Goodness-of-fit filtering in classical metric multidimensional scaling with large datasets

Jan Graffelman ^{a,b}

^aDepartment of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain;

^bDepartment of Biostatistics, University of Washington, Seattle, WA, USA

ABSTRACT

Metric multidimensional scaling (MDS) is a widely used multivariate method with applications in almost all scientific disciplines. Eigenvalues obtained in the analysis are usually reported in order to calculate the overall goodness-of-fit of the distance matrix. In this paper, we refine MDS goodness-of-fit calculations, proposing additional point and pairwise goodness-of-fit statistics that can be used to filter poorly represented observations in MDS maps. The proposed statistics are especially relevant for large data sets that contain outliers, with typically many poorly fitted observations, and are helpful for improving MDS output and emphasizing the most important features of the dataset. Several goodness-of-fit statistics are considered, and both Euclidean and non-Euclidean distance matrices are considered. Some examples with data from demographic, genetic and geographic studies are shown.

ARTICLE HISTORY

Received 19 July 2019

Accepted 3 December 2019

KEYWORDS

Plot brushing; outlier; attractor point; eigenvalue; Manhattan distance; allele sharing distance

1. Introduction

Multidimensional scaling (MDS) is a versatile multivariate technique that has found application in many branches of science. The goal of the method is to construct a configuration of points in a low-dimensional space, such that interpoint distances in this configuration approximate the entries of a given distance matrix [10,15,23]. In this era of large datasets, MDS applications can involve distance matrices with hundreds or thousands of rows, leading to dense maps with many points. Split-and-combine [24] and projection algorithms [16] have been proposed to make it computationally feasible to analyse very large datasets by metric MDS. Not all observations are equally well represented in MDS maps, and some observations may have a poor goodness-of-fit, in the sense that their interpoint distances with other observations poorly approximate the corresponding entries in the original distance matrix. In classical metric MDS, eigenvalues (which represent variance accounted for [3]) are used to assess the overall goodness-of-fit of the map, but goodness-of-fit statistics at the level of individual points or for a single pair of observations are lacking. The main idea of this paper is to develop point and pairwise goodness-of-fit statistics, in

CONTACT Jan Graffelman jan.graffelman@upc.edu Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Avinguda Diagonal 647, Barcelona 08028, Spain

Supplemental data for this article can be accessed here. <https://doi.org/10.1080/02664763.2019.1702929>

order to use them for *plot brushing*: by not plotting the poorly represented points the final map will be less dense, and the main features of the data set are better emphasized. At the same time, brushing also avoids misinterpretations based on poorly represented observations. In applications, one is often tempted to interpret all points in the same way, as if all points were equally well represented, but in practice, large differences in goodness-of-fit among observations do often exist. The remainder of this paper is structured as follows. In Section 2, we briefly summarize the theory of classical MDS and develop point and pairwise goodness-of-fit statistics. In Section 3, we apply our statistics to examples with datasets taken from demography, genetics and geography. Section 4 finishes the article with a discussion.

2. Theory

We summarize classical metric MDS and establish our notation in Section 2.1. We address point and pairwise goodness-of-fit measures for (pairs of) observations of Euclidean distance matrices in Section 2.2, and discuss analogous measures for non-Euclidean distance matrices in Section 2.3.

2.1. Classical metric MDS

Classical metric MDS, also known as *classical scaling* or *principal coordinate analysis* (PCO), is a standard topic in courses on multivariate analysis. Text books on multivariate analysis usually dedicate a chapter to this method [10,15]. The books of Borg and Groenen [1] and Cox and Cox [4] are entirely dedicated to the topic. In this paper, we confine ourselves to principal coordinate analysis, which is popular in many fields. Non-metric methods [11] and metric methods based on iterative stress minimization [5,12] are not considered here. In PCO [6], the approximation of the distance matrix is indirect, via the scalar product matrix $\mathbf{B}$, which is obtained by scaling and double-centring the matrix of squared distances

$$\mathbf{A} = -\frac{1}{2}\mathbf{D}^{(2)}, \quad \mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}, \quad (1)$$

where $\mathbf{D}^{(2)}$ refers to the $n \times n$ distance matrix with squared entries and $\mathbf{H}$ is the centring matrix $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$. The scalar product matrix $\mathbf{B}$ is decomposed by the spectral decomposition

$$\mathbf{B} = \mathbf{V}\mathbf{D}_\lambda\mathbf{V}', \quad (2)$$

where $\mathbf{V}$ is the $n \times n$ matrix of orthogonal eigenvectors ($\mathbf{V}'\mathbf{V} = \mathbf{I}_n$) and $\mathbf{D}_\lambda$ is an $n \times n$ diagonal matrix containing the eigenvalues of $\mathbf{B}$ in non-increasing order of magnitude ($\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$). The coordinates of the observations in the MDS map are obtained by

$$\mathbf{X} = \mathbf{V}\mathbf{D}_\lambda^{\frac{1}{2}}. \quad (3)$$

Typically, a two-dimensional representation is made by only using the first two eigenvectors and eigenvalues. The Euclidean distances between the rows of $\mathbf{X}$, which we represent by $\hat{\mathbf{D}}$, are then used to approximate the original entries of $\mathbf{D}$. If there are negative eigenvalues,

then only the eigenvectors and eigenvalues corresponding to non-negative eigenvalues are used to calculate the solution.

We will use the following notation in our goodness-of-fit calculations, and let g indicate the overall goodness-of-fit of a k -dimensional map, g_i the goodness-of-fit of the i th observation and g_{ij} the goodness-of-fit of the distance between observations i and j . After creating an MDS plot as outlined above, the question arises whether the map gives a good approximation to the original distance matrix. If the distance matrix is Euclidean, with $\mathbf{B}$ having only non-negative eigenvalues, then goodness-of-fit of a k -dimensional solution is usually assessed by computing the statistic

$$g = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^n \lambda_i} = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^K \lambda_i}, \quad (4)$$

where K is the rank of $\mathbf{B}$, such that the trailing $n-K$ zero eigenvalues can be ignored. A common choice is $k = 2$, though in many applications dimensions beyond the second can be informative. If $\mathbf{B}$ has negative eigenvalues, the denominator of Equation (4) is often adjusted by taking absolute values of the eigenvalues or by considering positive eigenvalues only. Statistic g corresponds to the fraction of the total sum-of-squares of the squared distances that is accounted for by the k -dimensional approximation, and this fraction can also be obtained as

$$g = \frac{\mathbf{1}'\hat{\mathbf{D}}^{(2)}\mathbf{1}}{\mathbf{1}'\mathbf{D}^{(2)}\mathbf{1}}, \quad (5)$$

where $\hat{\mathbf{D}}^{(2)}$ is the approximation obtained by using the first k columns of $\mathbf{X}$ only. The eigenvalues obtained in PCO are proportional to the eigenvalues obtained by a principal component analysis (PCA) of the original data matrix from which the distance matrix $\mathbf{D}$ has been calculated, in case such a data matrix is available [15]. In PCA, these eigenvalues represent the goodness-of-fit of the centred data matrix, leading to the surprising situation that the latter would *equal* the goodness-of-fit of the distance matrix if Equation (4) or (5) is used as the criterion.

2.2. Euclidean distance matrices

The total variability in the distance matrix can be expressed as the sum of all squared distances, and this variability is also proportional to the sum of all eigenvalues and the trace of $\mathbf{B}$. We have

$$\text{tr}(\mathbf{B}) = 2n \sum_{i=1}^n \lambda_i = \sum_{i,j}^n d_{ij}^2. \quad (6)$$

This quantity can be decomposed over dimensions and over observations. This decomposition is given by the $n \times n$ matrix $\mathbf{Q}_d$, obtained by

$$\mathbf{Q}_d = \mathbf{X} \odot \mathbf{X}, \quad (7)$$

with $\mathbf{X} = \mathbf{V}\mathbf{D}_{\lambda}^{\frac{1}{2}}$, and $\odot$ representing the Hadamard product. We use subindex d to emphasize this is the decomposition of the total sum of squared distances. Matrix $\mathbf{Q}_d$

satisfies

$$\mathbf{1}'\mathbf{Q}_d = \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_{n-1}, \lambda_n = 0).$$

Each row makes a contribution to the total sum-of-squares, and these contributions are given by $\mathbf{w} = \mathbf{Q}_d\mathbf{1}$. For the i th row, this contribution is

$$w_i = \sum_{j=1}^n x_{ij}^2 \quad (8)$$

where x_{ij} represents the entry of row i and column j of the solution matrix $\mathbf{X}$. The goodness-of-fit for a particular point (g_i) in a k -dimensional solution can then be calculated as

$$g_i = \frac{\sum_{j=1}^k x_{ij}^2}{\sum_{j=1}^n x_{ij}^2} = \frac{\sum_{j=1}^k \lambda_j v_{ij}^2}{\sum_{j=1}^n \lambda_j v_{ij}^2}, \quad (9)$$

where v_{ij} represents the i th element of eigenvector j , the j th column of $\mathbf{V}$. Statistic g_i indicates how well the contribution of the i th row is accounted for in k dimensions. At the same time, g_i is the ratio of the squared Euclidean distance between point i and the origin in the map and the squared Euclidean distance between point i and the origin in the full space. A point will have a large goodness-of-fit if the first k eigenvalues are large, and also if it has a large distance from the origin. The overall goodness-of-fit of a k -dimensional approximation to the distance matrix is seen to be a weighted average of the goodness-of-fit of each row:

$$\frac{\sum_{i=1}^n g_i w_i}{\sum_{i=1}^n w_i} = \frac{\sum_{i=1}^n \frac{\sum_{j=1}^k \lambda_j v_{ij}^2}{w_i} w_i}{\sum_{j=1}^n \lambda_j} = \frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^n \lambda_j}, \quad (10)$$

where the weights are the contributions of each row to the total sum-of-squares. Pair-wise goodness-of-fit statistics can also be developed. They indicate how well the distance between a particular pair of points is represented. They may be considered more interesting, since our goal is the representation of interpoint distances. We use g_{ij} to refer to the goodness-of-fit of the distance between points i and j . A natural measure is

$$g_{ij} = \frac{\hat{d}_{ij}^2}{d_{ij}^2} = \frac{\sum_{l=1}^k (x_{il} - x_{jl})^2}{\sum_{l=1}^n (x_{il} - x_{jl})^2}, \quad (11)$$

which for Euclidean distance matrices satisfies $0 \leq g_{ij} \leq 1$.

2.3. Non-Euclidean distance matrices

For non-Euclidean distance matrices, a perfect representation of the distance matrix in high-dimensional space is not possible, as is evidenced by the existence of negative eigenvalues in the solution. This complicates the definition of adequate goodness-of-fit measures, both globally, as well as for individual observations and pairs of observations. For an Euclidean distance matrix, the amount of error in a low dimensional approximation will decrease as more dimensions are considered. Finally, an Euclidean distance matrix will

be perfectly represented if all non-trivial dimensions with $\lambda > 0$ are considered. We consider several approaches for obtaining point and pair-wise goodness-of-fit statistics in the non-Euclidean case outlined in the subsections below. We subsequently use an Euclidean subspace, the scalar product matrix and error statistics for goodness-of-fit calculations.

Euclidean subspace

For a non-Euclidean distance matrix, the amount of error in a low dimensional representation will at first decrease as more dimensions are considered, but only up to a certain limiting number of dimensions ℓ . This number ℓ is given by the number of positive eigenvalues that are larger than the absolute value of the last most negative eigenvalue obtained in the analysis. Another particularity is that the coordinates of the solution, as calculated by Equation (3), are only defined for those dimensions that have non-negative eigenvalues. Let p be the number of non-negative eigenvalues. We will have n eigenvalues, but only p coordinates are defined, and only up to ℓ coordinates improve the representation of the distance matrix. Let X_ℓ contain the first ℓ coordinates of the solution only. We can compute an estimate of the original distance matrix by computing the Euclidean distances between the observations in ℓ dimensions only. One could use the same pointwise goodness-of-fit index as in Equation (9), but using only ℓ dimensions in the denominator, and using at most $k \leq \ell$ dimensions.

$$g_i^\ell = \frac{\sum_{j=1}^k x_{ij}^2}{\sum_{j=1}^\ell x_{ij}^2} = \frac{\sum_{j=1}^k \lambda_j v_{ij}^2}{\sum_{j=1}^\ell \lambda_j v_{ij}^2}. \quad (12)$$

A pairwise measure, analogous to Equation (11), is

$$g_{ij}^\ell = \frac{\hat{d}_{ij}^2}{\tilde{d}_{ij}^2} = \frac{\sum_{l=1}^k (x_{il} - x_{jl})^2}{\sum_{l=1}^\ell (x_{il} - x_{jl})^2}. \quad (13)$$

These measures assume that $\ell \geq 2$.

Scalar product matrix

Mardia [14,15] has suggested the use of the *squared* eigenvalues for non-Euclidean distance matrices. Because the distances are approximated indirectly, via the scalar product matrix $\mathbf{B}$, one could report the goodness-of-fit of the latter instead. The total sum-of-squares of $\mathbf{B}$ is given by $\text{tr}(\mathbf{B}'\mathbf{B}) = \sum_{i=1}^n \lambda_i^2$, and therefore the goodness-of-fit of $\mathbf{B}$ is obtained by using the *squared* eigenvalues

$$g = \frac{\sum_{i=1}^k \lambda_i^2}{\sum_{i=1}^n \lambda_i^2}. \quad (14)$$

Akin to Equation (7), we now have the decomposition

$$\mathbf{Q}_b = \mathbf{Y} \odot \mathbf{Y}, \quad (15)$$

with $\mathbf{Y} = \mathbf{V}\mathbf{D}_\lambda$. We use subindex b to emphasize this is the decomposition of the total sum of squares of the scalar product matrix $\mathbf{B}$. Matrix $\mathbf{Q}_b$ satisfies

$$\mathbf{1}'\mathbf{Q}_b = \boldsymbol{\lambda}^{(2)} = (\lambda_1^2, \lambda_2^2, \dots, \lambda_{n-1}^2, \lambda_n^2).$$

Contributions of each row to the total sum-of-squares of $\mathbf{B}$ can be obtained as

$$w_i = \sum_{j=1}^n y_{ij}^2. \quad (16)$$

The goodness-of-fit for a particular point (g_i) in the k -dimensional solution can then be calculated as

$$g_i^b = \frac{\sum_{j=1}^k y_{ij}^2}{\sum_{j=1}^n y_{ij}^2} = \frac{\sum_{j=1}^k \lambda_j^2 v_{ij}^2}{\sum_{j=1}^n \lambda_j^2 v_{ij}^2}. \quad (17)$$

As before, a weighted average of these point-wise measures gives the overall goodness-of-fit. It is harder to develop a useful measure for the goodness-of-fit of pairwise distances in the non-Euclidean case. For Euclidean distance matrices, the approximation to the observed distances is always from below, and therefore Equation (11) seems a sensible measure, with an upper bound of 1 if the distance of the corresponding pair is perfectly represented. In the non-Euclidean case, the fitted distances can exceed the originally observed distances (see the geographical example in the next section).

Error statistics

Instead of focusing on goodness-of-fit, one can also focus on error. If the map obtained by MDS is good approximation, then errors will be small, and concentrated around zero. Most points may be expected to make a small and similar contribution to the error sum-of-squares (ESS). A simple measure of poorness-of-fit, indicated by $\tilde{g}_{ij}$, is the contribution to the total ESS,

$$\tilde{g}_{ij} = \frac{e_{ij}^2}{\sum_{i>j} e_{ij}^2}, \quad (18)$$

where e_{ij} is an element of $\mathbf{E} = \mathbf{D} - \hat{\mathbf{D}}$. Large outliers on $\tilde{g}_{ij}$ correspond to poorly fitted pairs. These error contributions satisfy $0 \leq \tilde{g}_{ij} \leq 1$. Equation (18) can also be used to develop a pointwise statistic, by summing errors that pertain to a particular observation, that is

$$\tilde{g}_i = \frac{\sum_{j=1}^n e_{ij}^2}{2 \sum_{i>j} e_{ij}^2}, \quad (19)$$

and large outliers on $\tilde{g}_i$ would correspond to poorly fitted observations. By counting each error twice in the denominator, we achieve that $0 \leq \tilde{g}_i \leq 1$.

3. Examples

We illustrate the proposed goodness-of-fit statistics with three different data sets taken from demography, genetics and geography and discuss these in the following sections.

3.1. Demographic distances between countries

We discuss the analysis of an Euclidean distance matrix obtained from a demographic dataset of six variables (live birth rate, death rate, infant death rate, male and female life

expectancy and gross national product (GNP)) for 97 countries described by Rouncefield [21]. Euclidean distances between countries were calculated using the standardized data. GNP was log-transformed to linearize its relationship with the other variables prior to standardization. An MDS map of the Euclidean distance matrix of the countries is shown in Figure 1(A). The original variables have been mapped into the MDS plot by regression to aid interpretation [7]. This shows the first dimension is a wealth dimension separating rich countries with high life expectancies on the left from poor countries with high infant death rates and high birth rates on the right. In Figure 1(B), the countries are colour coded according to their goodness-of-fit, and this reveals some countries, mainly in the centre of the map, with, according to Equation (9) a low goodness-of-fit (Saudi Arabia 0.26; Libya 0.29; Oman 0.29). These points maybe better ignored (or brushed away) to avoid misinterpretation of the map. Closeness to the origin does not necessarily imply poor fit, as illustrated by Lebanon, which is close to the origin, but has good fit (0.94). In Figure 1(C), we use pairwise goodness-of-fit statistics to reveal relatively poorly displayed inter-country distances that have $g_{ij} < 0.50$; these countries are connected by lines. This reveals that many countries that appear as neighbours in the map are in reality farther away from each other than the map suggests, being their squared full space distance at least two times larger. In Figure 1(D), we focus on well-represented inter-country distances, showing all distances for Sierra Leone that have $g_{ij} > 0.90$. This shows almost all distances with respect to Sierra Leone are very well represented, and that the country acts as an *attractor point*. Similarly, Gambia, Malawi, Ethiopia, Somalia, Angola and Mexico are also well represented attractor points in terms of inter-country distances. These countries have, in terms of the original distance matrix, a large average distance with respect to the rest of the countries. In this analysis, the larger distances have in general, a better fit than the smaller distances, as is also revealed by a scatter plot of observed against fitted distances (see supplementary Figure S1).

3.2. Genetics distances between individuals

MDS is widely used in genetics for the detection of population substructure, which refers to the existence of groups of individuals in a genetic database that come from different human populations. Many examples of the use of MDS for this purpose can be found in the genetic literature [9,17,18,22,25]. MDS studies in genetics often use the allele sharing distance. The possible genotypes of bi-allelic genetic variables are, in generic notation, the homozygotes AA and BB and the heterozygote AB. If one of the alleles is counted, usually the minor allele, the genotype data can be coded into 0,1,2 format. The allele sharing distance is defined as

$$d_{ij} = \frac{1}{K} \sum_{k=1}^K (2 - x_{ijk}), \quad (20)$$

where x_{ijk} is the number of alleles shared by individuals i and j at genetic variable k , taking only values in the set (0,1,2). Equation (20), when applied to (0,1,2) data, is actually equivalent to the Manhattan distance, which is a well-known metric in the statistical literature [15]. This metric is often also referred to as the city-block distance or the taxicab metric. The Manhattan distance is directly proportional to the allele sharing distance. Geneticists often perform MDS of genetic data with the PLINK software [19]; this program

Figure 1. MDS of the poverty data set. (A) MDS map with added variables (LM: life expectancy of males, LF: life expectancy of females, GNP: gross national product, Birth: birth rate, Infant: infant death rate, Death: death rate). (B) MDS map colour-coded according to the goodness-of-fit of the countries. (C) MDS map showing inter-country distances with fit below 0.50. (D) MDS map showing inter-country distances with fit above 0.90 for Sierra Leone.

does classical metric MDS as described above on the Manhattan distances between the individuals of a genetic database. We present here an MDS of 109 initially presumed unrelated individuals of a sample of Chinese in Metropolitan Denver, CO, USA; the CHD sample of the 1000 Genomes project (www.internationalgenome.org). Genetic variants were filtered prior to MDS by selecting only autosomal variants without missing values, with a minor allele frequency above 0.40, non-significant in an exact test for Hardy–Weinberg equilibrium ($\alpha = 0.05$), and without strong correlations with flanking markers. These filters left 28,158 genetic variables in the analysis. Figure 2(A) shows the MDS map of the individuals. All eigenvalues obtained were non-negative. The overall goodness-of-fit for a two-dimensional display is low, only 3.0%, which is typical of genetic applications with large datasets. Figure 2(B) applies a goodness-of-fit filter of 0.25 at the individual level. This reveals that most individuals have a poor fit, and only the four outliers are well represented.

Figure 2. MDS of CHD data set. Individuals are colour coded according to their goodness-of-fit. (A) MDS of the Manhattan distances. (B) Filtered MDS map where individuals with goodness-of-fit below 0.25 are not shown. (C) MDS map where individuals with a pairwise goodness-of-fit below 0.25 are connected by blue lines. (D) MDS map where individuals with a pairwise goodness-of-fit above 0.50 are connected by blue lines.

The four outliers correspond to individuals for which a family relationship has been identified by Pemberton *et al.* [18]. In particular, the outlying pair in the first dimension has been estimated to be a full sib (FS) pair (identifiers NA17981 and NA17986), whereas the outlier in the second dimension has been estimated to be a parent–offspring (PO) pair (identifiers NA17976 and NA18166). Figure 2(C) connects individuals with a poorly fitted genetic distance (goodness-of-fit < 0.25), whereas Figure 2(D) connects individuals with relatively better represented genetic distances (goodness-of-fit > 0.50). This reveals that most genetic distances between unrelated individuals around the origin are poorly displayed, and that only the genetic distances of the related individuals are reasonably well fitted. Supplementary Figure S2 shows the scatterplot of observed against fitted Manhattan distances, and also confirms that all distances involving one or more individuals that form

a part of the single PO or FS family relationship are better fitted. Interestingly, the third dimension of the MDS map shows another pair of outliers, corresponding to a second degree relationship that was also uncovered by Pemberton *et al.* [18].

3.3. Geographic distances between Spanish cities

A classical example of metric MDS is the construction of a geographical map of a set of cities based on a table of inter-city distances. Such distances can be given in different form: as straight line geographical distances, road or railway distances, or as travel times. Many examples of this kind have been described in the literature [10,13,15]. We consider the road distances in kilometres between 47 cities in Spain. Figure 3(A) shows the result of a classical metric MDS of the data. The first principal axis is shown in the vertical dimension in order to better match the geographical map of Spain. There are 24 positive eigenvalues, the 25th eigenvalue is the structural zero and the remaining 22 are negative. Of all positive eigenvalues, three exceed the largest negative eigenvalue in absolute value. Using the standard adjustments, the goodness-of-fit of the distance matrix in two dimensions is 0.699 (using absolute values) or 0.815 (using positive eigenvalues only). The goodness-of-fit of the scalar products, obtained by using squared eigenvalues, is 0.979. Symbols in Figure 3(A) are colour coded to indicate goodness-of-fit of the observations according to Equation (17), which is based on squared eigenvalues. This suggests that the more outlying peripheral cities in Catalonia, Andalusia and the Basque country have a better fit than the central cities. The most central cities, Guadalajara and Madrid, have the poorest goodness-of-fit, 0.23 and 0.45 respectively. Figure 3(B) shows boxplots of the pointwise statistics of equations (17) (left) and (12) (right). Both statistics suggest about three relatively poorly fitted cities: Madrid, Cuenca and Guadalajara. Figure 3(C) shows these cities have a lower average distance with respect to the other Spanish cities, and that cities with larger average distances tend to have better fit. Figure 3(D) shows the pairs of cities that contribute more than 1% to the total error sum-of-squares, according to Equation (18); the distance between Guadalajara and Cuenca has the worst fit. A scatter plot of fitted against observed distances is shown in supplementary Figure S3 and shows graphically, despite the negative eigenvalues, an excellent fit with only a few poorly fitted intercity distances. The outliers in Figure S3 effectively correspond to the distance traced by lines in Figure 3(D).

4. Discussion

We have developed statistics that quantify the goodness-of-fit of observations and of pairs of observations in classical metric MDS. Nowadays, MDS is applied to increasingly large datasets, and the proposed statistics can be used to identify poorly represented (pairs of) points. Such points may be brushed away in order to emphasize the most salient aspects of the analysis and to avoid misinterpretations. We stress that the brushing of a point is not the same as its elimination from the analysis. A brushed point has been used in the analysis, but is simply not shown because it is poorly fitted. We do not suggest a re-analysis of the data without the poorly fitted points, as this would give an entirely new map, where again poorly fitted observations can be expected to be present. The practical applications in this paper show that poorly represented observations often cluster around the origin of the MDS map. However, goodness-of-fit filtering in MDS is not equivalent to brushing

Figure 3. Metric MDS of intercity distances in Spain. (A) MDS map with cities colour-coded according to the goodness-of-fit of the city. (B) Boxplots of pointwise goodness-of-fit measures according to Equations (17) and (12). (C) Goodness-of-fit g^b as a function of the average distance. (D) MDS map with poorly represented distances, as identified by Equation (18) with $\tilde{g}_{ij} > 0.01$, connected by blue lines.

away all points within a circle around the origin. Indeed, a point close to the origin can be well represented if its distance from the origin in the full space is also small.

In many applications of classical metric MDS, negative eigenvalues arise. Overall goodness-of-fit is then usually expressed by an ad-hoc adjustment, e.g. considering only positive eigenvalues or taking absolute values of the eigenvalues. These adjustments do not have a sound theoretical foundation and are used merely to avoid that the goodness-of-fit statistic in Equation (4) exceeds one. When the squared eigenvalues (14) are used, no ad-hoc adjustments are needed, as the index is always neatly in the 0–1 range. The use of squared eigenvalues implies we report the goodness-of-fit of **B** instead of the goodness-of-fit of the distance matrix. Because the distances are approximated indirectly, via scalar product matrix **B**, it is then understood that a better fit of **B** will generally imply a better fit of **D**, without quantifying exactly how well **D** is actually represented. A disadvantage

is that, beyond ℓ dimensions, the goodness-of-fit of $\mathbf{B}$ increases as more dimensions are included, while the goodness-of-fit of $\mathbf{D}$ is actually deteriorating. Overall goodness-of-fit statistics will of course look more favourable if squared eigenvalues are used.

Cailliez [2] proposed an adjustment for metric MDS with non-Euclidean distance matrices. By adding a constant to its off-diagonal entries, all eigenvalues can be rendered non-negative. If the adjustment is used, the point and pairwise goodness-of-fit measures for Euclidean distances in Section 2.2 can be used. However, negative eigenvalues are naturally expected if the distances are subject to error, and the use of an adjustment may introduce distortion in the real configuration of the observations [1].

An attractive property of classical metric MDS is that it provides a solution that is in the same scale as the original distance matrix. If the original distance matrix is in kilometres, the MDS map is also in kilometres, which facilitates interpretation of the map. This is obvious for the geographical data set analysed above, but also for the genetic data, as described next. For genetic data coded in (0,1,2) format, it is convenient to scale the Manhattan distance matrix by $1/k$ (or, accordingly, scale by $1/\sqrt{k}$ if Euclidean distances are used). This scaling will not affect the configuration of the points in the map, and neither its goodness-of-fit, but it will render the axes and distances of the map more interpretable. Two individuals that are a unit distance apart in the map now differ on average by one allele per locus. This scaling will typically bring all the coordinates in the MDS map within the $(-1, 1)$ interval, as the maximum difference in number of alleles between two individuals is two (see Figure 2). This interpretation is hampered by the fact that the map is a low-dimensional approximation to the original distance matrix, and therefore the property will not hold exactly, but only approximately so. For distance matrices that have the Euclidean property, and that therefore only have non-negative eigenvalues [15, Chapter 14], the map distances will approximate the true distances from below, and one can say that two individuals that are one unit apart, will differ on average by at least one minor allele per locus.

The ability of classical MDS to detect related individuals arises from its sensitivity to outliers: the related individuals have the smallest possible distance in the whole database (see Figure S2) and the plane fitted by MDS is tilted toward these individuals. Consequently, all distances involving these outliers are relatively better represented. At first sight, the MDS map may suggest the outliers to be different from the rest of the sample, potentially stemming from a different human population. This is not the case, and in fact the original distances of the outlying individuals with respect to the other individuals of the sample are not larger than the average original distance between just any two unrelated individuals. If, as in the genetic example, the MDS solution is dominated by a few outliers, and the latter are understood, then a re-analysis of the data without outliers can be applied in order to better understand the structure of the remaining observations.

5. Software and datasets

The function `cmdscale` of the statistical environment R [20], actually version 3.5.1, is widely used to perform classical MDS. This function does not allow the calculation of all statistics proposed in Section 2. We supply R code (function `PrinCoor`) included in R package `calibrate` [8] that implements all goodness-of-fit statistics discussed in this paper. All datasets used in this paper are accessible online. The poverty data set

is available at <http://jse.amstat.org/v3n2/datasets.rouncefield.html>, the genetic data set is available at <http://www.internationalgenome.org> and the geographical data set is available at <http://www-eio.upc.es/jan/data/SpainDist.dat>.

Acknowledgments

The anonymous reviewers are gratefully acknowledged for their comments on the paper.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was partially supported by grants RTI2018-095518-B-C22 of the Spanish Ministry of Science, Innovation and Universities and the European Regional Development Fund, and by grant R01 GM075091 from the United States National Institutes of Health.

ORCID

Jan Graffelman <http://orcid.org/0000-0003-3900-0780>

References

- [1] I. Borg and P.J.F Groenen, *Modern Multidimensional Scaling*, 2nd ed. Springer, New York, 2005.
- [2] F. Cailliez, *The analytical solution of the additive constant problem*, Psychometrika 48 (1983), pp. 305–308.
- [3] J. Carroll and J. Chang, *Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition*, Psychometrika 35 (1970), pp. 283–319.
- [4] T.F. Cox and M.A.A Cox, *Multidimensional Scaling*, 2nd ed. Chapman & Hall, Boca Raton, 2001.
- [5] J. De Leeuw and P. Mair, *Multidimensional scaling using majorization: SMACOF in R*, J. Stat. Softw. 31 (2009), pp. 1–30.
- [6] J.C. Gower, *Some distance properties of latent root and vector methods used in multivariate analysis*, Biometrika 53 (1966), pp. 325–338.
- [7] J. Graffelman and T. Aluja-Banet, *Optimal representation of supplementary variables in biplots from principal component analysis and correspondence analysis*, Biometric. J. 45 (2003), pp. 491–509.
- [8] J. Graffelman and F.A. van Eeuwijk, *Calibration of multivariate scatter plots for exploratory analysis of relations within and between sets of variables in genomic research*, Biometric. J. 47 (2005), pp. 863–879.
- [9] M. Jakobsson, S. Scholz, P. Scheet, J. Gibbs, J. VanLiere, H. Fung, Z. Szpiech, J. Degnan, K. Wang, R. Guerreiro, J. Bras, J. Schymick, D. Hernandez, B. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. Cann, J. Hardy, N. Rosenberg and A. Singleton, *Genotype, haplotype and copy-number variation in worldwide human populations*, Nature 451 (2008), pp. 998–1003.
- [10] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, 5th ed. New Jersey, Prentice Hall, 2002.
- [11] J. Kruskal, *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis*, Psychometrika 29 (1964), pp. 1–27.
- [12] P. Mair, I. Borg and T. Rusch, *Goodness-of-fit assessment in multidimensional scaling and unfolding*, Multivariate Behav. Res. 51 (2016), pp. 772–789.
- [13] B.F.J Manly, *Multivariate Statistical Methods: a Primer*, Chapman and Hall, London, 1989.

- [14] K.V. Mardia, *Some properties of classical multi-dimensional scaling*, Comm. Statist. Theory Methods 7 (1978), pp. 1233–1241.
- [15] K.V. Mardia, J.T. Kent and J.M. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.
- [16] E. Paradis, *Multidimensional scaling with very large datasets*, J. Comput. Graph. Stat. 27 (2018), pp. 935–939.
- [17] T. Pemberton, M. DeGiorgio and N. Rosenberg, *Population structure in a comprehensive genomic data set on human microsatellite variation*, G3: Genes Genomes Genetics 3 (2013), pp. 891–907.
- [18] T. Pemberton, C. Wang, J. Li and N. Rosenberg, *Inference of unexpected genetic relatedness among individuals in HapMap Phase III*, Am. J. Hum. Genet. 87 (2010), pp. 457–464.
- [19] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. de Bakker, M. Daly and P. Sham, *Plink: A toolset for whole-genome association and population-based linkage analysis*, Am. J. Hum. Genet. 81 (2007), pp. 559–575.
- [20] R Development Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2004. ISBN 3-900051-00-3
- [21] M Rouncefield, *The statistics of poverty and inequality*, J. Stat. Educ. 3 (1995). Available at <http://jse.amstat.org/v3n2/datasets.rouncefield.html>.
- [22] C. Sabatti, S. Service, A. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. Jones, N. Zaitlen, T. Varilo, M. Kaakinen, U. Sovio, A. Ruokonen, J. Laitinen, E. Jakkula, L. Coin, C. Hoggart, A. Collins, H. Turunen, S. Gabriel, P. Elliot, M. McCarthy, M. Daly, M. Järvelin, N. Freimer and L. Peltonen, *Genome-wide association analysis of metabolic traits in a birth cohort from a founder population*, Nat. Genet. 41 (2009), pp. 35–46.
- [23] W Torgerson, *Theory and Methods of Scaling*, Wiley, New York, 1958.
- [24] J. Tzeng, H. Lu and W. Li, *Multidimensional scaling for large genomic data sets*, BMC. Bioinformatics. 9 (2008), 179.
- [25] C. Wang, Z. Szpiech, J. Degnan, M. Jakobsson, T. Pemberton, J. Hardy, A. Singleton and N. Rosenberg, 2010. *Comparing spatial maps of human population-genetic variation using procrustes analysis*. Stat. Appl. Genet. Mol. Biol. 9. Article 13