I^t \mathbf{P_c} \vec{\mathbf{g}}_I = {}_{(n-1)}\, \overline{p}_{[I,I]}$, *where $\overline{p}_{[I,I]}$ is the mean value of the elements of the submatrix of $\mathbf{P_c}$ whose row and column numbers are in set I;*

(b) $d_I^2 = \frac{n-1}{n_I}\cos^2(\vec{\mathbf{i}}_I, \mathbf{P_c}\vec{\mathbf{i}}_I) = \frac{n-1}{n_I}\frac{n-n_I}{n}\cos^2(\vec{\mathbf{i}}_I^*, \mathbf{P_c}\vec{\mathbf{i}}_I^*) = \frac{n-1}{n}\frac{n-n_I}{n_I}R_I^2$, *where* $\vec{\mathbf{i}}_I^* = (\mathbf{I}_n - \mathbf{P}_{\vec{\mathbf{1}}_n})\vec{\mathbf{i}}_I$ *and* R_I^2 *is the coefficient of determination in the multiple linear regression of the dummy variable* $\vec{\mathbf{i}}_I$ *on the* p *variables in* $\mathbf{X}$.

(c) The maximum MD of a group I *to the center,* $d_I^2 = \frac{n-1}{n}\frac{n-n_I}{n_I}$, *is attained iff*

$$R_I^2 = 1 \quad \Leftrightarrow \quad \vec{\mathbf{i}}_I \in \mathcal{C}(\mathbf{X}_m) \quad \Leftrightarrow \quad \vec{\mathbf{i}}_I^* \in \mathcal{C}(\mathbf{X}_c).$$

2. For the Mahalanobis distance between groups I *and* J, d_{IJ}^2, *we have, with similar notation:*

(a) $d_{IJ}^2 = {}_{(n-1)}(\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J)^t\mathbf{P_c}(\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J) = {}_{(n-1)}(\overline{p}_{[I,I]} + \overline{p}_{[J,J]} - 2\overline{p}_{[I,J]})$ *where* $\overline{p}_{[I,J]}$ *is the mean value of the elements of the submatrix of* $\mathbf{P_c}$ *with row numbers in set* I *and column numbers in set* J.

(b) $d_{IJ}^2 = {}_{(n-1)}\frac{n_I+n_J}{n_I\,n_J}\cos^2\left(\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J, \mathbf{P_c}(\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J)\right) = {}_{(n-1)}\frac{n_I+n_J}{n_I\,n_J}R_{IJ}^2$, *where* R_{IJ}^2 *is the coefficient of determination of the multiple linear regression of* $\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J$ *on the* p *variables in* $\mathbf{X}$.

(c) The maximum MD between groups I *and* J, $d_{IJ}^2 = {}_{(n-1)}\left(\frac{1}{n_I} + \frac{1}{n_J}\right)$, *is attained iff*

$$R_{IJ}^2 = 1 \quad \Leftrightarrow \quad \vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J \in \mathcal{C}(\mathbf{X}_m) \quad \Leftrightarrow \quad \vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J \in \mathcal{C}(\mathbf{X}_c).$$

Proof

1. (a) $\vec{\mathbf{g}}_I^t\mathbf{P_c}\vec{\mathbf{g}}_I = \frac{1}{n_I^2}\vec{\mathbf{i}}_I^t\mathbf{P_c}\vec{\mathbf{i}}_I$, with $\vec{\mathbf{i}}_I^t\mathbf{P_c}\vec{\mathbf{i}}_I$ giving the sum of the n_I^2 values of the principal $n_I \times n_I$ submatrix of $\mathbf{P_c}$ associated with the rows/columns of individuals in group I.

   (b) From the previous point, $d_I^2 = (n-1)\,\vec{\mathbf{g}}_I^t\mathbf{P_c}\vec{\mathbf{g}}_I = \frac{n-1}{n_I^2}\vec{\mathbf{i}}_I^t\mathbf{P_c}\vec{\mathbf{i}}_I$. From equation (5), $\vec{\mathbf{i}}_I^t\mathbf{P_c}\vec{\mathbf{i}}_I = \vec{\mathbf{i}}_I^t\vec{\mathbf{i}}_I \cdot \cos^2(\vec{\mathbf{i}}_I, \mathbf{P_c}\vec{\mathbf{i}}_I)$. Now $\vec{\mathbf{i}}_I^t\vec{\mathbf{i}}_I = \|\vec{\mathbf{i}}_I\|^2 = n_I$, so $d_I^2 = \frac{n-1}{n_I}\cos^2(\vec{\mathbf{i}}_I, \mathbf{P_c}\vec{\mathbf{i}}_I)$, which gives the first equation. Since $\mathcal{C}(\mathbf{X}_c) \subseteq \mathcal{C}(\vec{\mathbf{1}}_n)^\perp$, it follows from equation (7) that $\cos^2(\vec{\mathbf{i}}_I, \mathbf{P_c}\vec{\mathbf{i}}_I) = \cos^2(\vec{\mathbf{i}}_I, \vec{\mathbf{i}}_I^*)\cdot\cos^2(\vec{\mathbf{i}}_I^*, \mathbf{P_c}\vec{\mathbf{i}}_I)$, defining $\vec{\mathbf{i}}_I^* = (\mathbf{I}_n - \mathbf{P}_{\vec{\mathbf{1}}_n})\vec{\mathbf{i}}_I$. But, again because $\mathcal{C}(\mathbf{X}_c) \subseteq \mathcal{C}(\vec{\mathbf{1}}_n)^\perp$, $\mathbf{P_c}(\mathbf{I}_n - \mathbf{P}_{\vec{\mathbf{1}}_n}) = \mathbf{P_c}$, hence $\mathbf{P_c}\vec{\mathbf{i}}_I = \mathbf{P_c}\vec{\mathbf{i}}_I^*$. Therefore, $\cos^2(\vec{\mathbf{i}}_I, \mathbf{P_c}\vec{\mathbf{i}}_I) = \cos^2(\vec{\mathbf{i}}_I, \vec{\mathbf{i}}_I^*) \cdot \cos^2(\vec{\mathbf{i}}_I^*, \mathbf{P_c}\vec{\mathbf{i}}_I^*)$. From equation (5), $\cos^2(\vec{\mathbf{i}}_I, \vec{\mathbf{i}}_I^*) = \frac{\|\vec{\mathbf{i}}_I^*\|^2}{\|\vec{\mathbf{i}}_I\|^2} = \frac{n-n_I}{n}$, since $\vec{\mathbf{i}}_I^*$ has n_I elements equal to $\frac{n-n_I}{n}$ and $n - n_I$ elements equal to $\frac{-n_I}{n}$, so that $\|\vec{\mathbf{i}}_I^*\|^2 = \frac{n_I}{n}(n - n_I)$. Hence, $d_I^2 = \frac{n-1}{n_I}\cos^2(\vec{\mathbf{i}}_I, \mathbf{P_c}\vec{\mathbf{i}}_I) = \frac{n-1}{n_I}\frac{n-n_I}{n}\cos^2(\vec{\mathbf{i}}_I^*, \mathbf{P_c}\vec{\mathbf{i}}_I^*)$. Finally, from equation (8), $\cos^2(\vec{\mathbf{i}}_I^*, \mathbf{P_c}\vec{\mathbf{i}}_I^*) = R_I^2$.

   (c) The maximum value for d_I^2 is attained iff $R_I^2 = 1$, in which case $\vec{\mathbf{i}}_I \in \mathcal{C}(\mathbf{X}_m)$. A reasoning similar to that used in point 11 of Proposition 3.1 proves that this is equivalent to stating that $\vec{\mathbf{i}}_I^* \in \mathcal{C}(\mathbf{X}_c)$.

2. (a) $(\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J)^t\mathbf{P_c}(\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J) = \vec{\mathbf{g}}_I^t\mathbf{P_c}\vec{\mathbf{g}}_I + \vec{\mathbf{g}}_J^t\mathbf{P_c}\vec{\mathbf{g}}_J - 2\vec{\mathbf{g}}_I^t\mathbf{P_c}\vec{\mathbf{g}}_J = \frac{1}{n_I^2}\vec{\mathbf{i}}_I^t\mathbf{P_c}\vec{\mathbf{i}}_I + \frac{1}{n_J^2}\vec{\mathbf{i}}_J^t\mathbf{P_c}\vec{\mathbf{i}}_J - 2\frac{1}{n_I n_J}\vec{\mathbf{i}}_I^t\mathbf{P_c}\vec{\mathbf{i}}_J$. As above, the value of an expression such as $\vec{\mathbf{i}}_I^t\mathbf{P_c}\vec{\mathbf{i}}_J$ is the sum of all elements in the $n_I \times n_J$ submatrix of $\mathbf{P_c}$ with row numbers in I and column numbers in J, hence the result.

   (b) The first equality results from the point above and equation (5), since $\|\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J\|^2 = \frac{n_I+n_J}{n_I n_J}$. In fact, vector $\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J$ has n_I elements of value $\frac{1}{n_I}$ and n_J elements $\frac{-1}{n_J}$, with any remaining elements equal to zero. Since vector $\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J \in \mathcal{C}(\vec{\mathbf{1}}_n)^\perp$, we have $\cos^2(\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J, \mathbf{P_c}(\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J)) = R_{IJ}^2$, the coefficient of determination of the linear regression of vector $\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J$ on the p columns of $\mathbf{X}$.

   (c) The maximum value of d_{IJ}^2 is achieved if $R_{IJ}^2 = 1$, that is if $\vec{\mathbf{g}}_I - \vec{\mathbf{g}}_J \in \mathcal{C}(\mathbf{X}_m)$. The other equivalence follows as in previous cases.

Point 1c) implies that the maximum MD of a group to the center, $\frac{n-1}{n}\frac{n-n_I}{n_I}$, is attained if and only if there is some set of coefficients $\{b_j\}_{j=0}^p$ such that $\vec{\mathbf{i}}_I = b_0\vec{\mathbf{1}}_n + b_1\vec{\mathbf{x}}_1 + \ldots + b_p\vec{\mathbf{x}}_p$, i.e., iff there is a linear discriminant axis $b_1\vec{\mathbf{x}}_1 + \ldots + b_p\vec{\mathbf{x}}_p$ which perfectly separates the individuals in group I (with common coefficient $1-b_0$) from the others (with coefficient $-b_0$), for some $b_0 \in \mathbb{R}$. Equivalently, there are two parallel affine hyperplanes in

$\mathbb{R}^p$, one of which (with equation $b_0 + b_1\, x_1 + ... + b_p\, x_p = 1$) contains the points in group I and the other (with equation $b_0 + b_1\, x_1 + ... + b_p\, x_p = 0$) contains the remaining points.

Similarly, point 2c) states that the maximum MD between two groups, $d^2_{IJ} = {}_{(n-1)}\frac{n_I+n_J}{n_I\, n_J}$, is attained if and only if there is a linear discriminant axis $b_1\vec{\mathbf{x}}_1 + ... + b_p\vec{\mathbf{x}}_p$, which perfectly separates the individuals in group I (with common coefficient $-b_0+\frac{1}{n_I}$) and the individuals in group J (with common coefficient $-b_0-\frac{1}{n_J}$) from the remaining individuals (with coefficient $-b_0$). Equivalently, if $n > n_I+n_J$, there are three parallel affine hyperplanes in $\mathbb{R}^p$ containing all points: the hyperplane with equation $b_0 + b_1\, x_1 + ... + b_p\, x_p = 1$ on which lie the n_I points corresponding to individuals in group I, the hyperplane $b_0+b_1\, x_1+...+b_p\, x_p = -1$ on which lie the n_J points corresponding to individuals in group J, and the hyperplane $b_0+b_1\, x_1+...+b_p\, x_p = 0$ on which lie the remaining $n-(n_I+n_J)$ points.

7 Discussion

Mahalanobis distances have traditionally been interpreted in the space of individuals, where each axis is associated with one of p variables, and each observed individual with a point in $\mathbb{R}^p$. Here, we argue that sample-based MDs defined by an $n \times p$ data set can be better understood by also considering the alternative setting of the space of variables, $\mathbb{R}^n$, where each axis corresponds to an observed individual and each variable is represented by a vector. In this alternative setting, Mahalanobis distances of individuals to the centre depend only on the sample size and the angles formed by each axis with the column-space of the column-centred data matrix $\mathbf{X}_c$. The squared cosines of the angles between the centred canonical vectors for each axis and $\mathcal{C}(\mathbf{X}_c)$ are a scaled MD, in the interval $[0, 1]$, which can be used as an index for the severity of outliers that does not depend on any assumptions regarding a probability distribution that may be associated with the dataset. In turn, MDs between individuals are defined by the sample size and the angle formed by the difference in the canonical basis vectors associated with the two individuals and the subspace $\mathcal{C}(\mathbf{X}_c)$. This setting provides geometric insight and explanation for results regarding MDs that are surprisingly recent (Gath & Hayes, 2006; Pires & Branco, 2018).

None of the above mentioned angles corresponds to what Mardia (1977) called "Mahalanobis angles". Mardia defined a matrix $\mathbf{G}$ whose generic element is $g_{ij} = (\vec{\mathbf{x}}_{[i]} - \vec{\mathbf{m}})^t\mathbf{S}^{-1}(\vec{\mathbf{x}}_{[j]} - \vec{\mathbf{m}})$. In the notation used here, $\mathbf{G} = (n-1)\mathbf{P_c}$, though Mardia (1977) did not explore either the fact that matrix $\frac{1}{n-1}\mathbf{G}$ is an orthogonal projection matrix, or the geometry of MDs in $\mathbb{R}^n$. Mardia's Mahalanobis angles have cosines $\frac{g_{ij}}{\sqrt{g_{ii}\cdot g_{jj}}} = \frac{p_{ij}}{\sqrt{p_{ii}\cdot p_{jj}}} = \frac{(\mathbf{P_c}\vec{\mathbf{e}}_i)^t(\mathbf{P_c}\vec{\mathbf{e}}_j)}{\|\mathbf{P_c}\vec{\mathbf{e}}_i\|\cdot\|\mathbf{P_c}\vec{\mathbf{e}}_j\|}$ and are therefore the angles in $\mathbb{R}^n$ between vectors $\mathbf{P_c}\vec{\mathbf{e}}_i$ and $\mathbf{P_c}\vec{\mathbf{e}}_j$. These angles appear to be less useful than those described above in understanding the properties of sample-based MDs. The vectors $\mathbf{P_c}\vec{\mathbf{e}}_i$ and $\mathbf{P_c}\vec{\mathbf{e}}_j$ are directly tied to MDs through norms, since from equations (3) and (4) we have $d^2_i = {}_{(n-1)}\,\|\mathbf{P_c}\vec{\mathbf{e}}_i\|^2$ and $d^2_{ij} = {}_{(n-1)}\,\|\mathbf{P_c}\vec{\mathbf{e}}_i - \mathbf{P_c}\vec{\mathbf{e}}_j\|^2$.

The connection between MDs to the centre and the linear regression of the canonical vectors of $\mathbb{R}^n$ on the p variables in the dataset suggests a simple procedure to identify those variables that are mainly responsible for the value of any given MD, which was discussed and illustrated by examples in Section 5. A good low-dimensional subset of variables may (if it preserves a large proportion of the original MD) assist in interpreting outliers. Conceptually, the methodology suggested in Section 5 can be applied to datasets of any size. From a computational point of view, the limiting factor is the search for the optimal subsets of predictors in the linear regressions, and therefore the number of variables p in the dataset. For datasets without too many variables (p up to approximately 30), efficient branch-and-bound algorithms can identify the best subsets of each cardinality. The `eleaps` function in `subselect` package is used in Section 5, but the `leaps` function in the `R` package with the same name (Lumley, 2017) provides an alternative. For datasets with large p, the simple and well-known stepwise search methods in regressions can drastically reduce computation times. Any alternative search method for the optimal variable subsets may be used, as illustrated in Section 5.

Similar procedures may be used to highlight the variables that are chiefly responsible for large MDs between pairs of individuals. As was seen in Section 4, the starting point in that case is the linear regression of the difference between the canonical vectors for the two individuals of interest, on the p variables in the dataset.

Inevitably, for datasets with a large number of variables, many different subsets of a given cardinality may preserve similar proportions of the original MDs, so that low-dimensional interpretations (and possibly visualizations) of the reasons for an outlying individual are not unique. Alternative sub-optimal subsets may be of interest. There is scope for future work in exploring this issue.

The methodology proposed in Section 5 may be fully automated if rules are provided to specify which MDs are of interest and what proportion of the original MD must be retained in a subset. However, sound data analysis still requires human intervention in making those decisions, by someone familiar with the problem under consideration. Hence, the procedure indicated here is best considered semi-automated.

The results for MDs associated with individual observations can be generalised to mean vectors of groups of individuals by replacing the canonical vectors in $\mathbb{R}^n$, that correspond to each observation, with the dummy vectors for the observations in each group, as discussed in Section 6.

Proposition 6.1 has important implications for MDs of individuals to the centre and between individuals, in data matrices with repeated rows. If there is a group I of n_I repeated rows in $\mathbf{X}$, their vector of group means will coincide with each of those rows. Thus, for any individual i in group I, we have $\vec{\mathbf{x}}_{[i]} = \vec{\mathbf{m}}_I$, hence:

$$d_i^2 \;=\; d_I^2 \;\leq\; \frac{n-1}{n} \cdot \frac{n-n_I}{n_I} \;<\; \frac{1}{n_I}\, M \;, \tag{16}$$

where $M = \frac{(n-1)^2}{n}$ is the maximum possible distance to the centre of any non-repeated individual.

Likewise, there is a smaller upper bound than $2(n-1)$ for the MD between two individuals i and j if other individuals have rows in the data matrix identical to those of individuals i and/or j. In fact, if individuals i and j belong to groups I and J of n_I and n_J repeated individuals, then $\vec{\mathbf{x}}_{[i]} = \vec{\mathbf{m}}_I$ and $\vec{\mathbf{x}}_{[j]} = \vec{\mathbf{m}}_J$. Hence,

$$d_{ij}^2 \;=\; d_{IJ}^2 \;\leq\; (n-1) \cdot \left[\frac{1}{n_I} + \frac{1}{n_J}\right]. \tag{17}$$

For any pair of observations i and j in a tightly-knit group of observations, $d_{ij}^2 \approx 0$ and $d_i^2 \approx d_j^2$. In that case, equation (4) implies that $p_{ij} \approx \frac{d_i^2}{n-1} = p_{ii} \approx p_{jj}$. Thus, all elements in the submatrix of $\mathbf{P_c}$ with row/column numbers in I will be approximately equal. The closer these similar values are to the maximum group MD to the centre divided by $n-1$, i.e., to $\frac{n-n_I}{n \cdot n_I}$, the more outlying will be that group of individuals.

These comments regarding vectors of group means suggest future lines of work to uncover further interesting properties of Mahalanobis distances.

Another interesting line for future work is the study of analogous properties of robust alternatives to standard Mahalanobis distances, which may also have geometric interpretations in the space of variables.

More generally, the results above confirm that the space of variables is a natural space for understanding and interpreting statistical methodologies. Its widespread use is to be recommended, given that many statistical concepts have clear geometric meaning in this space. Geometric intuition may assist in understanding their properties.

A The `xploremaha` `R` function

```
xploremaha <- function(data, rank, plot=TRUE, switch=30, kmin=1, kmax=dim(data)[2]-1,
+                                        niter=150000, force=FALSE, ...){
require(subselect)
data.maha <- mahalanobis(data, center=apply(data,2,mean), cov=var(data))
n <- dim(data)[1]
p <- dim(data)[2]
if (kmax == p) {kmax <- kmax-1}
valor <- rev(sort(data.maha))[rank]
valorR2 <- valor*n/(n-1)^2
qual <- which(data.maha == valor)
ei <- function(n,i){rep(c(0,1,0),c(i-1,1,n-i))}
```

```
aux <- lmHmat(ei(n=n,i=qual) ~ . , data=data)
if (p <= switch){
temp <- eleaps(mat=aux$mat, kmin=kmin, kmax=kmax, H=aux$H, r=1, crit="ccr12")
}
else
{
temp <- anneal(mat=aux$mat, kmin=kmin, kmax=kmax, H=aux$H, r=1, crit="ccr12",
+                            niter=niter, force=force, cooling=0.01, coolfreq=50)
}
md <- temp$bestvalues*(n-1)^2/n
if (plot){
plot(kmin:kmax, md, pch=16, ylim=c(0,valor), xlab="Number of variables in subset",
+                                    ylab="Mahalanobis distance (best variable subset)")
text(kmin:kmax, md-valor/30, labels=round(md/valor, d=3))
}
list(bestsets=temp$bestsets, bestvalues=temp$bestvalues)
}
```

Acknowledgements

This research was supported by the Portuguese Foundation FCT - UID/MAT/00006/2019. The author thanks the reviewers and editor for their suggestions.

References

Anderson, T. (1958). *An introduction to multivariate statistical analysis*. John Wiley & Sons.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (Third ed.). John Wiley & Sons.

Branco, J., & Pires, A. (2011). *Travelling through multivariate data spaces with Mahalanobis distance.* (Presentation slides in JOCLAD 2011 conference, in Vila Real, Portugal)

Cadima, J., Cerdeira, J., & Minhoto, M. (2004). Computational aspects of algorithms for variable selection in the context of principal components. *Computational Statistics and Data Analysis*, *47*, 225-236.

Dheeru, D., & Karra Taniskidou, E. (2017). *UCI Machine Learning Repository.* Retrieved from `http://archive.ics.uci.edu/ml`

Duarte Silva, A. (2001). Efficient variable screening for multivariate analysis. *Journal of Multivariate Analysis*, *76(1)*, 35-62.

Duarte Silva, A. (2002). Discarding variables in a principal component analysis: algorithms for all-subsets comparisons. *Computational Statistics*, *17*, 251-271.

Fanty, M., & Cole, R. (1991). Spoken letter recognition. In *Advances in neural information processing systems 3, Lippman, R. P., Moody, J., and Touretzky, D. S. (eds).* (p. 220-226). Morgan Kaufmann, San Mateo, CA.

Furnival, G., & Wilson, R. (1974). Regressions by leaps and bounds. *Technometrics*, *16*, 499-511.

Gath, E., & Hayes, K. (2006). Bounds for the largest Mahalanobis distance. *Linear Algebra and its Applications*, *419*, 93-106.

Horn, R., & Johnson, C. (1985). *Matrix analysis.* Cambridge University Press.

Johnson, R., & Wichern, D. (2007). *Applied multivariate statistical analysis* (Sixth ed.). Pearson Prentice-Hall, NJ.

Kurgan, L., Cios, K., Tadeusiewicz, R., Ogiela, M., & Goodenday, L. (2001). Knowledge discovery approach to automated cardiac SPECT diagnosis. *Artificial Intelligence in Medicine*, *23:2*, 149-169.

Lumley, T. (2017). leaps: Regression subset selection [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=leaps` (R package version 3.0, based on Fortran code by Alan Miller)

Mahalanobis, P. (1936). On the generalized distance in statistics. *Proc.Nat.Inst.Sci.India (Calcutta)*, *2*, 49-55.

Mardia, K. (1977). Mahalanobis distances and angles. In *Multivariate Analysis - IV* (p. 495-511). North-Holland Publishing Company.

Morrison, D. (1990). *Multivariate statistical methods* (third ed.). McGraw-Hill International Editions.

Nash, W., Sellers, T., Talbot, S., Cawthorn, A., & Ford, W. (1994). *The population biology of abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait* (Tech. Rep.). Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288).

Olkin, I. (1992). A matrix formulation for how deviant an observation can be. *The American Statistician*, *46*, 205-209.

Orestes Cerdeira, J., Duarte Silva, P., Cadima, J., & Minhoto, M. (2023). subselect: Selecting variable subsets [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=subselect` (R package version 0.15.4)

Pires, A. M., & Branco, J. A. (2018). *High dimensionality: the latest challenge to data analysis.* (Working paper in Statistics, Instituto Superior Técnico, University of Lisbon)

Puntanen, S., Styan, G., & Isotalo, J. (2011). *Matrix tricks for linear statistical models.* Springer-Verlag, Berlin Heidelberg.

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`