The Visual Computer
https://doi.org/10.1007/s00371-020-01793-w

ORIGINAL ARTICLE

# Feature selection based on star coordinates plots associated with eigenvalue problems

**Alberto Sanchez Campos[1,2] · Laura Raya[3] · Miguel A. Mohedano-Munoz[1] · Manuel Rubio-Sánchez[1]**



**Abstract**

Feature selection consists of choosing a smaller number of variables to work with when analyzing high-dimensional data sets. Recently, several visualization tools, techniques, and feature relevance measures have been developed in order to help users carry out the feature selection. Some of these approaches are based on radial axes methods, where analysts perform backward feature elimination by discarding features that have a low impact on the visualizations. Similarly, in this paper, we propose a new feature relevance measure for star coordinates plots associated with the class of linear dimensionality reduction mappings defined through the solutions of eigenvalue problems, such as linear discriminant analysis or principal component analysis. We show that the approach leads to enhanced feature subsets for class separation or variance maximization in the plots for numerous data sets of the UCI repository. Lastly, in practice, the tool allows analysts to decide which features to discard by examining their relevance and by taking into account previous domain knowledge.

**Keywords** Feature selection · Eigenvalue problems · Linear projections · Multidimensional visualization · Star coordinates · Principal component analysis · Linear discriminant analysis

## 1 Introduction

Data preprocessing is an important operation in the fields of statistics, data mining, or machine learning. Nowadays, many data sets contain hundreds or thousands of features, many of which can be redundant or irrelevant [25]. Thus, an initial data set is typically simplified in order to work with an alternative one that contains a smaller number of features. There are two main approaches for reducing the dimensionality of the data: feature transformation [29] and feature selection [28]. Feature transformation (in the context of dimensionality reduction) consists of mapping the original data features to a new space of lower dimensionality. In contrast, feature selection is concerned with choosing a subset of original features to work with. This preprocessing step can be beneficial since using the resulting smaller subset of features can reduce overfitting, enhance performance, shorten computational runtimes, or lead to simpler and more interpretable models [38].

Finding an optimal subset of features generally requires examining an exponential number of subsets. Thus, most feature selection approaches rely on efficient greedy algorithms [13,28] that select or discard features progressively. In this paper, we focus on combining automatic procedures with interactive visualization approaches, where analysts can make decisions regarding which features to discard by considering both the output of an automatic method and their previous domain knowledge. Specifically, we study feature selection aided by star coordinates (SC) [22,23], which is a multivariate visualization method based on radial axes [10,11,37]. SC not only generates linear projections of the data onto a two-dimensional plane, but also displays a set of axis vectors associated with the features. This provides additional information about the features' relation to the data samples and to themselves. In practice, users can select and

✉ Alberto Sanchez Campos
alberto.sanchez@urjc.es

Laura Raya
laura.raya@u-tad.com

Miguel A. Mohedano-Munoz
miguel.munoz@urjc.es

Manuel Rubio-Sánchez
manuel.rubio@urjc.es

1 Universidad Rey Juan Carlos, Madrid, Spain

2 Research Center for Computational Simulation, Madrid, Spain

3 U-tad, Madrid, Spain

place the axis vectors arbitrarily in the plot in order to generate any linear mapping. In this work, we focus on a different alternative that consists of computing the axis vectors through automatic procedures related to linear dimensionality reduction algorithms [35], such as principal component analysis (PCA) [21] or linear discriminant analysis (LDA) [31].

Recently, several works have proposed feature reduction procedures that take into account the length of the axis vectors to determine the importance of a data variable in SC plots [35, 43]. In addition, the work in [38] measures the influence of a variable in a visualization by computing a measure of the displacement of the plotted points when a feature (i.e., an axis vector) is eliminated from the data set. In short, it measures how much a plot would change when discarding features.

While the previous approaches are valid for arbitrary SC plots, we present a feature relevance measure for enhancing the feature elimination process for several commonly used SC visualizations. Specifically, we propose a strategy for determining the influence of features in SC plots associated with linear dimensionality reduction transformations that are the result of solving eigenvalue problems. The approach not only takes into account the magnitude and the orientation of the axis vectors, but it also considers the eigenvalues associated with the eigenvectors that solve the problem and constitute the linear mapping. The results show that the proposed measure outperforms related approaches based on SC plots described in the literature.

Lastly, we describe a simple graphical interface that ranks the features according to their relevance and allows users to visualize the SC plots and to discard variables interactively. Our proposal offers the analyst a better estimate of the importance of each feature in the linear mapping. This allows domain experts to acquire insight into the data and guides them toward obtaining a reliable set of features.

The rest of the paper is organized as follows. Section 2 describes the most relevant methods related to our proposal, while Sect. 3 includes basic background. In Sect. 4, we describe our measure for determining the importance of a feature in a SC plot for eigenvalue problems, while Sect. 5 presents the results. Finally, Sect. 6 presents the conclusions.

## 2 Related work

The previous work on feature selection has mainly focused on automatic techniques [4,5,13]. However, recently, the data visualization community has developed methods that involve interactive visualizations and graphical interfaces, in order to integrate users and their expertise into the data analysis process. There are different ways to categorize visual methods for feature selection. Firstly, their goal can be to choose smaller sets of variables for: classification [26,32], clustering [2,12,18,20,40,41,47], outlier detection [20], gaining insight regarding features or relations among them [8,18, 20,26,30], or simply to rank variables [30,39,45]. In addition, some methods rank the features in order to carry out the attribute selection [18,20,26,32,39], but others are aimed at searching for subsets of variables without relying exclusively on a particular ranking [2,6,12,41,45,47]. A more complete state-of-the-art review of these techniques can be found in [38].

Many of these techniques rely on different quality metrics and heuristics (including estimations of feature similarity, goodness of a clustering, uniformity, interestingness, number of outliers, entropy, and many others) and are, therefore, very diverse (see [3] for an overview of some of these approaches). Some of them can also be regarded as feature ranking methods, since they sort the data attributes according to some measure, and either select or discard them progressively.

In this paper, we present an approach that falls within this category of feature selection methods. Thus, we focus here on feature ranking methods that use measures of feature relevance, which are not just based on quality metrics on the visualizations [19,40]. Yang et al. present interactive hierarchical displays [46,47] to visualize large multivariate data sets. Users can group similar features to display data with a lower set of dimensions in parallel coordinates, star glyphs, scatterplot matrices, and dimensional stacking. Another proposal for ranking features [39] relies on different heuristics, such as uniformity or number of outliers. Specifically, it is based on ordering histograms and scatter plots. The work in [20] proposes a tool based on parallel coordinates where the order and number of axes can be interactively manipulated according to a ranking algorithm. The method combines user-defined and weighted quality metrics like measures of correlation, or outlier and cluster detection. The method proposed in [1] sorts features in RadViz by comparing the results of a cluster density metric on visualizations obtained by adding a single new feature to an existing plot. INFUSE [26] helps users to understand how features are ranked. The tool displays a circular glyph for each feature, showing information related to various measures commonly used for feature selection such as the Fisher score or the information gain.

Finally, it is important to note that the feature ranking measures used in the literature are usually specific for certain data analysis tasks (classification, clustering, etc.). In contrast, the measure that we propose in the paper is general in the sense that it applies to SC visualizations, which can be used for a wide variety of analysis tasks.

## 3 Background

Dimensionality reduction mappings can be categorized as linear or nonlinear. In general, although nonlinear mappings may be able to represent data more faithfully, it is usually dif-





Journal: 371 MS: 1793 TYPESET DISK LE CP Disp.:2020/1/10 Pages: 14 Layout: Large

ficult to understand the influence of the original features in the mapping. Thus, in this paper, we employ SC plots, which generate linear projections of the data and also show information about the features, which allows users to understand how they affect the linear mappings.

In particular, SC is an exploratory data analysis technique that has been used to inspect correlations, cluster structure, class separation, or searching for outliers or data with desired characteristics. Specifically, it is a projection method that maps high $n$-dimensional data points (i.e., individual samples) linearly onto a plane. In particular, the linear mapping is defined through a set of $n$ 2-dimensional axis vectors $\mathbf{v_i}$, for $i = 1, \ldots, n$, where $\mathbf{v}_i$ is associated with the $i$th data variable. The representation $\mathbf{p} \in \mathbb{R}^2$ of a data point $\mathbf{x} \in \mathbb{R}^n$ is a linear combination of the vectors $\mathbf{v}_i$. Formally:

$$\mathbf{p} = x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \cdots + x_n\mathbf{v}_n = \mathbf{V}^{\mathrm{T}}\mathbf{x}, \tag{1}$$

where $\mathbf{V}$ is the $n \times 2$ matrix whose rows are the vectors $\mathbf{v}_i$, and $x_i$, for $i = 1, \ldots, n$, are the attribute values of $\mathbf{x}$. It is important to note that the linear mapping is completely specified by the matrix $\mathbf{V}$. Lastly, the mapping of an entire data set of cardinality $N$ can also be expressed in matrix form as:

$$\mathbf{P} = \mathbf{X}\mathbf{V}, \tag{2}$$

where $\mathbf{X}$ is the $N \times n$ data matrix and $\mathbf{P}$ is the corresponding $N \times 2$ matrix of projected points.

There are two ways to choose the axis vectors (i.e., the matrix $\mathbf{V}$) when working with SC. On the one hand, they can be specified manually and interactively by analysts, for instance, through some graphical user interface. In this regard, it would be possible to generate any linear mapping of the data onto a plane, since users can choose arbitrary axis vectors that define matrix $\mathbf{V}$. On the other hand, we can also obtain a $2 \times n$ transformation matrix $\mathbf{A}$ that maps $n$-dimensional data points onto a plane through some automatic procedure (e.g., PCA). In that case, we can build a SC plot that produces the same mapping by setting $\mathbf{V} = \mathbf{A}^{\mathrm{T}}$, where the axis vectors would simply be the columns of $\mathbf{A}$, due to (1). Thus, given any linear projection, possibly obtained through some sophisticated computational procedure, we can always build an analogous SC plot. The resulting visualization will not only show the projected points, but will also depict information regarding the $n$ original features in the form of axis vectors.

There are numerous linear techniques that can be useful for data analysis, data mining, and machine learning tasks, such as projection pursuit [16] or independent component analysis (ICA) [17]. In this paper, our focus will be on linear mappings that are the result of solving eigenvalue problems. Here, we detail the methods used to better understand their objective functions, which we maximize.

One of the most common methods is PCA, which can be interpreted in several ways from an optimization point of view (see [33]). PCA is appealing for data analysis since the projected points will represent the best rank-2 approximation of the high-dimensional data. In particular, PCA finds the orthogonal $n \times 2$ matrix $\mathbf{V}$ that solves the following optimization problem:

$$\begin{array}{ll} \underset{\mathbf{V} \in \mathbb{R}^{n\times 2}}{\text{maximize}} & \mathrm{Tr}\left[\frac{1}{N-1}\mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}\right] \\ \text{subject to} & \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I} \end{array} \tag{3}$$

where Tr denotes trace, $\mathbf{I}$ is the $(2 \times 2)$ identity matrix, and $\mathbf{X}$ is the $N \times n$ data matrix that has been previously centered (i.e., the mean of the original data has been subtracted from each data point). The solution to (3) is the matrix whose columns are the two eigenvectors associated with the two largest eigenvalues $\lambda_1$ and $\lambda_2$ of the sample covariance matrix of the data $\mathbf{X}^{\mathrm{T}}\mathbf{X}/(N-1)$ (see [24]). These eigenvalues represent the maximum variances of the data along the orthogonal directions specified by the eigenvectors in the data space. They also represent the variances along the canonical axes of the SC plot. Lastly, the optimum value of the objective function in (3) will be the sum of the eigenvalues: $\lambda_1 + \lambda_2$.

Another popular linear approach that can be used when the data is categorized (i.e., labeled) into $C$ different classes is LDA. The technique projects the data onto a subspace of lower dimensionality in an effort to achieve good class separability. Specifically, LDA tries to maximize a ratio of a measure of the between-class scatter over a measure of the within-class scatter. For visualization purposes on a plane (which requires $C > 2$), LDA finds an orthogonal projection matrix $\mathbf{V}$ that solves the following optimization problem:

$$\begin{array}{ll} \underset{\mathbf{V} \in \mathbb{R}^{n\times 2}}{\text{maximize}} & \mathrm{Tr}\left[\frac{\mathbf{V}^{\mathrm{T}}\mathbf{S}_b\mathbf{V}}{\mathbf{V}^{\mathrm{T}}\mathbf{S}_w\mathbf{V}}\right] \\ \text{subject to} & \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I} \end{array} \tag{4}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are between-class and within-class scatter matrices, respectively. If $\mathbf{S}_w$ is nonsingular, then the columns of the matrix $\mathbf{V}$ that optimizes (4) will be the two eigenvectors associated with the two largest eigenvalues $\lambda_1$ and $\lambda_2$ of $\mathbf{S}_w^{-1}\mathbf{S}_b$ (see [24]). For LDA, the eigenvalues indicate a measure of the class separability along the directions specified by their corresponding eigenvectors. Thus, the classes will tend to be more separated along the direction of the first eigenvector. Lastly, as in PCA, the optimum value of the objective function (4) is $\lambda_1 + \lambda_2$.

Finally, if analysts are interested in obtaining a reduced set of $m$ features that approximate the data as well as possible in PCA, or that better separate the classes in LDA, they





Journal: 371 MS: 1793 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2020/1/10 Pages: 14 Layout: Large

can progressively discard the variables that contribute less to forming these plots. In other words, they can discard the features that reduce $\lambda_1 + \lambda_2$ the least when they are eliminated. Naturally, this greedy approach does not guarantee finding the optimal subset of exactly $m$ features (note that finding an optimal subset of features is usually NP-hard [7]).

## 4 Weighted displacement feature relevance measure

The interpretation of how SC maps high-dimensional data onto a plane is fairly straightforward. Firstly, the orientation of an axis vector indicates in which direction a plotted point would move when increasing the value of the associated feature. In addition, the relative magnitude of an axis vector, in comparison with the rest, provides intuition regarding the amount of contribution of a particular variable in the resulting visualization, given that all variables are scaled similarly. Note that in SC the features should share a similar scaling, since otherwise the ones with larger ranges would have a greater impact on the resulting plots. In this paper, we work with standardized data (i.e., the features have zero mean and unit variance). Other possibilities include transforming each feature to lie in the [0,1] interval, or centering and normalizing them to have unit range [36].

The possibility to visualize the feature axis vectors in SC, and to determine their relative contributions to a plot, allows us to perform a visual feature selection. For instance, we can progressively discard the most irrelevant variables, while also maintaining others according to domain knowledge. Recently, several works in the literature have proposed measures for establishing this contribution or importance of a variable in a SC plot, and therefore, on the analysis task for which the visualization is intended. In [35,43], the feature selection process is guided exclusively by the length of the axis vectors, where the shortest ones constitute the candidates to be discarded. Sanchez et al. [38] propose the average displacement of the low-dimensional points when a feature is discarded as a measure to determine the influence of that variable in the plot. Specifically, this measure is defined as:

$$f(\mathbf{v}_i) = \frac{1}{N}\sum_{j=1}^{N} \|\mathbf{p}^{(j)} - \mathbf{q}_{\mathbf{v}_i}^{(j)}\|\,, \tag{5}$$

where $\mathbf{p}^{(j)}$ is the projection of the $j$th sample and $\mathbf{q}_{\mathbf{v}_i}^{(j)}$ is the corresponding low-dimensional point when removing the feature associated with the axis vector $\mathbf{v}_i$. Note that it is also possible to use the median point displacement, which is more robust. However, in the remainder of the paper, we will use the definition in (5), since it is the one described in [38].

Instead, we propose a new measure to guide the process of visual feature selection. The following result shows how on any SC plot the average point displacement when a feature is discarded depends not only on the axis vector length, but also on the mean of the absolute values of the associated feature components of all of the data samples.

**Proposition 1** *In SC, the average displacement of the low-dimensional points when a feature is discarded is $f(\mathbf{v}_i) = \alpha_i\|\mathbf{v}_i\|$, where $\mathbf{v}_i$ is the SC axis associated with the feature and $\alpha_i$ is the mean of the absolute values of the $i$th component of all the data samples.*

***Proof*** Let $\mathbf{x}^{(j)} = \left(x_1^{(j)}, \ldots, x_n^{(j)}\right)$, for $j = 1, \ldots, N$, be the samples in our data set, then:

$$\mathbf{q}_{\mathbf{v}_i}^{(j)} = \sum_{k=1,\, k\neq i}^{n} x_k^{(j)}\mathbf{v}_k = \mathbf{p}^{(j)} - x_i^{(j)}\mathbf{v}_i\,, \tag{6}$$

is the low-dimensional location of $\mathbf{x}^{(j)}$ when discarding the $i$th feature from the SC model. In that case, the average point displacement can be expressed as:

$$\begin{aligned} f(\mathbf{v}_i) &= \frac{1}{N}\sum_{j=1}^{N}\|\mathbf{p}^{(j)} - \mathbf{q}_{\mathbf{v}_i}^{(j)}\| \\ &= \frac{1}{N}\sum_{j=1}^{N}\|\mathbf{p}^{(j)} - \left(\mathbf{p}^{(j)} - x_i^{(j)}\mathbf{v}_i\right)\| \\ &= \frac{1}{N}\sum_{j=1}^{N}\|x_i^{(j)}\mathbf{v}_i\| = \frac{1}{N}\sum_{j=1}^{N}|x_i^{(j)}|\,\|\mathbf{v}_i\| \\ &= \|\mathbf{v}_i\|\frac{1}{N}\sum_{j=1}^{N}|x_i^{(j)}| = \alpha_i\|\mathbf{v}_i\|. \end{aligned} \tag{7}$$

□

Note that even if each feature is standardized to have mean 0 (and standard deviation 1), $\alpha_i$ is generally different for each feature as it is the mean of the absolute values of the $i$th component.

When the SC projection is given by a linear dimensionality reduction algorithm like LDA or PCA (based on eigenvalue problems), the horizontal and vertical axes of the SC plot represent the optimal directions (defined by the eigenvectors) associated with the optimization problem. Therefore, $\lambda_1$ and $\lambda_2$ represent the variance (for PCA) or a measure of class separability (for LDA) in the $X$ and $Y$ axes of the SC plot, respectively. Our approach is based on the key insight that if $\lambda_1 > \lambda_2$ then a larger point displacement on the $X$ axis (after removing a feature) is likely to have a stronger impact on the problem's objective function. Therefore, our proposed novel measure will take into account the relative importance of each of the canonical axes when determining the influence of each original feature.

To compute this measure, we first break down the average displacement when a feature is discarded into horizontal and vertical components. These displacements will depend not only on the length of the axis vector to discard, but also on its direction, and on the associated feature's values (i.e., its probability distribution). The following proposition provides simplified expressions of the displacements.

**Proposition 2** *In SC, the average horizontal and vertical displacements of the low-dimensional points when the ith feature is discarded are* $f_1(\mathbf{v}_i) = f(\mathbf{v}_i)\,|\cos(\theta_i)|$ *and* $f_2(\mathbf{v}_i) = f(\mathbf{v}_i)|\sin(\theta_i)|$*, respectively, where* $\theta_i$ *is the angle between* $\mathbf{v}_i$ *and the (1,0) vector (i.e., the positive horizontal axis).*

***Proof*** Let $\mathbf{x}^{(j)} = \left(x_1^{(j)}, \ldots, x_n^{(j)}\right)$, for $j = 1, \ldots, N$, be the samples in our data set. According to (7), the average horizontal and vertical displacements of the low-dimensional points when a feature is discarded can be computed as:

$$\begin{aligned} f_1(\mathbf{v}_i) &= \frac{1}{N}\sum_{j=1}^{N} |p_1^{(j)} - q_{\mathbf{v}_i,1}^{(j)}| = \frac{1}{N}\sum_{j=1}^{N} |x_i^{(j)} v_{i,1}| \\ &= |v_{i,1}| \frac{1}{N}\sum_{j=1}^{N} |x_i^{(j)}| = \|\mathbf{v}_i\|\,|\cos(\theta_i)| \frac{1}{N}\sum_{j=1}^{N} |x_i^{(j)}| \\ &= \|\mathbf{v}_i\|\,|\cos(\theta_i)|\,\alpha_i = f(\mathbf{v}_i)\,|\cos(\theta_i)|. \end{aligned} \tag{8}$$

Similarly,

$$f_2(\mathbf{v}_i) = \frac{1}{N}\sum_{j=1}^{N} |p_2^{(j)} - q_{\mathbf{v}_i,2}^{(j)}| = f(\mathbf{v}_i)|\sin(\theta_i)|, \tag{9}$$

where $p_k^{(j)}$, $q_{\mathbf{v}_i,k}^{(j)}$, and $v_{i,k}$ are the $k$th components of $\mathbf{p}^{(j)}$, $\mathbf{q}_{\mathbf{v}_i}^{(j)}$, and $\mathbf{v}_i$, respectively. □

Having decomposed the total displacement into horizontal and vertical components, we propose using a weighted sum of each displacement. Specifically, the weights correspond to the eigenvalues associated with the plot's axes, which encode the importance of these canonical directions. Formally:

$$\begin{aligned} g(\mathbf{v_i}) &= \lambda_1 f_1(\mathbf{v}_i) + \lambda_2 f_2(\mathbf{v}_i) \\ &= f(\mathbf{v}_i)\,(\lambda_1|\cos(\theta_i)| + \lambda_2|\sin(\theta_i)|) \\ &= \|\mathbf{v}_i\|\,\alpha_i\,(\lambda_1|\cos(\theta_i)| + \lambda_2|\sin(\theta_i)|)\,. \end{aligned} \tag{10}$$

Since $\lambda_1 \geq \lambda_2$, the horizontal displacement will usually have more relevance than the vertical one for the algorithm's objective. Thus, although the length of an axis vector plays a role in determining the importance of a feature (and therefore has been used in [35,36,44]), its orientation should be considered as well. For example, if $\lambda_1$ is substantially greater than $\lambda_2$ then a feature with a long axis vector that is nearly perpendicular to the horizontal axis may not be particularly relevant for the algorithm's objective (e.g., to separate classes when using LDA). Note that in that case $\cos(\theta_i) \approx 0$ while $|\sin(\theta_i)| \approx 1$, and therefore $g(\mathbf{v_i}) \approx \|\mathbf{v}_i\|\alpha_i\lambda_2$. Similarly, if the feature's axis vector $\mathbf{v}_i$ is nearly horizontal (i.e., perpendicular to the $Y$ axis) then $g(\mathbf{v_i}) \approx \|\mathbf{v}_i\|\alpha_i\lambda_1$.

Figure 1a shows the LDA projection of a four feature subsets (I0, DA, DR, P) from the breast tissue data set of the UCI repository [9], which contains 106 samples categorized into six classes. We automatically scale the figure to make the data (colored dots according to their class label) and the axis vectors (red line segments) occupy all of the available space. Our tool also includes a unit circle that can be useful in other SC plots (e.g., the length of the axis vectors in orthographic star coordinates [27] must be at most 1). The eigenvalues are $\lambda_1 = 13.46$ and $\lambda_2 = 0.90$, which represent 92% and 6%, respectively, of the sum of the four eigenvalues. This means that the horizontal axis is an order of magnitude more important than the $Y$ axis for separating the classes, according to the objective function of LDA. This is also noticeable in Fig. 1a, where if the points are projected onto the horizontal axis it is still possible to separate the "adi" (blue) and "con" (green) classes from the rest. Instead, it would be difficult to separate any of the classes where the points had been projected onto the vertical axis.

In this example, not only the lengths of the axis vectors related to features I0 and P are very similar, but the total average displacement after eliminating each one is also similar. Nevertheless, P has a smaller impact on the visualization since $\lambda_1$ is considerably greater than $\lambda_2$, and because its associated axis vector is almost perpendicular to the horizontal axis. Concretely, our proposed measure $g(\mathbf{v})$ for I0 is 4 times that P, as is shown in Table 1. We can observe this graphically in Fig. 1b, c. In (b), we have ignored P and created a new LDA plot. In this case, the classes are separated nearly as well as in (a). However, in (c), the classes are not as well separated when removing feature I0: the "con" class (green), which was fairly well separated in (a) and (b), now overlaps with several other classes. Lastly, for this data set, our metric recommends removing DR, which is associated with the smallest value of $g(\mathbf{v})$ (see Table 1).

Finally, in practice, it is typical to discard various features at the same time. If that is the case, the total relevance of a subset of features can also be characterized by the displacement of the low-dimensional points when discarding that subset of features. A naive approach for selecting $k$ variables to remove consists of computing the displacements when discarding the entire subsets of features. However, this would be time-consuming since it would require computing $n$ choose $k$ linear mappings, where in this case $n$ is the number of variables that remain (i.e., that have not yet been discarded)





Journal: 371 MS: 1793 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2020/1/10 Pages: 14 Layout: Large

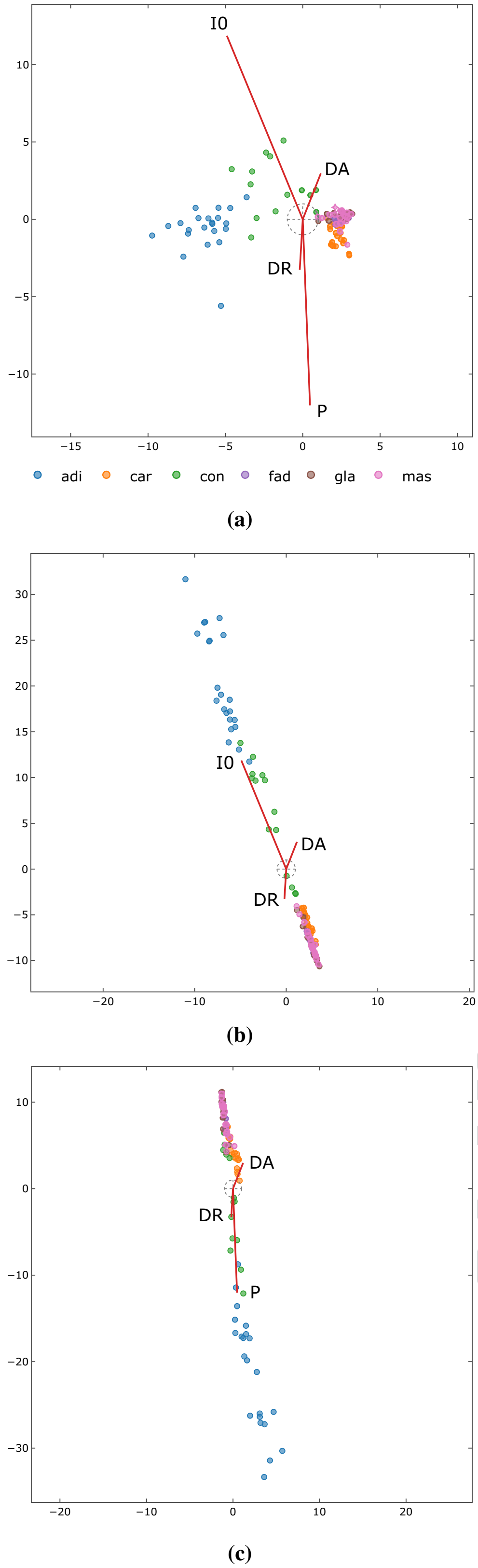



**Fig. 1** SC plots related to LDA of subsets of the breast tissue data set. In **a**, the plot uses features I0, DA, DR, P. In **b**, the feature P is discarded while in **c** I0 is eliminated. The classes are separated better in **b** than in **c** as suggested by our proposed metric, since $g(\mathbf{v})$ is smaller for P

**Table 1** Feature relevance measures for the SC plot in Fig. 1a, which is an LDA projection of a subset of four features (I0, DA, DR, P) of the breast tissue data set

| Feature ($\mathbf{v}$) | $\|\mathbf{v}\|$ | $f(\mathbf{v})$ | $g(\mathbf{v})$ |
|---|---|---|---|
| I0 | 12.84 | 11.12 | 66.52 |
| P | 12.03 | 10.10 | 14.73 |
| DA | 3.19 | 2.42 | 14.00 |
| DR | 3.27 | 2.45 | 4.23 |

at a certain stage of the feature selection process. Instead, a faster approach that only requires computing $n$ new plots consists of using sums of our proposed measure when applied to individual features. The theoretical foundation for this faster strategy relies on the fact that the weighted displacement measure $g$ applied to some set of features is bounded above by the sum of $g$ applied on the individual features of the set, as we show in the following result.

**Proposition 3** *Let $S = \{\mathbf{v}_{i_1}, \ldots, \mathbf{v}_{i_k}\}$ represent a set of $k$ axis vectors in a SC plot, where $I = \{i_1,\ldots,i_k\}$ simply contains the feature indices. When the features related to $S$ are discarded simultaneously, the measure $g(S)$ is bounded above by the sum of the $g(\mathbf{v})$ measures for each feature, i.e., $g(S) \le g(\mathbf{v}_{i_1}) + \cdots + g(\mathbf{v}_{i_k})$.*

***Proof*** Firstly, let:

$$\mathbf{q}_S^{(j)} = \sum_{k=1,\, k\notin I}^{n} x_k^{(j)} \mathbf{v}_k = \mathbf{p}^{(j)} - \sum_{i_k \in I} x_{i_k}^{(j)} \mathbf{v}_{i_k} \tag{11}$$

denote the low-dimensional point when discarding the features included in $S$. In that case, the average horizontal displacement of the low-dimensional points can be expressed as:

$$\begin{aligned} f_1(S) &= \frac{1}{N} \sum_{j=1}^{N} |p_1^{(j)} - q_{S,1}^{(j)}| \\ &= \frac{1}{N} \sum_{j=1}^{N} |x_{i_1}^{(j)} v_{i_1,1} + \cdots + x_{i_k}^{(j)} v_{i_k,1}| \\ &\le \frac{1}{N} \sum_{j=1}^{N} |x_{i_1}^{(j)} v_{i_1,1}| + \cdots + |x_{i_k}^{(j)} v_{i_k,1}| \\ &= f_1(\mathbf{v}_{i_1}) + \cdots + f_1(\mathbf{v}_{i_k}), \end{aligned} \tag{12}$$

where $q_{S,1}^{(j)}$ is the horizontal component of $q_S^{(j)}$. Analogously, the vertical displacement is bounded as follows:

$$f_2(S) \le f_2(\mathbf{v}_{i_1}) + \cdots + f_2(\mathbf{v}_{i_k}). \tag{13}$$





Journal: 371 MS: 1793 TYPESET DISK LE CP Disp.:2020/1/10 Pages: 14 Layout: Large

Finally, the total relevance of the group of features associated with $S$ is bounded by the sum of the relevance of each feature:

$$
\begin{aligned}
g(S) &= \lambda_1 f_1(S) + \lambda_2 f_2(S) \\
&\leq \lambda_1 f_1(\mathbf{v}_{i_1}) + \cdots + \lambda_1 f_1(\mathbf{v}_{i_k}) \\
&\quad + \lambda_2 f_2(\mathbf{v}_{i_1}) + \cdots + \lambda_2 f_2(\mathbf{v}_{i_k}) \\
&= g(\mathbf{v}_{i_1}) + \cdots + g(\mathbf{v}_{i_k}).
\end{aligned} \tag{14}
$$

□

Therefore, although it is possible to find a set of $k$ features to discard that minimizes $g$, an approximate but more efficient strategy consists of minimizing upper bounds on $g$.

## 5 Results

We have developed a tool using plotly, dash, scikit-learn, and pandas that shows SC plots and enables users to observe the influence of features in the SC projection by using an additional bar chart. The tool includes point-and-click and selection mechanisms to interact with the bar chart, which allow analysts to make decisions easily regarding which features to remove.

The bar chart shows, for every feature at a particular stage of the feature selection process, the value of the proposed measure $g(\mathbf{v})$. In addition, for the purpose of this paper, the bar chart can include the average displacement measure $f(\mathbf{v})$ (see 5), and the length of the axis vectors $\|\mathbf{v}\|$, which allow us to compare the different feature relevance measures. The tool allows us to sort the features according to one of the three measures. Furthermore, in order to compare the metrics effectively, we normalize each one by dividing it by its maximum value. Thus, the values of the metrics will be between 0 and 1, where 1 represents the greatest contribution to the SC plot for a particular metric. Lastly, each time the analyst removes features, the linear dimensionality reduction algorithm is applied again to the remaining features and the measures are recalculated. Figure 2 illustrates the bar chart through an example based on a PCA plot of the well-known Iris data set from the UCI repository [9].

Figure 3 shows the effect of discarding features according to the different measures, and how this affects the maximization of the variance (i.e., the objective function of PCA). Removing the least important variable modifies the projection, and therefore, the variance obtained in the low-dimensional space. The variance, as established by the sum of the two eigenvalues, is initially 3.86 when considering the four data variables. For this data set, the proposed approach recommends removing "sepal-width," in which case the variance decreases to 2.99. However, point displacement and axis length recommend removing "petal-width" (it is the smallest bar of the bar chart for both measures), which causes the variance to drop to 2.95.

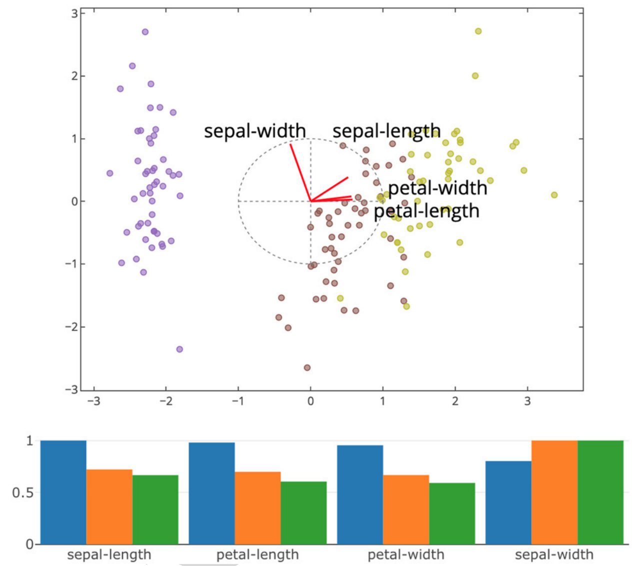
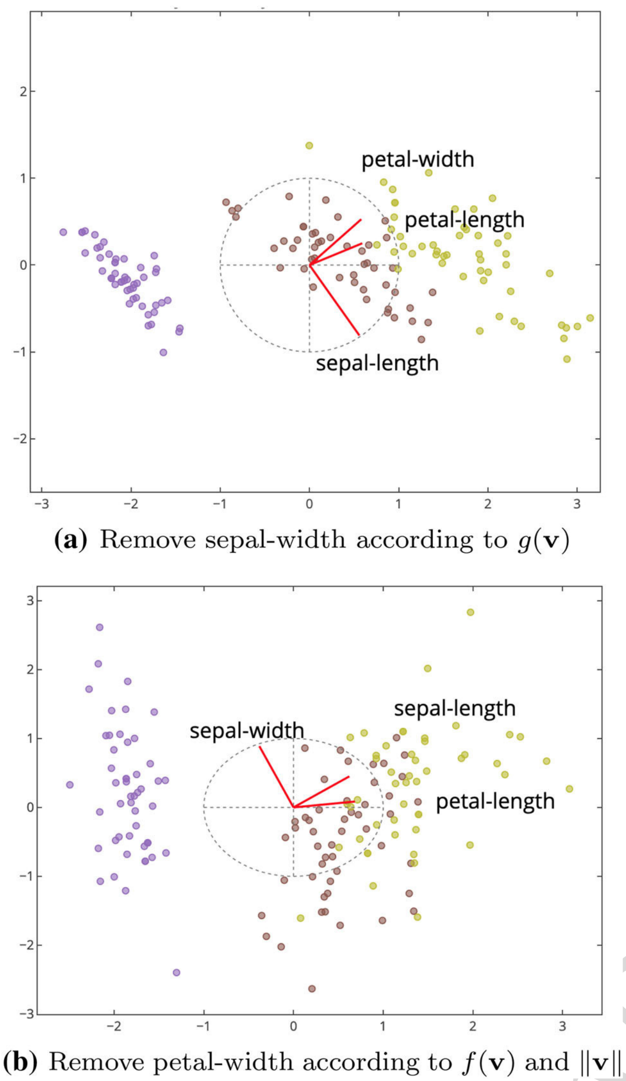



**Fig. 2** SC plot related to PCA of the Iris data set (setosa in purple, versicolor brown, virginica yellow). The corresponding bar chart shows the three different studied measures (the proposed approach $g(\mathbf{v_i})$ in blue, the average point displacement $f(\mathbf{v_i})$ in orange, and the axis length $\|\mathbf{v}_i\|$ in green), which guide the feature selection process. In this case, the features arranged according to $g(\mathbf{v_i})$ in decreasing order

Figure 4 shows a more complex scenario which uses the Olives data set [48], composed of information (8 features) about 572 olive oils. It presents a flow chart showing the greedy procedures that reduce the number of features from eight to four, depending on each of the three metrics. At the top, we show the initial PCA plot, its related variance, and the bar chart with the three metrics for the whole feature set. The ovals in the graphic indicate the least important feature regarding each metric (i.e., the shortest bar). The arrows indicate the greedy decisions when the analyst follows the recommendation to remove a selected feature. Subsequently, a new SC plot related to PCA is computed with the remaining features. For simplicity, we only show the resulting bar chart together with the obtained variance. Note that in some cases the least important feature is the same. For example, the first feature to be discarded is "stearic" for the three metrics. In practice, the decisions taken and the stopping criterion depend on the metrics and on the user's domain knowledge. For illustration purposes, we have only shown all possible decisions that lead to a subset of four features. Finally, the PCA plots obtained for each metric are shown at the bottom. Note that in this example $g(\mathbf{v_i})$ allows analysts to obtain larger variance values.

Although the differences in the resulting variances may seem small, they are relevant if we consider the largest variance that could be obtained at every stage by making an




**(a)** Remove sepal-width according to $g(\mathbf{v})$

**(b)** Remove petal-width according to $f(\mathbf{v})$ and $\|\mathbf{v}\|$

**Fig. 3** Effect of discarding features from the plot in Fig. 2 according to different measures. The plot in **a** is obtained after removing the variable "sepal-width," as suggested by the proposed approach, for which it is the feature with the least influence. Instead, when using point displacement or axis length, the feature to remove is "petal-width," which leads to the plot in **b**. The sum of variances along the $X$ and $Y$ axes of the plot when using the proposed measure (2.99) is greater than the one for the other two approaches (2.95)

optimal choice when discarding a variable. Note that, given a set of $n$ features, it could be possible to compute $n$ new plots where each feature is discarded, and subsequently select the feature for which $\lambda_1 + \lambda_2$ is maximized. However, this strategy is clearly inefficient. Figure 5 illustrates a comparison of the three feature relevance measures and also shows how close they are to the optimal choice, for the standardized Auto MPG data set from [9] that contains eight features. The graphic shows variances associated with PCA plots as variables are discarded from the initial feature subset (one by one, following a greedy approach, as explained in Fig. 4), according to the three measures and the optimal choice strategy. In the example, our metric $g(\mathbf{v_i})$ provides feature subsets that lead to larger variances in general, which are very close to the ones obtained by discarding the optimal variables. Naturally, since the variance of each variable is one (because the data is standardized), the three curves take the value 2 when reducing the selected set to two single features in a two-dimensional plot.

We also tested the performance of the feature relevance measures on a broader experiment involving PCA and LDA plots for randomly selected feature subsets of numerous data sets. For PCA, we used: "Iris," "Auto MPG," "Breast Cancer Wisconsin," "Ecoli," "Glass Identification," "Mice Protein Expression," "Parkinsons," "Spambase," "SPECTF Heart," "Statlog," "Wine," "Forest Types," "Wall-Following Robot Navigation," "Letter Recognition," and "Weight Lifting Exercises" available at repository [9]. For LDA, we used: "Glass Identification," "Iris," "Mice Protein Expression," "Wine," "Letter Recognition," "Weight Lifting Exercises," "Optical Recognition of Handwritten Digits," and "Olives," which include class labels.

The experiments involved 200 trials where in each one we selected a data set at random, and a subset of features, also randomly (with $n > 2$). Subsequently, we applied the three metrics in order to discard a single feature and evaluated the resulting subset. This allows us to compare the performance of each metric on the same subset of features. For both PCA and LDA, we considered that a feature subset is superior to another if the value of its objective function (i.e., $\lambda_1 + \lambda_2$) is larger (for PCA it is the variance, while for LDA it is a measure of class separation).

Since this experiment involves repeated measures, we performed nonparametric Friedman tests to determine whether there were statistically significant differences between the feature relevance measures. These tests were followed up by a multiple comparison analysis to test for individual differences between the metrics. We found statistically significant differences between our approach and the other two described in the literature. Figure 6 shows summary diagrams of comparison intervals of the mean ranks, where there are statistically significant differences (we have used a default significance level of $\alpha = 0.05$) if the intervals do not overlap.

Finally, we present an example in which we discard several features at the same time. Figure 7 shows the initial LDA mapping for the (larger) Weight Lifting Exercises data set [42]. This data set contains 4024 samples of exercises monitored through 53 numerical features and categorized into five classes that indicate the way of executing the exercise. At the top of the plot, we have indicated the value of the objective function for LDA. In addition, we have also included the average silhouette coefficient score [34] of the projected points, which is a popular measure of cluster (or

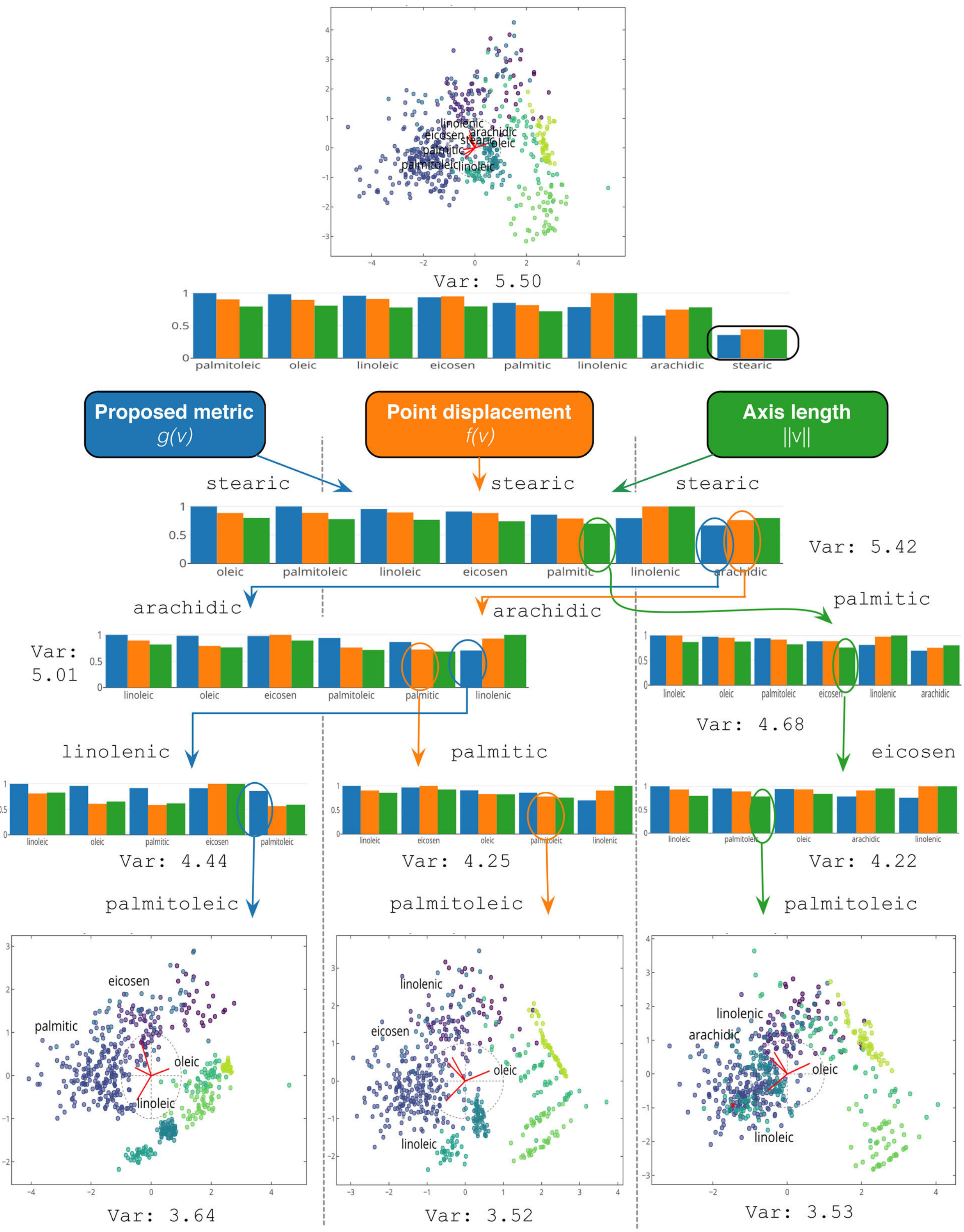



**Fig. 4** Visual feature selection process of a PCA plot with the Olives data set (variance 5.50) according to $\|\mathbf{v}_i\|$, $f(\mathbf{v_i})$, and $g(\mathbf{v_i})$. During the first step, the three metrics recommend to discard the same feature: "stearic." In the second step, $f(\mathbf{v_i})$ and $g(\mathbf{v_i})$ recommend to discard "arachidic" (variance 5.01). Then, "linolenic" would be eliminated by $g(\mathbf{v_i})$ (variance 4.44) while point displacement would discard "palmitic" (variance 4.25). In the last step, both would discard "palmitoleic," which yields a reduced model of 4 variables. Instead, in the second step, the recommended feature to discard by $\|\mathbf{v}_i\|$ is "palmitic," leading to a variance of 4.68. Subsequently, "eicosen" and "palmitoleic" would be discarded due to their length. Finally, by comparison, $g(\mathbf{v_i})$ is able to obtain a plot with a larger variance (3.64) using a subset of 4 features



Journal: 371 MS: 1793 ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:2020/1/10 Pages: 14 Layout: Large

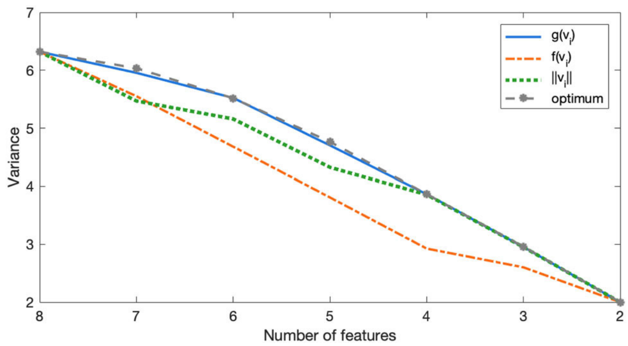



**Fig. 5** Variance reduction obtained by applying feature selection on PCA plots of the standardized Auto MPG data set. In general, our approach is able to obtain feature subsets for which the corresponding variance is greater than the one for the other metrics and is very close to the variance for optimal subsets. Specifically, we computed the sequence of optimal subsets (of seven down to two variables) by discarding the feature that leads to the plot with the largest variance at each step

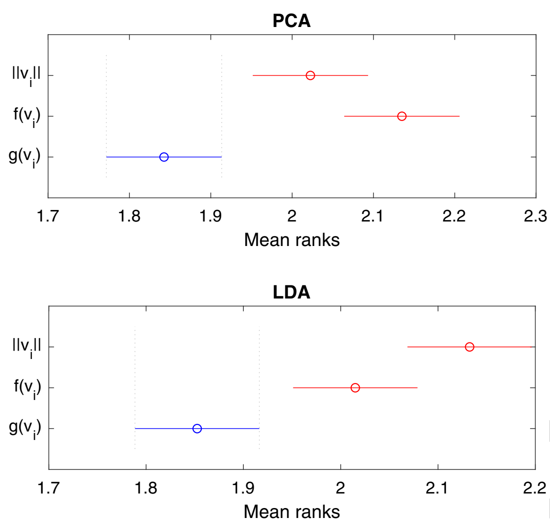



**Fig. 6** Multiple comparison post-hoc analysis of the mean rank differences between the three relevance measures for feature selection on SC plots, where a smaller rank indicates a better performance. Our proposed measure generally leads to PCA plots with greater variance and LDA plots with a higher class separation

class) separation quality. A higher average silhouette coefficient score is associated with denser and more separated clusters. For this particular LDA plot that uses all of the feature in the data set, its value is 0.56. The LDA plot shows the projected data points colored according to their class, together with the axis vectors. We have only included the names of seven features for clarity. Lastly, the figure includes the bar chart with the values of the three analyzed metrics, sorted according to $g(\mathbf{v})$. Note that for certain features (e.g., the seventh, from left to right) the metrics can be quite different.

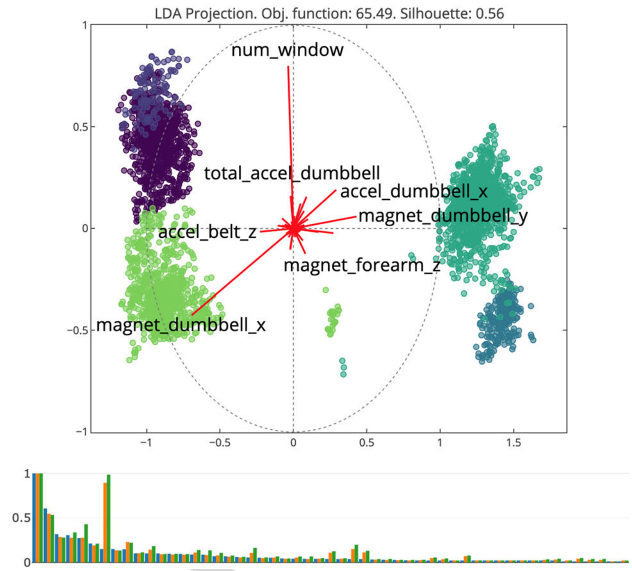



**Fig. 7** SC plot related to LDA of the Weight Lifting Exercises data set, which has 53 features and five different classes. Most axis vectors are clumped in the center of the plot (we have omitted most names of the features for visual clarity). The bar chart shows the importance of all of the features according to the three approaches (the features are ordered according to $g(\mathbf{v})$). Lastly, the LDA objective is 65.49, while the silhouette score is 0.56

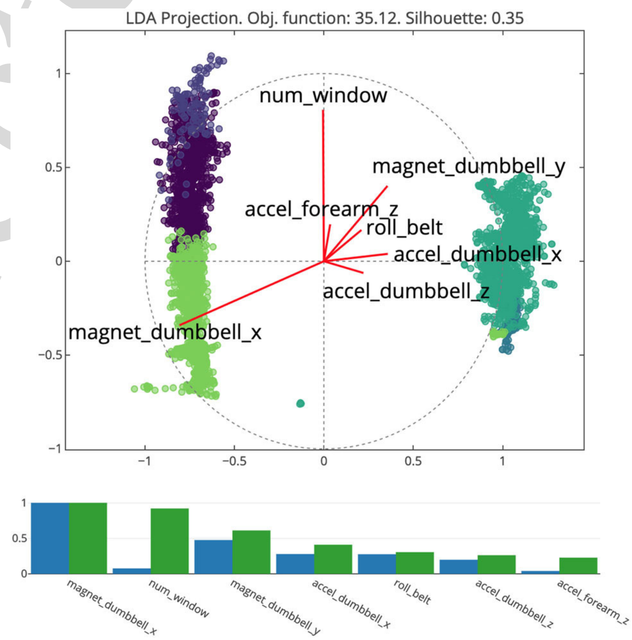



**Fig. 8** SC plot related to LDA of the Weight Lifting Exercises data set, for seven features selected by considering the lengths of the axis vectors $\|\mathbf{v}\|$. Instead of discarding variables one by one, we have eliminated the 46 features with the shortest axis vectors in a few steps. Specifically, we discarded groups of features that shared a similar importance score. The LDA objective is 35.12, while the silhouette score is 0.35. The bar chart shows the influence of the seven remaining features sorted by the length of the axis vectors. We have included the values of the proposed metric (blue bars) for comparison

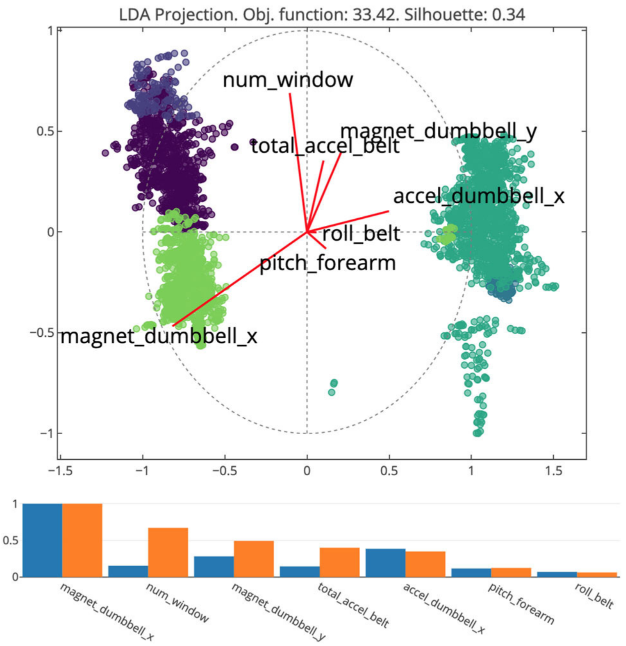



**Fig. 9** SC plot related to LDA of the Weight Lifting Exercises data set, for seven features selected by considering the average point displacement metric $f(\mathbf{v})$. The sorted orange bars in the bar chart show the values of the metric for the seven variables, while the blue bars indicate the value of our proposed measure $g(\mathbf{v})$. In this example, the LDA objective is 33.42, while the silhouette score is 0.34

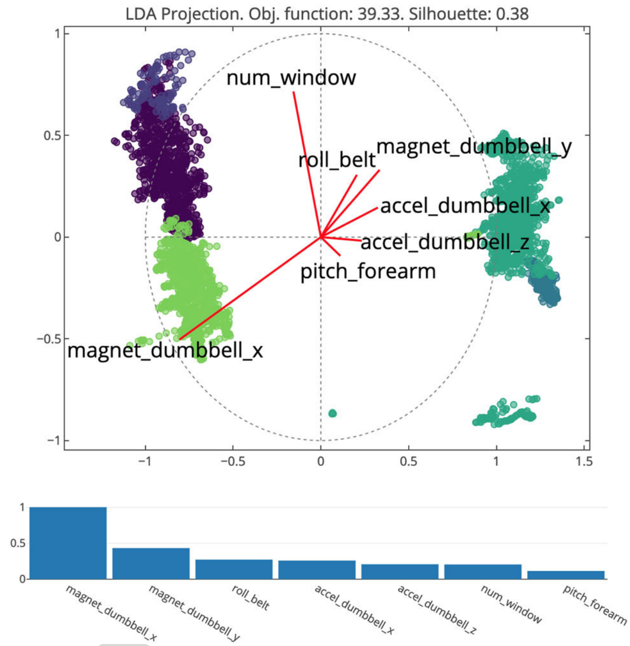



**Fig. 10** SC plot related to LDA of the Weight Lifting Exercises data set, for seven features selected by considering the proposed weighted displacement metric $g(\mathbf{v})$. The sorted bars show the values of the metric for the seven variables. In this case, the LDA objective is 39.33, while the silhouette score is 0.38. These values that measure the quality of class separation are greater than when using $\|\mathbf{v}\|$ or $f(\mathbf{v})$

Figures 8, 9, and 10 show the feature selection processes that results from analyzing the bar charts for the LDA plots regarding $\|\mathbf{v}_i\|$, $f(\mathbf{v_i})$, and $g(\mathbf{v_i})$, respectively. Subsequently, we have discarded groups of features with similar low measure values, in a few iterations, until obtaining a final selection of seven features. The figures also show the corresponding bar chart of each subset of the seven selected variables. Note that some, but not all, of the variables appear in each of the selected subsets. The values of the LDA objective function ($\lambda_1 + \lambda_2$) are 35.12, 33.42, and 39.33, while the silhouette scores are 0.35, 0.34, and 0.38, respectively. Thus, we obtained greater values when using proposed measure $g(\mathbf{v_i})$. This example shows that it is possible to use the metric to remove groups of several features simultaneously.

## 6 Discussion and conclusions

In this paper, we have presented a feature relevance measure for visual feature selection based on SC plots associated with linear projections related to eigenvalue problems like PCA or LDA. In contrast to other approaches in the literature, the measure uses information about the eigenvalues, which are related to the problems' objective function, to determine the most important features for a particular plot. The feature selection is carried out by discarding the least important features, either one by one, or by considering groups of variables. Results show that the proposed measure outperforms other methods based on SC plots described in the literature.

The goal of the approach is to involve the user in order to benefit from its domain knowledge when making decisions regarding which variables to discard. If the number of variables is very large, it can be extremely difficult for users to consider all or most of them simultaneously. In those cases, users could rely on the proposed metric, but would essentially apply it without taking advantage of their expertise. Thus, in those scenarios, it is preferable to first employ an automatic feature selection procedure (e.g., based on entropy) in order to reduce the number of variables to a more manageable amount (around 50 or less if possible), and only then use our visualization approach on the remaining features.

In addition, in principle, $g(\mathbf{v})$, as well as $\|\mathbf{v}\|$ and $f(\mathbf{v})$, could be applied in an automatic manner. However, it is meant to provide suggestions to expert users, within a visualization framework, where they can intuitively decide whether to discard the proposed variable, or to retain it according to their domain knowledge, and to the information shown in the plots. It is important to note that $g(\mathbf{v})$ not only ranks the features (as do many other methods), but the approach is coupled with a plot where users can obtain additional information from the visualizations (e.g., which variables are related and could therefore be redundant, the distribution of data points and





Journal: 371 MS: 1793 TYPESET DISK LE CP Disp.:2020/1/10 Pages: 14 Layout: Large

their attribute values, which variables are related to outliers, etc.).

Although nonlinear mappings are generally be able to represent data more faithfully, it is difficult to understand how the original variables affect the mappings. Instead, when working with linear mappings, we can represent the original features as (axis) vectors in SC plots. Although their lengths and orientations provide information and possibly new insight, we have shown that it is beneficial to consider additional aspects as well, such as the point displacement together with the problem's eigenvalues, when performing feature selection. Specifically, we have shown that the proposed feature relevance measure leads to PCA plots with greater variance, and LDA plots that separate classes better, after removing the least important features for the linear mappings. Nevertheless, the approach can be applied to many other linear methods for dimensionality reduction that are based on eigenvalue problems (e.g., variants of LDA and PCA, locality preserving projections (LPP) [15], neighborhood preserving embedding (NPE) [14], etc.).

The effectiveness of the method depends on how well the linear mappings represent the data. In practice, analysts should examine the relative values of the obtained eigenvalues (e.g., through a typical scree plot) and verify that the values of the first two ($\lambda_1$ and $\lambda_2$) are relatively greater than the rest. If $\lambda_3$ was also relatively large, users could also examine additional SC plots involving the third eigenvector. For example, they could form projection matrices whose columns correspond to eigenvectors 1 and 3, or 2 and 3. Another option consists of creating a three-dimensional SC plot. In that case, the formula for $g(\mathbf{v})$ in (10) can be easily extended in order to involve a third eigenvalue and the average point displacement on the $Z$ axis.

The invariance of our approach with respect to rotations and scalings depends exclusively on the invariance of the linear methods. For example, PCA and LDA are invariant to rotations, but not to scalings. Users must therefore be aware that the data normalization will affect the feature selection. We recommend standardizing the data, since in SC plots the scales of the features should be similar.

We have developed a prototype tool in dash and plotly using scikit-learn and pandas, to compare our measure with previous alternatives. The visual interface can run several linear dimensionality reduction methods and provides a bar chart where analysts can analyze and compare the different feature relevance measures. This combination of an automatically calculated measure, together with an interactive visual tool, allows users to discard features based on the automatic recommendations and on their own expertise about the features. The code of the tool is freely available (http://monkey.etsii.urjc.es/vfsc/VFSC).

We have also tested the tool with experts from the fields of medicine and monitoring of distributed systems, who compared the feature relevance measures analyzed in the paper. They obtained similar results with the measures, since they relied on their domain knowledge to select the features. However, they indicated that the proposed measure provided more reasonable feature candidates to discard. Thus, they were able to select the final feature subsets considerably faster.

Finally, the proposed feature relevance measure allows users to carry out feature selection through a backward elimination approach. We have not defined a stopping criterion for this iterative process, since it depends on the particular analysis task and on domain knowledge. For example, when using LDA, users could discard variables until the classes begin to overlap, or while the performance of a classifier trained on the selected features is above a certain threshold. 6

**Acknowledgements** This work has been supported by the Spanish Ministry of Economy (Grant RTI2018-098694-B-I00). The authors would like to thank Diego Rojo for constructive criticism of the manuscript.



## References

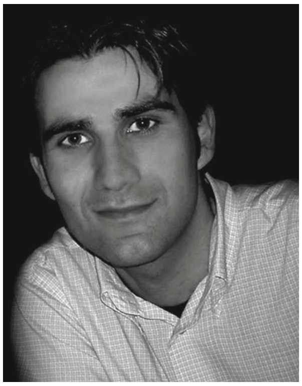

1. Albuquerque, G., Eisemann, M., Lehmann, D., Theisel, H., Magnor, M.: Improving the visual analysis of high-dimensional datasets using quality measures. In: IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 19–26 (2010). https://doi.org/10.1109/VAST.2010.5652433
2. Baumgartner, C., Plant, C., Kailing, K., Kriegel, H.P., Kröger, P.: Subspace selection for clustering high-dimensional data. In: Proceedings of the Fourth IEEE International Conference on Data Mining, ICDM'04, pp. 11–18. IEEE Computer Society, Washington, DC (2004)
3. Bertini, E., Tatu, A., Keim, D.: Quality metrics in high-dimensional data visualization: an overview and systematization. IEEE Trans. Vis. Comput. Graph. **17**(12), 2203–2212 (2011). https://doi.org/10.1109/TVCG.2011.229
4. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. Artif. Intell. **97**(1), 245–271 (1997)
5. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**(1), 16–28 (2014)
6. Chegini, M., Shao, L., Gregor, R., Lehmann, D.J., Andrews, K., Schreck, T.: Interactive visual exploration of local patterns in large scatterplot spaces. Comput. Graph. Forum **37**(3), 99–109 (2018). https://doi.org/10.1111/cgf.13404
7. Chen, B., Hong, J., Wang, Y.: The minimum feature subset selection problem. J. Comput. Sci. Technol. **12**(2), 145–153 (1997). https://doi.org/10.1007/BF02951333
8. Choo, J., Lee, H., Kihm, J., Park, H.: iVisClassifier: an interactive visual analytics system for classification based on supervised dimension reduction. In: IEEE Symposium on Visual Analytics Science and Technology, pp. 27–34 (2010). https://doi.org/10.1109/VAST.2010.5652443
9. Dheeru, D., Karra Taniskidou, E.: UCI machine learning repository. http://archive.ics.uci.edu/ml (2017)
10. Diehl, S., Beck, F., Burch, M.: Uncovering strengths and weaknesses of radial visualizations: an empirical approach. IEEE Trans. Vis. Comput. Graph. **16**, 935–942 (2010)
11. Draper, G.M., Livnat, Y., Riesenfeld, R.F.: A survey of radial methods for information visualization. IEEE Trans. Vis. Comput. Graph. **15**, 759–776 (2009)

12. Guo, D.: Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. Inf. Vis. **2**(4), 232–246 (2003). https://doi.org/10.1057/palgrave.ivs.9500053
13. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
14. He, X., Cai, D., Yan, S., Zhang, H.J.: Neighborhood preserving embedding. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) vol. 1, 2, pp. 1208–1213 (2005). https://doi.org/10.1109/ICCV.2005.167
15. He, X., Niyogi, P.: Locality preserving projections. In: Proceedings of the 16th International Conference on Neural Information Processing Systems, NIPS'03, pp. 153–160. MIT Press, Cambridge (2003). http://dl.acm.org/citation.cfm?id=2981345.2981365
16. Huber, P.J.: Projection pursuit. Ann. Stat. **13**(2), 435–475 (1985)
17. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley, Hoboken (2001)
18. Ingram, S., Munzner, T., Irvine, V., Tory, M., Bergner, S., Möller, T.: DimStiller: workflows for dimensional analysis and reduction. In: IEEE VAST, pp. 3–10. IEEE Computer Society (2010)
19. Jänicke, H., Chen, M.: A salience-based quality metric for visualization. In: Proceedings of the 12th Eurographics/IEEE-VGTC Conference on Visualization, EuroVis'10, pp. 1183–1192. The Eurographics Association, Wiley, Chichester (2010). https://doi.org/10.1111/j.1467-8659.2009.01667.x
20. Johansson, S., Johansson, J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. IEEE Trans. Vis. Comput. Graph. **15**, 993–1000 (2009)
21. Jolliffe, I.T.: Principal Component Analysis. Springer Series in Statistics. Springer, Berlin (2010)
22. Kandogan, E.: Star coordinates: a multi-dimensional visualization technique with uniform treatment of dimensions. In: Proceedings of the IEEE Information Visualization Symposium, Late Breaking Hot Topics, pp. 9–12 (2000)
23. Kandogan, E.: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'01, pp. 107–116. ACM, New York (2001)
24. Kokiopoulou, E., Chen, J., Saad, Y.: Trace optimization and eigenproblems in dimension reduction methods. Numer. Linear Algebra Appl. **18**(3), 565–602 (2011)
25. Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E.: Machine learning: a review of classification and combining techniques. Artif. Intell. Rev. **26**(3), 159–190 (2006). https://doi.org/10.1007/s10462-007-9052-3
26. Krause, J., Perer, A., Bertini, E.: Infuse: interactive feature selection for predictive modeling of high dimensional data. IEEE Trans. Vis. Comput. Graph. **20**(12), 1614–1623 (2014)
27. Lehmann, D.J., Theisel, H.: Orthographic star coordinates. IEEE Trans. Vis. Comput. Graph. **19**(12), 2615–2624 (2013)
28. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H.: Feature selection: a data perspective. ACM Comput. Surv. (CSUR) **50**(6), 94 (2017)
29. Markovitch, S., Rosenstein, D.: Feature generation using general constructor functions. Mach. Learn. **49**(1), 59–98 (2002)
30. May, T., Bannach, A., Davey, J., Ruppert, T., Kohlhammer, J.: Guiding feature subset selection with an interactive visualization. In: IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 111–120 (2011). https://doi.org/10.1109/VAST.2011.6102448
31. McLachlan, G.J.: Discriminant Analysis and Statistical Pattern Recognition. Wiley Series in Probability and Mathematical Statistics. Wiley, Hoboken (2004)
32. Rauber, P.E., da Silva, R.R.O., Feringa, S., Celebi, M.E., Falcão, A.X., Telea, A.C.: Interactive image feature selection aided by dimensionality reduction. In: EuroVis Workshop on Visual Analytics (EuroVA). The Eurographics Association (2015)
33. Reris, R., Brooks, J.P.: Principal component analysis and optimization: a tutorial. In: 14th INFORMS Computing Society Conference, pp. 200–211 (2015)
34. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**(1), 53–65 (1987). https://doi.org/10.1016/0377-0427(87)90125-7
35. Rubio-Sánchez, M., Raya, L., Díaz, F., Sanchez, A.: A comparative study between radviz and star coordinates. IEEE Trans. Vis. Comput. Graph. **22**(1), 619–628 (2016)
36. Rubio-Sánchez, M., Sanchez, A.: Axis calibration for improving data attribute estimation in star coordinates plots. IEEE Trans. Vis. Comput. Graph. **20**(12), 2013–2022 (2014)
37. Rubio-Sánchez, M., Sanchez, A., Lehmann, D.J.: Adaptable radial axes plots for improved multivariate data visualization. Comput. Graph. Forum **36**(3), 389–399 (2017). https://doi.org/10.1111/cgf.13196
38. Sanchez, A., Soguero-Ruiz, C., Mora-Jimenez, I., Rivas-Flores, F., Lehmann, D., Rubio-Sanchez, M.: Scaled radial axes for interactive visual feature selection: a case study for analyzing chronic conditions. Expert Syst. Appl. **100**, 182–196 (2018). https://doi.org/10.1016/j.eswa.2018.01.054
39. Seo, J., Shneiderman, B.: A rank-by-feature framework for interactive exploration of multidimensional data. Inf. Vis. **4**(2), 96–113 (2005). https://doi.org/10.1057/palgrave.ivs.9500091
40. Tatu, A., Bak, P., Bertini, E., Keim, D., Schneidewind, J.: Visual quality metrics and human perception: an initial study on 2d projections of large multidimensional data. In: Proceedings of the International Conference on Advanced Visual Interfaces, AVI '10, pp. 49–56. ACM, New York (2010). https://doi.org/10.1145/1842993.1843002
41. Tatu, A., Maaß, F., Färber, I., Bertini, E., Schreck, T., Seidl, T., Keim, D.A.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In: Proceedings IEEE Symposium on Visual Analytics Science and Technology, pp. 63–72. IEEE Computer Society (2012)
42. Velloso, E., Bulling, A., Gellersen, H., Ugulino, W., Fuks, H.: Qualitative activity recognition of weight lifting exercises. In: Proceedings of the 4th Augmented Human International Conference, AH '13, pp. 116–123. ACM, New York (2013).https://doi.org/10.1145/2459236.2459256
43. Wang, Y., Li, J., Nie, F., Theisel, H., Gong, M., Lehmann, D.J.: Linear discriminative star coordinates for exploring class and cluster separation of high dimensional data. Comput. Graph. Forum **36**, 401–410 (2017). https://doi.org/10.1111/cgf.13197
44. Wang, Y., Nie, F., Lehmann, D.J., Gong, M.: Discriminative star coordinates. Technical Report FIN-02-2016, Otto-von-Guericke-Universität Magdeburg (2016)
45. Yang, J., Peng, W., Ward, M.O., Rundensteiner, E.A.: Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In: Proceedings of the Ninth Annual IEEE Conference on Information Visualization, INFOVIS'03, pp. 105–112. IEEE Computer Society, Washington (2003)
46. Yang, J., Peng, W., Ward, M.O., Rundensteiner, E.A.: Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In: Proceedings of the Ninth Annual IEEE Conference on Information Visualization, INFOVIS'03, pp. 105–112. IEEE Computer Society, Washington, DC (2003). http://dl.acm.org/citation.cfm?id=1947368.1947390
47. Yang, J., Ward, M.O., Rundensteiner, E.A.: Interactive hierarchical displays: a general framework for visualization and exploration of large multivariate data sets. Comput. Graph. **27**, 265–283 (2003)
48. Zupan, J., Novic, M., Li, X., Gasteiger, J.: Classification of multicomponent analytical data of olive oils using different neural networks. Anal. Chim. Acta **292**(3), 219–234 (1994)




Journal: **371** MS: **1793** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2020/1/10** Pages: **14** Layout: **Large**

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Alberto Sanchez Campos** is an associate professor at Universidad Rey Juan Carlos and researcher at Research Center for Computational Simulation. He received M.S. and Ph.D. degrees, obtaining the Extraordinary Ph.D. Award, in Computer Science from Universidad Politécnica de Madrid (Spain) in 2004 and 2008, respectively. His primary research areas are data analysis and visualization, high-performance and large-scale computing, where he has published several journal papers, book chapters and articles in international conferences. He has also done long placement abroad in some prestigious international researching centers, such as CERN, NeSC, NRC-Canada and the University of Melbourne.

**Laura Raya** is a professor and researcher at Centro Universitario de Tecnología y Arte Digital (U-tad), Spain. She received M.S. degree in Computer Science, M.S. degree in Computer Graphics and Ph.D. degree in Computer Science from Universidad Rey Juan Carlos of Madrid in 2008, 2010, and 2014, respectively. Since 2013, she is the head of the master's degree and the manager of Virtual Reality Projects at the Department of Computer Science, at U-tad (Madrid, Spain).

**Miguel A. Mohedano-Munoz** received his degree on Environmental Sciences at Universidad Rey Juan Carlos (Spain) in 2017. Currently, he is doing his Ph.D. thesis about data analysis through dimensionality reduction and machine learning techniques at Universidad Rey Juan Carlos Department. His main research areas are information data visualization and exploratory data analysis.

**Manuel Rubio-Sánchez** received M.S. and Ph.D. degrees in Computer Science from Universidad Politécnica de Madrid in 1997 and 2004, respectively. In 1998, he was awarded a research assistant scholarship from the Spanish Ministry of Education that took place at the Oral Communication Laboratory Robert Wayne Newcomb until 2003. Since 2004, he has had a faculty position at Universidad Rey Juan Carlos (Madrid, Spain), where he is currently an associate. Since 2006, he has performed several research visits at University of California, San Diego. His research interests include exploratory data analysis and visualization, machine learning, and computer science education.