Expert Systems With Applications 100 (2018) 182–196

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Scaled radial axes for interactive visual feature selection: A case study for analyzing chronic conditions

A. Sanchez[a,d,*], C. Soguero-Ruiz[a], I. Mora-Jiménez[a], F.J. Rivas-Flores[b], D.J. Lehmann[c], M. Rubio-Sánchez[a]

[a] *Universidad Rey Juan Carlos, Madrid, Spain*
[b] *Hospital Universitario de Fuenlabrada, Madrid, Spain*
[c] *Otto von Guericke University Magdeburg, Magdeburg, Germany*
[d] *Research Center for Computational Simulation, Madrid, Spain*

ARTICLE INFO

Article history:
Received 30 October 2017
Revised 6 January 2018
Accepted 27 January 2018
Available online 6 February 2018

Keywords:
High-dimensional data visualization
Interactive feature selection
Visual analytics
Exploratory data analysis
Medical chronic conditions

ABSTRACT

In statistics, machine learning, and related fields, feature selection is the process of choosing a smaller subset of features to work with. This is an important topic since selecting a subset of features can help analysts to interpret models and data, and to decrease computational runtimes. While many techniques are purely automatic, the data visualization community has produced a number of interactive approaches where users can make decisions taking into account their domain knowledge. In this paper we propose a new visualization technique based on radial axes that allows analysts to perform feature selection effectively, in contrast to previous radial axes methods. This is achieved by employing alternative scaled axes that provide insight regarding the features that have a smaller contribution to the visualizations. Therefore, analysts can use the technique to carry out interactive backwards feature elimination, by discarding the least relevant features according to the information on the plots and their expertise. Our approach can be coupled with any linear dimensionality reduction method, and can be used when performing analyses of cluster structure, correlations, class separability, etc. Specifically, in this paper we focus on combining the proposed technique with methods designed for classification. Lastly, we illustrate the effectiveness of our proposal through a case study analyzing high-dimensional medical chronic conditions data. In particular, clinicians have used the technique for determining the most important features that discriminate between patients with diabetes and high blood pressure.

1. Introduction

The analysis of high-dimensional data sets is a complex and common problem in fields such as statistics, data mining, or machine learning. In practice, data sets may contain hundreds or thousands of features, many of which can be irrelevant, redundant, or simply add noise. Feature selection consists of the process of discarding those features. The topic is important since analyzing or using the resulting smaller subset can provide several benefits such as: simpler models that are easier to interpret, reduced overfitting, enhanced performance, or shorter computational runtimes.

While many feature selection techniques rely on purely automatic procedures (Guyon & Elisseeff, 2003), the data visualization community has produced a number of interactive approaches where users are integrated into the analysis process with the goal of benefiting from their perceptual capabilities, flexibility, and domain knowledge. With these visualization tools analysts are able to steer the selection process according to their expertise, obtaining subsets of features adapted to the specific problem and application domain, in contrast to automatic methods.

In this paper we focus on interactive visualization methods based on radial axes (Kandogan, 2000; 2001; Rubio-Sánchez, Sanchez, & Lehmann, 2017), which map high-dimensional samples onto a two-dimensional space. The transformations are defined through a set of radial axis vectors, each associated with a feature, which users can modify interactively in order to carry out diverse exploratory tasks, such as analyzing correlations, cluster structure, or class separation, or searching for outliers or data with desired characteristics. However, performing feature selection with these methods is cumbersome. On the one hand, a forward selection is impractical, especially for efficiency reasons. On the other hand, while a backwards selection could be implemented with current

[*] Corresponding author at: Universidad Rey Juan Carlos, Madrid, Spain.
E-mail addresses: alberto.sanchez@urjc.es (A. Sanchez), cristina.soguero@urjc.es (C. Soguero-Ruiz), inmaculada.mora@urjc.es (I. Mora-Jiménez), franciscojavier.rivas@salud.madrid.org (F.J. Rivas-Flores), dirk@isg.cs.uni-magdeburg.de (D.J. Lehmann), manuel.rubio@urjc.es (M. Rubio-Sánchez).

https://doi.org/10.1016/j.eswa.2018.01.054
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

techniques, the size of the axis vectors and the scale of the plots complicate determining which features should be discarded, from both a visual and an interactive point of view.

Alternatively, in this paper we introduce a new approach based on radial axes that is designed to facilitate performing backwards feature elimination, where users can progressively discard features with a small influence either on the visualizations or on a specific task (e.g., class or cluster separation). Specifically, this is accomplished by employing a set of scaled radial vectors that provide a clearer visual guidance for determining which features have the least impact on the low-dimensional plots, and therefore represent reasonable candidates to be discarded in a backwards elimination process. In practice, analysts determine the contribution of the features to the plots and their related analysis tasks by examining the lengths and orientations of the axis vectors. Moreover, they can also take into consideration their expertise when deciding whether a feature should belong to the final selected subset. Lastly, we illustrate the effectiveness of our approach through a case study related to a real medical chronic conditions data set. Concretely, clinicians have used the technique, in combination with their expert domain knowledge, in order to obtain insight regarding the discriminative power of the data features for classifying diabetes and/or high blood pressure patients.

The rest of the paper is organized as follows. Section 2 describes the most relevant methods related to our proposal. In Section 3 we describe our approach based on scaled axes, illustrating how the proposal can be used to perform visual feature selection. Section 4 shows its capabilities through the case study related to medical data. Finally, Section 5 presents a discussion with the main benefits and limitations of the proposal, while Section 6 presents the conclusions and future work.

2. Related work

In this section we present a brief introduction to feature selection methods (with emphasis on visual techniques), and describe the most relevant radial axes methods for multivariate visualization related to our proposal.

2.1. Feature selection

There is a vast literature on automatic feature selection techniques (Blum & Langley, 1997; Chandrashekar & Sahin, 2014; Guyon & Elisseeff, 2003). *Feature ranking* methods sort the features according to some criteria and then select the features progressively (*forward selection*), consider all of the features initially and discard them sequentially (*backwards elimination*), or simply apply some threshold to select the top-ranked features. If the ultimate goal is classification, these strategies are also called *filters*, and discard features as an independent preprocessing step before training a classifier. Alternatively, *wrapper* methods select subsets of features according to the accuracy of classification algorithms, which can be regarded as black boxes that score subsets of features. Lastly, *embedded* methods use a hybrid strategy that incorporates the feature selection process when training a particular classifier.

The method proposed in this paper can be regarded as a feature ranking procedure for backwards elimination feature selection. However, instead of defining an automatic algorithm, it relies on interactive visualizations of data where users can apply their domain knowledge to steer the process of discarding features. Recently, the data visualization community has developed several visual feature selection methods and tools that also take into account user interaction. Most of the approaches propose graphical user interfaces that show several visualizations simultaneously. Some contain well-known graphics in order to show overviews or properties of the data, while others constitute novel visualization methods. In order to perform feature selection many of these methods rely on *quality metrics*, which are measures that extract meaningful information about data. While some of these metrics are popular statistical estimates (correlation, Fisher score, or entropy gain, among others), many others constitute heuristic measures (May, Bannach, Davey, Ruppert, & Kohlhammer, 2011).

Several of the earliest proposals are due to Yang et al., which developed hierarchical methods for visual feature reduction. Yang, Peng, Ward, and Rundensteiner (2003a) propose a dimensionality reduction method based on InterRing visualizations (Yang, Ward, & Rundensteiner, 2002), which groups features hierarchically according to their similarity. The method was later extended to rank and filter out features (Yang, Ward, Rundensteiner, & Huang, 2003b). Guo (2003) describes an interactive tool using several visualizations (e.g., parallel coordinates (Inselberg & Dimsdale, 1990) and entropy matrices) to identify subspaces and high-dimensional (hierarchical) clusters. The approach uses various heuristics, including a measure of the “goodness of a clustering”, and orderings related to paths on minimal spanning trees (MST). An interactive framework for ranking features based on ordering histograms and scatter plots is proposed in Seo and Shneiderman (2005). The work relies on numerous heuristics related to the distributions that appear in the visualizations (e.g., uniformity, number of outliers or gaps, or modality). Similarly, Johansson and Johansson (2009) use heuristics related to the importance of a feature for correlation, outlier, and cluster detection. By weighting these measures interactively, users can generate feature orderings and reduce the number of features. Ingram et al. (2010) present the DimStiller system for feature reduction and analysis. It uses abstractions (e.g., operators, expressions, or workflows) to combine different visualization techniques, and structure and guide the data analysis process. In particular, the approach can be used to determine whether features are meaningful, relationships between features, or the validity of detected clusters. May et al. (2011) propose an interactive visualization technique denoted as SmartStripes for guiding the feature selection process, which can be used with categorical features. Tatu et al. (2012) examine clusterings in different sets of subspaces, which can be interactively explored by relying on subspace similarity and interestingness measures. The visualization tool allows to visualize features and subsets of features at various levels of detail, through parallel coordinates, lists of scatter plots, or multidimensional scaling (MDS) (Cox & Cox, 1994) visualizations. Krause, Perer, and Bertini (2014) describe the INFUSE system, which is designed to help interpret how predictive features are ranked across feature selection algorithms and classifiers. For each feature, the tool displays a circular glyph depicting information related to several feature selection methods, which are based on measures of information gain, Fisher score, odds ratios, and relative risks. In addition, the tool depicts the results of several classification algorithms for the feature selection methods, across several cross-validation folds. Lastly, Rauber et al. (2015) propose a tool for interactive image feature selection including five different views (observation, projection, feature, group, and feature scoring) that show information at various levels of detail. The tool uses recursive feature elimination (RFE) (Guyon, Weston, Barnhill, & Vapnik, 2002) and an ensemble of randomized decision trees (Geurts, Ernst, & Wehenkel, 2006), and the projection view employs the least square projection (LSP) (Paulovich, Nonato, Minghim, & Levkowitz, 2008) dimensionality reduction technique.

Table 1 presents a brief summary of the previous visual feature selection methods. In particular, the table considers: (a) the goal or task they are designed for, (b) the reduction approach, which can consist of progressively discarding features one by one, or of selecting entire subsets of features in a single step, (c) the auxil-

Table 1
Summary of visual feature selection methods in the literature.

Method	Task	Reduction approach	Auxiliary visualizations	Quality metric
Yang et al. (2003a)	Dimensionality reduction	Subset selection	InterRing	Similarity
Yang et al. (2003b)	Feature ranking	Subset selection	InterRing	Similarity Importance
Guo (2003)	Feature insight Clustering	Feature reduction Subset selection	Entropy matrix Parallel coordinates Interactive histograms Bar and line charts	Goodness of clustering Maximum conditional entropy MST ordering
Seo and Shneiderman (2005)	Feature ranking	Feature reduction	Score matrix Histograms Scatterplots Box plots	1 and 2-dimensional metrics Modality Outlierness Gaps
Johansson and Johansson (2009)	Feature ranking	Feature reduction	Score matrix Scatter plot matrix Parallel coordinates	Correlation Distribution density
Ingram et al. (2010)	Feature insight Cluster validation	Feature reduction	Scatter plot matrices Correlation matrices Scree plots	Intrinsic dimensionality Variance and correlation MDS stress
May et al. (2011)	Feature insight	Subset selection	Histograms	Mutual information
Tatu et al. (2012)	Clustering	Subset selection	Parallel coordinates Scatterplot lists MDS of subspaces	Subspace redundancy Subspace interestingness
Krause et al. (2014)	Feature insight Classification	Feature reduction Subset selection	Glyphs Bar charts	Information gain Fisher score Odds ratio Relative risk
Rauber et al. (2015)	Classification	Feature reduction	Scatterplots LSP	RFE Random forests

iary visualization methods, and (d), the quality metrics used. It is worth mentioning that the capability of a tool for feature selection not only depends on the different graphics and the associated interaction techniques, but also on the nature of the data set, and on the quality metrics used to rank the features (or feature subsets), which are remarkably diverse. Bertini, Tatu, and Keim (2011) carry out a thorough literature review in order to provide a unified picture of proposed quality metrics for high-dimensional data visualization).

2.2. Radial axes methods

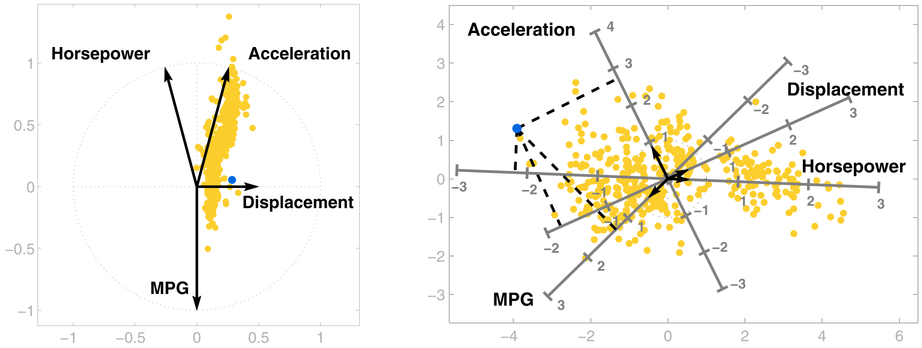
In this paper we propose a new approach based on radial axes visualizations that allows analysts to perform feature selection effectively. Radial axes methods are popular multivariate visualization techniques that produce dimensionality reduction mappings. The simplest method is star coordinates (SC) (Kandogan, 2000; 2001), which is an extension of the scatterplot for more than two features, and has been used for exploratory tasks such as analyzing cluster structure, outliers, or trends. Let $\mathbf{X}$ be an $N \times n$ data matrix, containing N samples, each characterized by n features. The method maps high-dimensional samples $\mathbf{x} \in \mathbb{R}^n$ onto a plane by relying on a set of n axis vectors $\mathbf{v}_i \in \mathbb{R}^2$, for $i = 1, \ldots, n$, with a common origin point. Each $\mathbf{v}_i$ is associated with the i-th feature. In particular, the low-dimensional representation $\mathbf{p} \in \mathbb{R}^2$ (also denoted as an "embedded point") of a sample $\mathbf{x} = [x_1, x_2, \cdots, x_n]^\top$ is a linear combination of the vectors $\mathbf{v}_i$. Formally,

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

where $\mathbf{V}$ is the $n \times 2$ matrix whose rows are the vectors $\mathbf{v}_i$. The method therefore generates linear mappings specified by $\mathbf{V}$. In SC, the orientation of an axis vector determines the direction in which a feature increases, while the length is related to its contribution to the plot. For illustration purposes, Fig. 1(a) shows an example using four features ('Acceleration', 'Horsepower', 'Displacement', and 'MPG') of the Auto MPG data set, available at the UCI Machine Learning Repository (Lichman, 2013). The axis vectors have been chosen to search for cars with large values of 'Horsepower' and 'Acceleration', but low values of 'MPG', which would be represented as dots at the top of the plot. The visualization also includes an axis vector for 'Displacement', which plays a role horizontally. It is important to note that although the length of its axis vector is smaller than the remaining lengths, its contribution to the plot is important since it has a larger component in the horizontal direction.

In practice, users can modify the axis vectors interactively in order to carry out diverse analysis tasks. However, another possibility is to automatically obtain sets of axis vectors from linear methods such as principal component analysis (PCA) (Jolliffe, 2010), independent component analysis (ICA) (Hyvärinen, Karhunen, & Oja, 2001), linear discriminant analysis (LDA) (McLachlan, 2004), and so forth. Consider a linear method that maps data points onto a plane through $\mathbf{p} = \mathbf{Ax}$, where $\mathbf{A}$ is a known $2 \times n$ matrix. Clearly, we can build a SC model that generates the same plot by setting $\mathbf{V} = \mathbf{A}^\top$, due to (1). In other words, we can recover the SC axis vectors (they would be the columns of $\mathbf{A}$) that lead to the plot related to the linear method. In the SC model, the possibility to visualize these axis vectors, together with the plotted points, allows us to determine relationships between the features and their contribution to the plots. Rubio-Sánchez, Raya, Díaz, and Sanchez (2016) introduced this idea to analyze plots based on LDA. Recently, Wang et al. (2017) have denoted it as discriminative star coordinates, and it has also been applied to the results of unsupervised LDA (Ding & Li, 2007), which combines k-means clustering (MacQueen, 1967) and LDA. Lastly, these works carry out feature selection by only comparing the lengths of the axis vectors. In other words, they do not take advantage of their orientations, which should also be considered (see Section 3.5).

Rubio-Sánchez et al. (2017) present a hybrid approach that bridges the gap between SC and principal component biplots (Gabriel, 1971; Gower, Gardner-Lubbe, & le Roux, 2011) called adaptable radial axes (ARA) plots. In SC, users can update the axis vectors freely, but it is difficult to recover high-dimensional data values accurately, which is one of the main disadvantages of the



(a) (b)

Fig. 1. Radial axes plots of the Auto MPG data set: (a) SC plot; (b) ARA plot, where the axis vectors have been selected to generate the PCA projection of the data onto a plane.

method (Draper, Livnat, & Riesenfeld, 2009). Alternatively, with principal component biplots users can approximate the feature (i.e., data) values of an entire data set as accurately as possible (in a least squares sense) through projections of the embedded points onto ticked axes (see Fig. 1(b)). However, since the axis vectors are fixed in these visualizations, users cannot modify them in order to carry out several exploratory analysis tasks (e.g., searching for data with certain features, or creating different mappings in order to detect outliers or visualize clusters). In ARA plots analysts can update the axis vectors freely, and also approximate data values through projections onto ticked axes. Fig 1(b) shows an example that uses standardized data. In this case, the means (which are 0) are represented at the origin, and the difference between consecutive tick marks corresponds to one standard deviation of the corresponding feature. Taking this interpretation into consideration, we can approximately determine through orthogonal projections that the car associated with the darker blue point (which is also depicted in the SC plot) has a large value of ‘Acceleration’ (approximately 2.8), and low values of ‘MPG’, ‘Horsepower’ and ‘Displacement’. Although the estimated values are simply approximations, it is considerably simpler to obtain them visually using ticked axes than in the SC graphic (see Rubio-Sánchez & Sanchez (2014)). Additionally, it is also worth mentioning that, similarly to SC, it is possible to configure the axis vectors to generate any linear mapping. In this example, the particular choice of axis vectors leads to a PCA plot of the data.

Formally, given a set of axis vectors coded in $\mathbf{V}$, ARA plots find the low-dimensional embedded point $\mathbf{p}$ of a data point $\mathbf{x}$ by solving the following optimization problem:

$$\underset{\mathbf{p}\in\mathbb{R}^2}{\text{minimize}} \quad \|\mathbf{V}\mathbf{p}-\mathbf{x}\|, \tag{2}$$

where $\mathbf{V}\mathbf{p}$ is the vector of approximated values for the data point $\mathbf{x}$. Therefore, in ARA plots the approximated feature values are the dot products between the embedded points $\mathbf{p}$ and the axis vectors $\mathbf{v}_i$. In this scenario, the value represented at the endpoint of the axis vector is $\|\mathbf{v}\|^2$. In addition, a unit of the original feature is located at $1/\|\mathbf{v}\|$ along the axis, which implies that the distance between tick marks separating consecutive integers is also $1/\|\mathbf{v}\|$. Since the length of $\mathbf{v}$ does not correspond to a unit of a feature (unless $\|\mathbf{v}\| = 1$), it cannot be used as a visual reference to indicate the location along the axis where a unit would be represented (see Fig. 2(a) for details). Therefore, the method requires drawing axis lines together with tick marks representing integers of the features. Without these tick marks, users would not be able to approximate data features properly, since it is difficult to visually estimate the reciprocal of the length of an axis vector (i.e., $1/\|\mathbf{v}\|$). Lastly, drawing these ticked axes can produce crowded plots even for a small number of features (see Section 3.4). The method proposed in this work mitigates this drawback.

3. Scaled radial axes plots

For the purpose of analyzing high-dimensional data and carrying out visual feature selection, we propose here a new radial axes method called Scaled Radial Axes (SRA) plots. In this section we describe the approach and indicate the main differences with other techniques based on radial axes.

3.1. Description and mathematical formulation

Users in SRA plots will also be able to recover feature values (x_i) by relying on orthogonal projections onto axes, similarly to ARA plots. In ARA the approximated values correspond to dot products between embedded points and axis vectors, which require axes lines and tick marks to indicate the locations associated with integer approximations. Alternatively, in SRA we consider a more intuitive strategy that uses scaled axes, where a unit of a feature is located exactly at the endpoint of its axis vector. Therefore, in this scenario the length of an axis vector determines the distance between consecutive integers of its corresponding feature. This is illustrated in Fig. 2, which shows the relationships between the distances on the plots and the corresponding approximations on the axes, for ARA and SRA.

In SRA the idea is implemented by recovering the i-th data feature of a data point through the following scaled dot product:

$$\frac{\mathbf{v}_i^\top \mathbf{p}}{\|\mathbf{v}_i\|^2}.$$

By dividing by the squared Euclidean norm of an axis vector, its endpoint now represents a unit of its associated feature, as shown in Fig. 2(b). This allows us to omit drawing line axes when the approximations are small (see Section 3.4). Therefore, we define SRA formally through the following optimization problem:

$$\underset{\mathbf{p}\in\mathbb{R}^2}{\text{minimize}} \quad \|\bar{\mathbf{V}}\mathbf{p}-\mathbf{x}\|_2^2, \tag{3}$$

where $\bar{\mathbf{V}}$ is similar to $\mathbf{V}$, but in this case each row is divided by its squared norm. Specifically, the rows of $\bar{\mathbf{V}}$ are:

$$\bar{\mathbf{v}}_i = \begin{cases} \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2^2} & \text{if } \mathbf{v}_i \neq \mathbf{0}, \\ \mathbf{0} & \text{if } \mathbf{v}_i = \mathbf{0}. \end{cases} \tag{4}$$

The optimal solution to (3) is given by:

$$\mathbf{p} = \bar{\mathbf{V}}^\dagger \mathbf{x}, \tag{5}$$

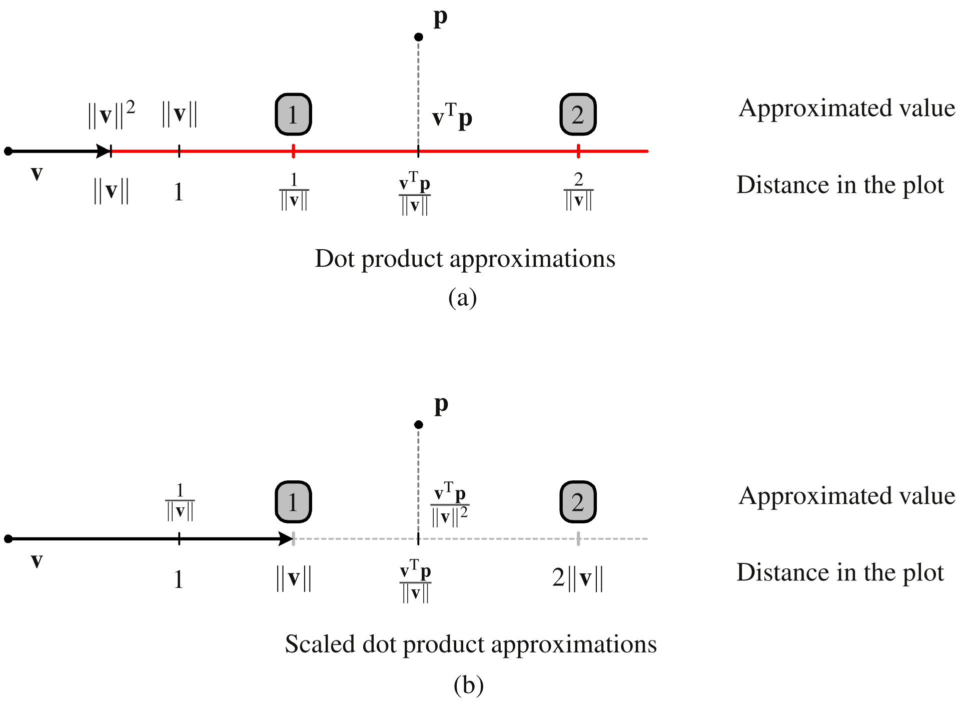

p
$\|\mathbf{v}\|^2$
$\|\mathbf{v}\|$
1
$\mathbf{v}^{\mathrm{T}}\mathbf{p}$
2
Approximated value
v
$\|\mathbf{v}\|$
1
$\frac{1}{\|\mathbf{v}\|}$
$\frac{\mathbf{v}^{\mathrm{T}}\mathbf{p}}{\|\mathbf{v}\|}$
$\frac{2}{\|\mathbf{v}\|}$
Distance in the plot
Dot product approximations
(a)
p
$\frac{1}{\|\mathbf{v}\|}$
1
$\frac{\mathbf{v}^{\mathrm{T}}\mathbf{p}}{\|\mathbf{v}\|^2}$
2
Approximated value
v
1
$\|\mathbf{v}\|$
$\frac{\mathbf{v}^{\mathrm{T}}\mathbf{p}}{\|\mathbf{v}\|}$
$2\|\mathbf{v}\|$
Distance in the plot
Scaled dot product approximations
(b)

Fig. 2. Relationships between approximated values (indicated on the upper part of the horizontal line) and distances in the plots (shown on the lower part of the horizontal line) for: (a) ARA, and (b) SRA. Note that ARA requires axes lines and tick marks (in red) to indicate the values of the approximations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where † denotes the Moore–Penrose pseudoinverse. The method therefore builds a linear mapping from the data space onto the observable plane characterized by the matrix $\bar{\mathbf{V}}^{\dagger}$. We can define the projection of an entire data set in matrix notation through:

$$\mathbf{P} = \mathbf{X}(\bar{\mathbf{V}}^{\dagger})^{\mathsf{T}}, \tag{6}$$

where $\mathbf{P}$ is the $N \times 2$ matrix whose rows consist of the embedded points. In practice it can be computed very efficiently, even for large values of n and N (see Section 5). Finally, when $\bar{\mathbf{V}}$ has full column rank (i.e., when the axis vectors are not all aligned along the same direction), $\bar{\mathbf{V}}^{\dagger} = (\bar{\mathbf{V}}^{\mathsf{T}}\bar{\mathbf{V}})^{-1}\bar{\mathbf{V}}^{\mathsf{T}}$.

3.2. Influence of the axis vectors on the plots

Using $\bar{\mathbf{V}}$ not only determines how the axes are scaled, but it also affects how the axis vectors influence the plots, and how users must interact with them. It is important to notice that shorter vectors will have a stronger impact on the SRA plots, in contrast to longer vectors when using other radial axes plots. Observe that, when searching for the optimal embedded point $\mathbf{p}$, the optimization problem in (3) naturally focuses on minimizing errors on shorter axis vectors. In particular, note that the objective function in (3) can be rewritten as:

$$\sum_{i=1}^{n} \left(\frac{1}{\|\mathbf{v}_i\|^2} \cdot \mathbf{v}_i^{\mathsf{T}}\mathbf{p} - x_i \right)^2. \tag{7}$$

Therefore, if the i-th axis vector $\mathbf{v}_i$ is long, $1/\|\mathbf{v}_i\|^2$ will be small and the choice of $\mathbf{p}$ will barely affect the i-th term of the sum in (7). The scaled axis vectors are useful for visual backwards feature selection since it is easier to spot the longest vectors, associated with features with a small influence on the plots.

However, the length of an axis vector is not the only factor determining the contribution of a feature to a plot. To illustrate this, in this work we compute the average displacement of the low-dimensional points when a feature is discarded as:

$$f(\mathbf{v}_i) = \frac{1}{N} \sum_{j=1}^{N} \|\mathbf{p}^{(j)} - \mathbf{q}_{\mathbf{v}_i}^{(j)}\|, \tag{8}$$

where N is the cardinality of the data set, $\mathbf{p}^{(j)}$ is the embedded point of the j-th data sample for a particular radial axes method, and $\mathbf{q}_{\mathbf{v}_i}^{(j)}$ is the corresponding low-dimensional point when removing the feature associated with the axis vector $\mathbf{v}_i$.

Fig. 3 shows an example of these average displacements for SC, ARA, and SRA plots. Specifically, we generated a random set of $n = 50$ axis vectors, and a random data set of $N = 100$ points. The components of the axis vectors and the values of the data points were drawn from a standard normal distribution. Subsequently, we computed the low-dimensional points associated with the three methods, and obtained their average displacements. The dots on the graphics represent pairs $(\|\mathbf{v}_i\|, f(\mathbf{v}_i))$ and illustrate the average displacement of the mapped points when $\mathbf{v}_i$ is removed from a radial axes plot, as defined in (8). The trend for SC and ARA is clearly increasing, but dots do not follow a strictly increasing pattern as $\|\mathbf{v}_i\|$ grows. Thus, there are features with longer axis vectors that do not contribute as much as others with shorter ones. Similarly, $f(\mathbf{v}_i)$ does not strictly decrease as $\|\mathbf{v}_i\|$ increases for SRA. For instance, the feature with the second shortest axis vector has less impact on the plot than the features with the third to sixth shortest axis vectors. Therefore, besides the length of an axis vector, it is necessary to take into account other factors such as the orientation of the axis vectors, the arrangement of clusters or classes in the plots, or domain knowledge (see Section 3.5). We emphasize this consideration since previous works in the literature have only focused on analyzing the lengths of the axis vectors.

3.3. Arbitrary linear mappings

Similarly to SC and ARA, it is also possible to select a set of axis vectors in SRA to generate any linear mapping from the data space

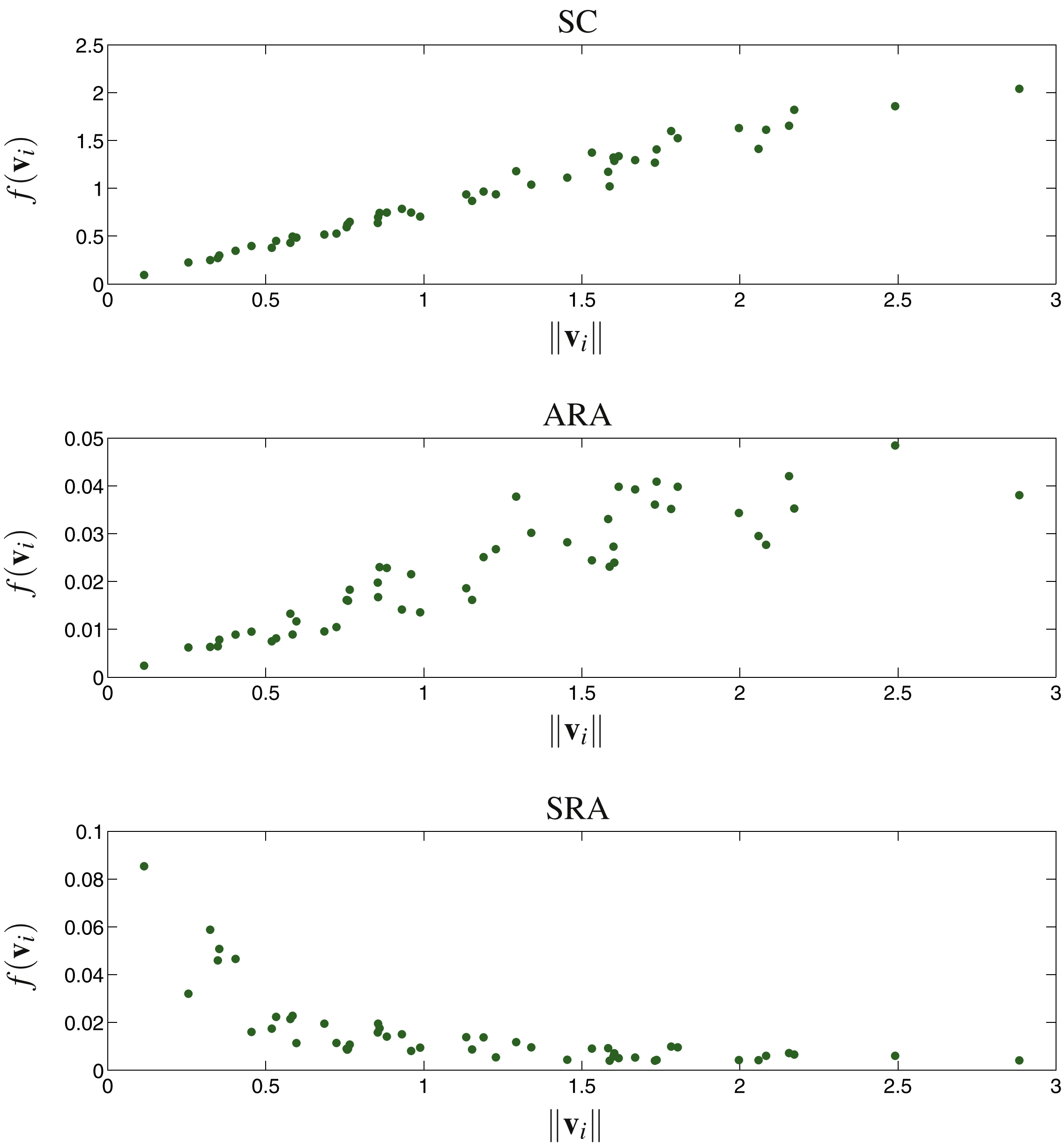


Fig. 3. Example of the contribution of axis vectors to plots (in terms of the average displacement of mapped points when removing a feature) depending on their length, for SC, ARA and SRA.

onto the plane. Let $\mathbf{A}$ be a known $2 \times n$ matrix defining the linear transformation to reproduce. Due to (5), we would need to find a set of axis vectors for which $\bar{\mathbf{V}}^{\dagger} = \mathbf{A}$. This can be accomplished by first computing the pseudoinverse of $\mathbf{A}$, which provides $\bar{\mathbf{V}}$:

$$\bar{\mathbf{V}} = \mathbf{A}^{\dagger}, \tag{9}$$

since $\mathbf{M} = (\mathbf{M}^{\dagger})^{\dagger}$ for any matrix $\mathbf{M}$. Subsequently, the axis vectors (that form $\mathbf{V}$) can be recovered through:

$$\mathbf{v}_i = \begin{cases} \frac{\bar{\mathbf{v}}_i}{\|\bar{\mathbf{v}}_i\|_2^2} & \text{if } \bar{\mathbf{v}}_i \neq \mathbf{0}, \\ \mathbf{0} & \text{if } \bar{\mathbf{v}}_i = \mathbf{0}, \end{cases} \tag{10}$$

which follows from (4), since it defines an involution. The axis vectors are therefore the rows of the pseudoinverse of $\mathbf{A}$, divided by their squared length. The special case in (10) is included by considering that $\mathbf{A}$ can be any matrix, where some rows of $\bar{\mathbf{V}}$ could be equal to $\mathbf{0}$. In those cases, the corresponding axes cannot be specified for the features. Thus, their axis vectors are set to $\mathbf{0}$, and the features are ignored when determining the optimal $\mathbf{p}$.

Fig. 4 shows radial axes plots that produce the LDA mapping of the well-known Iris data set (Lichman, 2013). It contains four data features (‘petal length’, ‘petal width’, ‘sepal length’, ‘sepal width’) and three classes (‘setosa’, ‘versicolour’, ‘virginica’) that identify three species of the iris flower. In particular, we generated the LDA transformation automatically (using standardized data) to separate the three classes, and recovered the layout of axis vectors that would generate that mapping for SC, ARA, and SRA, in (a), (b), and (c), respectively. Note that the plotted points are the same in the three visualizations. The SC plot does not incorporate line axes, and therefore users cannot recover feature values accurately. The ARA plot mitigates this issue by including ticked axes (but can lead to cluttered visualizations for data sets that contain more features). In SRA, the ticked line axes are not necessary and the visualization also allows users to recover feature values by using the vectors instead of line axes (the endpoints of the vectors indicate the location of the units on the axes). Moreover, it is easier to visually identify the less relevant features for the class separation task in

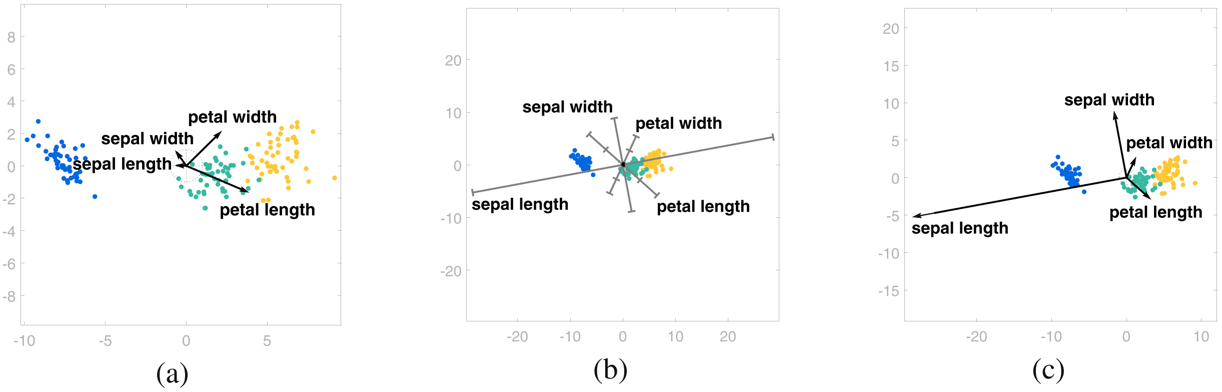


Fig. 4. Radial axes plots that produce the LDA mapping of the Iris data set for: (a) SC, (b) ARA, and (c) SRA. The embedded points are colored according to their class. The axis vectors in the ARA plot are very short and are depicted in black near the origin.

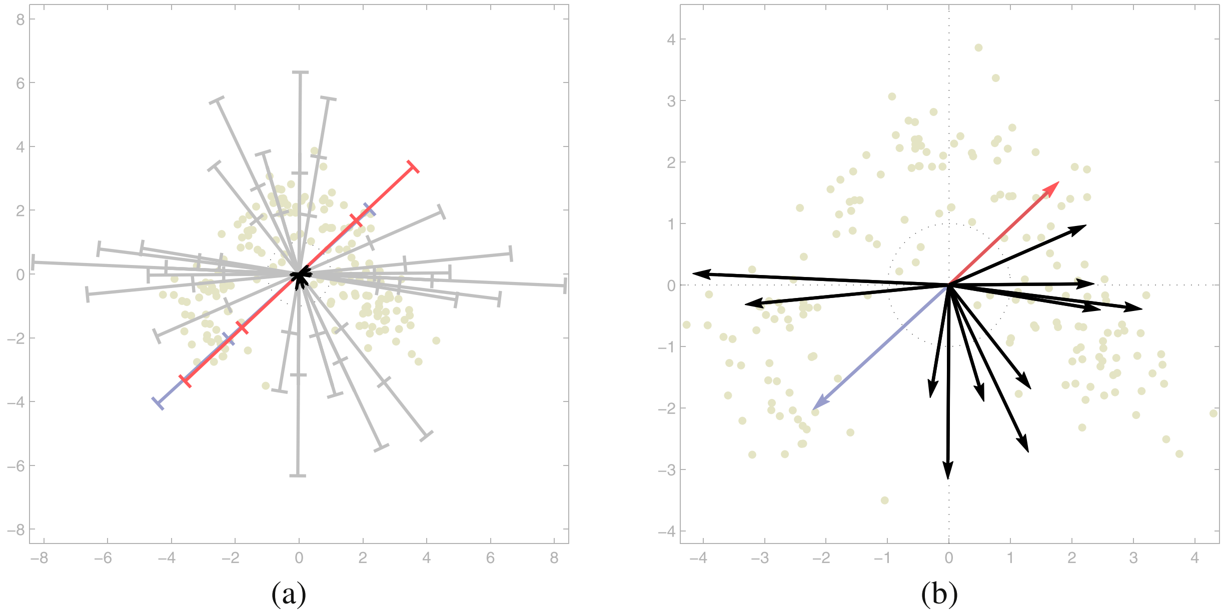


Fig. 5. Projection of the Wine data set, composed of 13 features, considering: (a) ARA plot, with axis vectors barely visible due to their small size (depicted in black near the origin), and axes with tick marks; (b) SRA plot using $\bar{\mathbf{V}}$, where the axis vectors provide enough visual information to recover original feature values. The clutter reduction when using SRA is apparent (due to the absence of axis lines).

SRA (longest vectors) than in ARA (shortest vectors), which is useful for backwards feature selection. Moreover, in this example the axis vectors in the ARA plot are barely visible.

3.4. Clutter reduction

The scaling of the axes is a key contribution regarding the usability of SRA: since the vector length visually encodes a unit of the particular feature, it provides the same information as the first tick mark on an ARA plot. This allows us to omit drawing line axes and their corresponding tick marks when values of the data features are small, which reduces clutter considerably.

Fig. 5 illustrates an example with the Wine data set available in Lichman (2013). This data set contains 13 features corresponding to the chemical analysis of three types of wine, which we have standardized in a preprocessing stage. The visualization in Fig. 5(a) is an ARA plot, where we have selected the axis vectors to obtain the PCA projection of the data onto a plane. The application of SRA in Fig. 5(b) points out some weaknesses of ARA: (1) greater overlap in the ARA plot due to the necessity of drawing the axis lines; (2) though the directions of axis vectors are provided by the axis lines, their specific orientations are barely visible; and (3) axes can share the same or very similar directions in some configurations (e.g., in regular layouts that are often used in the literature), making it difficult to distinguish which tick marks are associated with which features. This last issue is illustrated in Section 5(a), where the colored darker axes exhibit almost identical directions. Note that without colors it would not be trivial to identify which tick marks correspond to a particular axis. Alternatively, the analogous SRA plot in Fig. 5(b) is less cluttered since it does not contain line axes. We have also colored the two vectors that share almost identical directions for reference, though this coloring is not necessary in SRA for distinguishing the axes and approximating values of the corresponding features. Lastly, when axes are omitted it can be easier to incorporate names of features into the plots.

In practice, the absence of tick marks in the SRA plot in Fig. 5(b) does not hamper users' ability to visually compute projected values severely, in comparison with the radial ticked axes plot in Fig. 5(a), which requires them. Note that in radial axes methods the features should share a similar scaling, since otherwise features with larger ranges would have a greater impact on the resulting plots. Therefore, they are usually standardized, transformed to lie in the [0,1] interval, or centered and normalized to have unit range. In

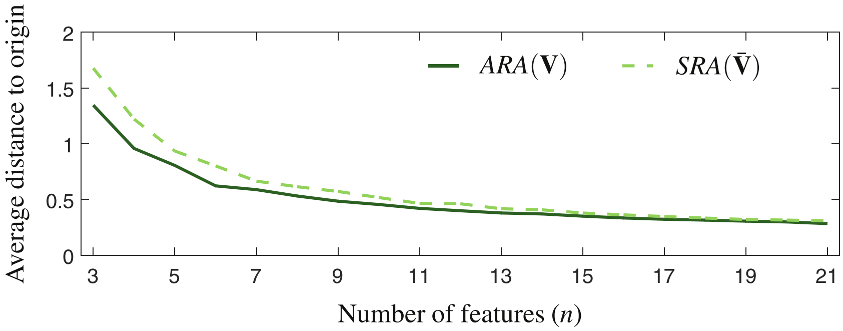


Fig. 6. Average distance from embedded points to the origin, for random configurations of vectors and data whose components were drawn from a standard normal distribution.

all of these cases the absolute values of the approximations corresponding to orthogonal projections onto the axes are generally lower than two. Therefore, users can approximate these values accurately by relying exclusively on the depicted axis vectors, whose endpoints are equivalent to one tick mark in a ticked axis.

Furthermore, the projections onto the axes in SRA are small not only because the data are standardized, but also due to the clumping effect of the projections, which tends to map points closer to the origin as the number of features increases. This effect is shown in Fig. 6, which shows average distances from embedded points to the origin as a function of the number of features (n). The results were averaged over 200 trials of random configurations of vectors, where we mapped 50 samples in each trial. The components of the axis vectors, and the values of the data points, were drawn from a standard normal distribution.

Finally, standardization has two main benefits. Firstly, a unit of a feature represents one standard deviation. Thus, the length of an axis vector in SRA, or the location of the first tick mark in ARA, have a clear statistical meaning. This is important to simplify the graphics, since it allows us to omit numerical labels next to the tick marks (see Fig. 1(b)). Secondly, Rubio-Sánchez and Sanchez (2014) showed that the approximations are more accurate when the data are centered.

3.5. Interactive visual feature selection for class separation

Since the scaling introduced in SRA highlights the least important features, the technique is appropriate for visual sequential backwards feature selection. In practice, users can eliminate features progressively by considering their contribution to a specific plot, which is affected by the lengths and directions of the axis vectors. They can also decide to maintain or discard features according to their domain knowledge.

In addition, assuming the data are categorized into several classes, it is possible to recover the axis vectors in SRA to generate plots related to linear methods designed to enhance classification performance. The most popular linear method is LDA, which maximizes the ratio between the inter-class and intra-class variance. In this paper we will also rely on metric learning approaches such as large margin nearest neighbor (LMNN) (Weinberger & Saul, 2009), and neighbourhood components analysis (NCA) (Goldberger, Roweis, Hinton, & Salakhutdinov, 2005), whose goal consists of enhancing nearest neighbor classification. The resulting SRA plots will provide insight regarding the less discriminative features in the data.

For instance, Fig. 7 shows an SRA plot associated with a LMNN mapping of the Breast Cancer Wisconsin Diagnostic data set (Alcala-Fdez et al., 2008), which includes 30 features from a digitized image of a fine needle aspirate of breast mass, used to determine if a tumor is benign (darker blue dots) or malignant (lighter orange dots). The data set contains information regarding 10 characteristics (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension) of the cell nuclei present in the image. For each characteristic the data set includes three types of measurements: (1) mean, (2) standard error, and (3) the mean just considering the three largest values for each image. In the plots we have appended a numerical suffix to the names of the features to indicate the type of measurement. Fig. 7(a) shows an SRA plot when using the 30 features of the data set. In contrast to SC or ARA plots, features with long vectors can be easily detected in SRA, and discarded in a backwards feature selection process. In this case, the axis vector for ‘Symmetry1’ is clearly larger than the rest. This implies that it barely affects the plot, and it is likely the least discriminative feature. After discarding ‘Symmetry1’, the SRA plot is shown in Fig. 7(b), where axis vectors related to ‘Smoothness3’, ‘Area1’, and ‘Concavity2’ are also longer than the rest. Thus, we can also omit these features by focusing on the lengths of the axis vectors, assuming it is appropriate according to domain knowledge. The resulting plot is shown in Fig. 7(c), where the locations of the points are very similar to those in Fig. 7(b).

As previously indicated, the direction of an axis vector also constitutes a key factor regarding the importance of a feature in a plot. Note that the low-dimensional points will move roughly in the direction of an axis vector when the corresponding feature is removed. Thus, for separating classes (or clusters) in the two-dimensional plot, we can also discard features whose axis vectors are roughly perpendicular to the direction separating these classes, even if those axis vectors are short. Fig. 8 illustrates this idea. In particular, Fig. 8(a) is just a zoomed version of the plot in Fig. 7(c), where both classes are separated fairly well horizontally. Observe that there are several axis vectors whose orientations are roughly perpendicular to the class separation direction. Therefore, although omitting them could originate large displacements of the plotted points, the two classes should remain fairly separated. Specifically, in the plot in Fig. 8(b) we have removed the features ‘Concave points1’ and ‘Concavity3’, which have relatively short axis vectors. The low-dimensional points therefore move vertically, but this barely alters the overlap between classes. Instead, in Fig. 8(c) we have eliminated ‘Radius2’ and ‘Perimeter1’, since their axis vectors point in the separation direction. In this case, although their lengths are similar to those for ‘Concave points1’ and ‘Concavity3’, the points move roughly horizontally. This substantially increases the overlap between the classes, which indicates that these features should belong to the final feature subset.

The process can continue by considering the lengths and orientations of other axis vectors (and possible domain knowledge), and by analyzing the class separation in the two-dimensional plots. The idea is to obtain a final subset of features that allows to separate classes reasonably well. Fig. 9 shows an example of an SRA plot

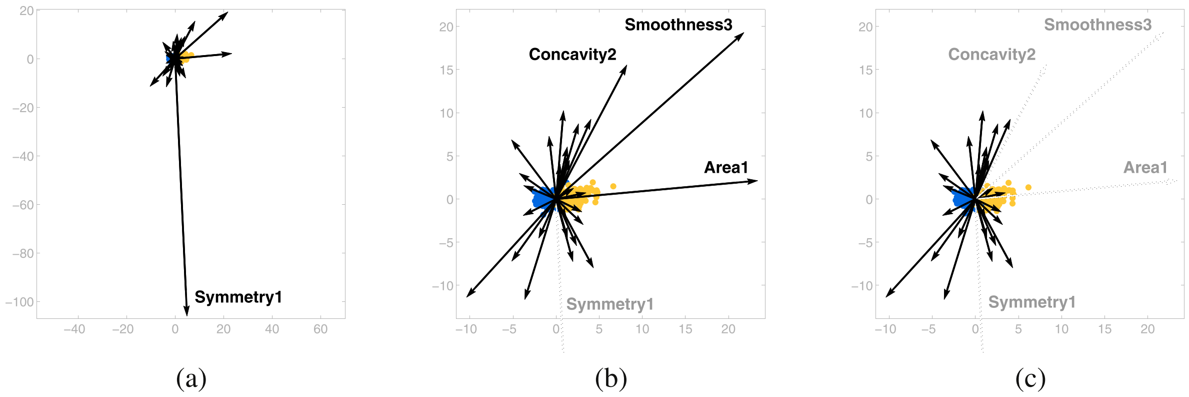


Fig. 7. Interactive visual feature selection. SRA plots related to LMNN for the Breast Cancer Wisconsin Diagnostic data set: (a) considering all features, (b) after removing the ‘Symmetry1’ feature; and (c) when removing features named ‘Smoothness3’, ‘Area1’, and ‘Concavity2’.

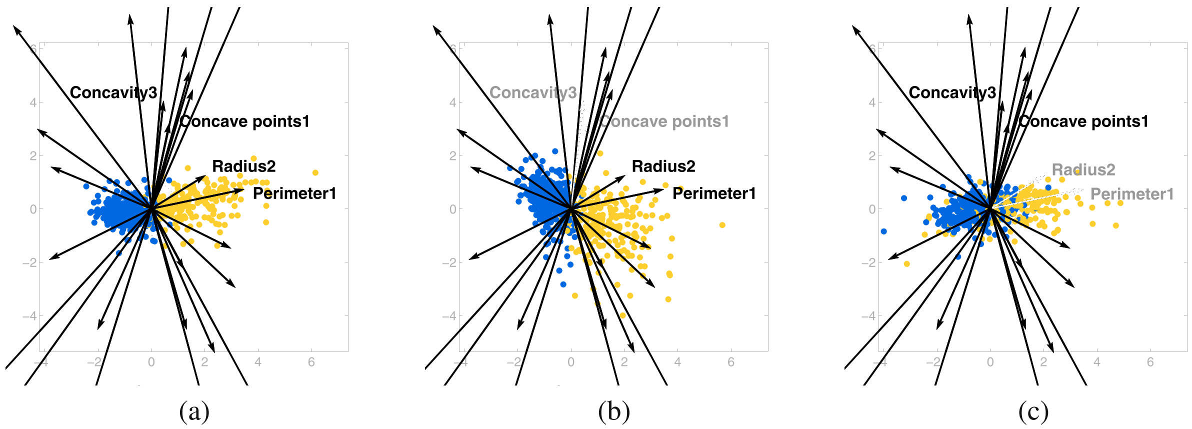


Fig. 8. SRA plots related to LMNN for the Breast Cancer Wisconsin Diagnostic data set: (a) zoom of Fig. 7(c); (b) effect of removing the ‘Concave points1’ and ‘Concavity3’ features in (a); (c) effect of discarding ‘Radius2’ and ‘Perimeter1’ in (a).

where we have retained seven of the original thirty features of the Breast Cancer Wisconsin Diagnostic data set.

Lastly, we measure the quality of SRA projections for class separation as carried out in Leban, Zupan, Vidmar, and Bratko (2006), by computing the leave-one-out accuracy of a voting k-nearest neighbor (k-nn) classifier (Duda, Hart, & Stork, 2001) applied on the plotted two-dimensional points. Specifically, we used $k = \sqrt{N}$, where N is the number of samples in the data set, as suggested by Dasarathy (1991). Thus, for the Breast Cancer Wisconsin Diagnostic data set we chose $k = 24$, since it contains $N = 569$ samples. We obtained a quality of class separation of 96.66% when considering the plot in Fig. 7(a) that involves all of the 30 features in the data set. The score only dropped to 93.32% when considering the plot in Fig. 9, which uses the reduced set of seven features.

4. Case study: analyzing chronic conditions

In this section we describe a case study in which clinicians used SRA for visual feature selection related to chronic conditions.

4.1. Chronic conditions fundamentals

Chronic diseases constitute a well-known problem in current societies, mainly due to the major demographic changes throughout the world over the past few years. On the one hand, the percentage of people over 65 years of age is expected to increase in developed regions (McNicoll, 2002). On the other hand, it is estimated that by the year 2050 about 20% of the whole world population will exceed 65 years. There are also clear positive correlations between age, chronic conditions, and the use of health services. According to World Health Organization (2006), chronic diseases account for 60% of global deaths, and trigger 75% of public health expenditure. Therefore, it is important to determine the diseases that present the highest prevalence, and to identify the factors that best characterize them.

Two diseases that highly contribute to the complex chronic patient group are diabetes mellitus (DM) and high blood pressure (HBP, also called essential arterial hypertension). Not only are they notoriously widespread, but their frequency increases with age, and patients maintain their chronic condition until their death. Specifically, DM is one of the leading chronic diseases in developed countries. It entails many consequences, both from a clinical and social viewpoint, since it increases the risk of many serious health problems. For example, vascular disease is the diabetes complication that can have a more severe prognosis, since it can be accompanied by damage to the coronary arteries, which may lead to myocardial infarction or limb amputation. Other complications of diabetes include kidney problems and blindness. HBP, which is diagnosed when diastolic/systolic blood pressure is 140/90 mmHg or greater, appears among 18% of those who suffer from chronic conditions (World Health Organization, 1999). It can be associated with the onset of other medical conditions such as chronic kidney disease, and it is also related to DM. The simultaneous presence of chronic diseases (comorbidities) can have dramatic consequences. For instance, HPB in patients with DM raises the risk of cardiovascular disease.

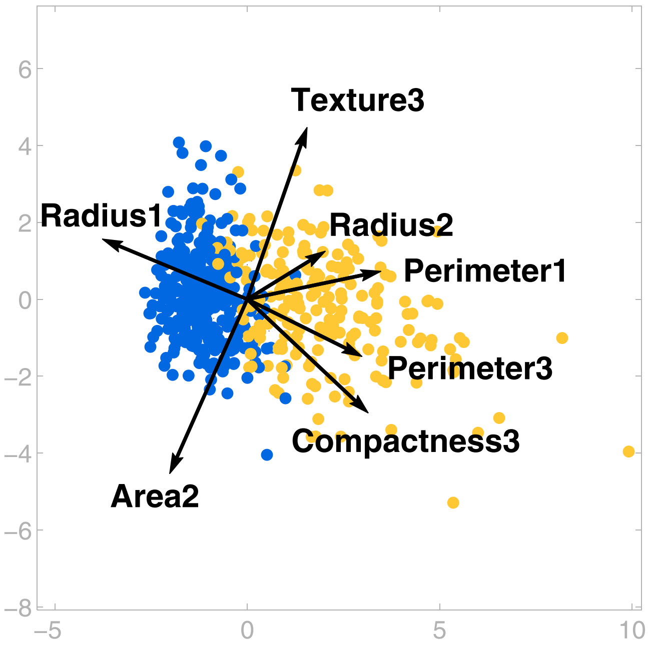


Fig. 9. SRA plot illustrating class separation after selecting seven out of the thirty features of the Breast Cancer Wisconsin Diagnostic data set.

4.2. Chronic conditions data

In this case study we used data provided by Hospital Universitario de Fuenlabrada (HUF) in Madrid, Spain. In order to identify patients with certain chronic diseases, a Patient Classification System (PCS) was applied. In essence, a PCS is a medical decision tree with clinically validated rules, which groups patients according to their health status and resource consumption. Berlinguet, Preyra, and Dean (2005) analyzed different PCS and concluded that the so-called Clinical Risk Groups (CRGs) offered the best performance according to three criteria: clinical relevance of the grouping, resource prediction, and ease of use. This was the reason for using the CRGs (Averill et al., 1999; Hughes et al., 2004) to determine a patient's health status. CRGs are hierarchically organized into nine core categories, from CRG-1 (healthy user) to CRG-9 (catastrophic).

Our data set contains information relative to demographic features (age and gender), diagnoses from primary and specialized care centers, and pharmaceutical drug dispensation during one year. Diagnoses were coded by considering three digits, as stated in the International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM) (Centers for Disease Control & Prevention, 2011). Medical drugs were specified through five characters, according to the Anatomical Therapeutic Chemical (ATC) Classification System (Norwegian Institute of Public Health, 2017) used in Europe. CRGs used this information to assign each patient to a single mutually exclusive health status or risk group.

In this paper we analyzed three chronic conditions (i.e., categories): crg-5192 (HBP), crg-5424 (DM), and crg-6144 (DM and HBP). The first digit of the CRG-code refers to the core group, while the next three digits are associated with the chronic condition category. Specifically, HUF provided us with data of 17,792 patients associated with the three chronic statuses of interest during the year 2012: 12,447 for crg-5192, 2166 for crg-5424, and 3179 for crg-6144. Since class-imbalance is a well-known issue in medical research (Fernández-Sánchez et al., 2017; Soguero-Ruiz, Hindberg, Mora-Jiménez, Rojo-Álvarez, & et al., 2016), we adopted an undersampling strategy taking into account the size of the minority group. Thus, we randomly selected 2166 patients from each group.

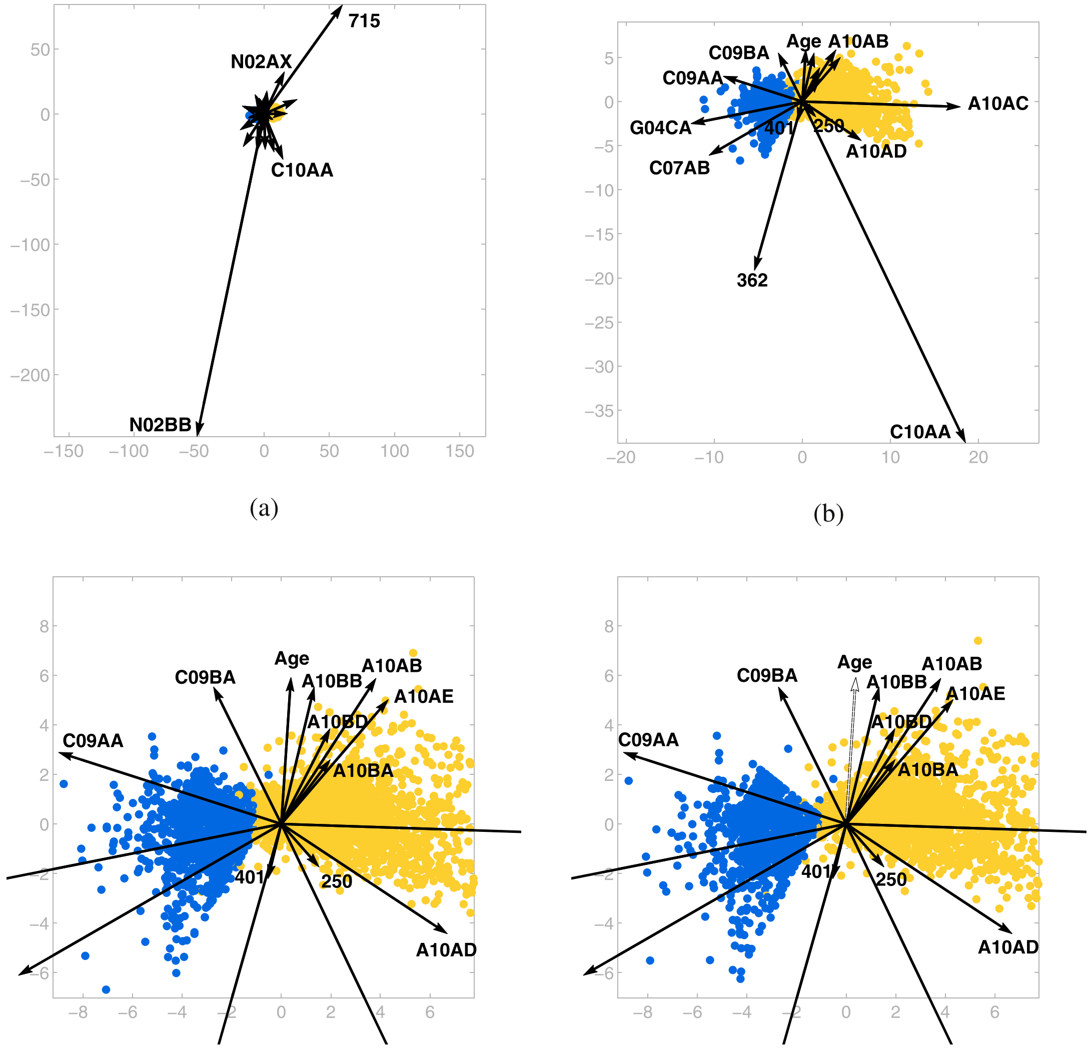
In a previous study we performed a descriptive analysis of diagnosis codes and demographic features in the group of only chronic hypertensive patients (Fernández-Sánchez et al., 2017). Regarding the features in the current work, we have also considered medical drugs apart from diagnosis. Each code of diagnosis and medical drug has been considered as a different feature. In particular, each patient is described by a total of 1517 features for diagnoses, and 746 for medical drugs. The features are integers that count the number of times that a particular patient has been diagnosed with a certain condition, or has been dispensed a particular drug. Around half of the features had a zero count for every single patient, and were therefore discarded. In addition, we reduced the data set even further by computing the entropy gain of each feature according to Rauber and Steiger-Garção (1993), and by selecting the 50 features with the highest gain. According to the domain knowledge of the clinicians who participated in the case study, the resulting subset of features contained the most relevant features related to the chronic conditions under study.

4.3. Visual feature selection with SRA

Since the dimensionality of the data (50) is still high, further feature selection procedures can be useful for identifying features with a greater clinical relevance for characterizing the chronic conditions. In our case study, the medical doctors used SRA, coupled with linear methods for classification, as a basis for performing a sequential backwards visual feature selection. Specifically, the goal was to determine which features were more helpful for discriminating between health statuses: (i) HBP, (ii) DM, and (iii) HBP and DM. Therefore, the clinicians used SRA to graphically identify different health groups, and to evaluate or confirm (in consonance with domain knowledge) the impact of each feature on the plots designed for class separation. Since clinicians were not experts in data visualization methods, we provided explanations of the main properties of SRA, as well as assistance throughout the process.

Firstly, the medical doctors analyzed which features contributed more to distinguishing between the hypertensive and diabetic groups (crg-5192 vs. crg-5424). This is the simplest scenario when considering chronic conditions, since the health statuses are characterized by only one chronic condition. Fig. 10 shows SRA plots associated with the LMNN mapping of the (standardized) data set, where the lighter (yellow) and darker (blue) points represent patients with DM, and HBP, respectively. In Fig. 10(a) we used the initial 50 features. The clinicians then progressively discarded features by relying on the visualizations and their own expertise until obtaining the plot in Fig. 10(b), which only contains 16 features. The quality of class separation only decreased from a score of 98.66% (when using the initial 50 features) to 98.61% when considering just 16 features (in this case we used the voting $66-nn$ classifier, since there are $N=4332$ samples).

The plot in Fig. 10(c) is simply a zoom of Fig. 10(b), where we can gain insight regarding the most relevant features for classifying patients with a single chronic condition. In this example, these features are mainly those oriented horizontally, since classes are separated along that direction. For instance, the features related to the drug codes 'G04CA' (alpha-adrenoreceptor antagonists) and 'C09AA' (angiotensin-converting-enzyme inhibitors, plain) point towards the crg-5192 class, as expected by the clinicians. Analogously, several axis vectors are oriented towards the crg-5424 class. Their contribution to the plots, as suggested by their lengths and orientations, was in accordance with the clinician's background knowledge. For example, the axis vectors for drug codes 'A10AB' (insulins and analogues for injection, fast-acting), 'A10AE' (insulins and analogues for injection, long-acting), 'A10BA' (biguanides), or 'A10BD' (combinations of oral blood glucose lowering drugs) all have positive components along the plot's X axis, since they point



(a)

(b)

(c)

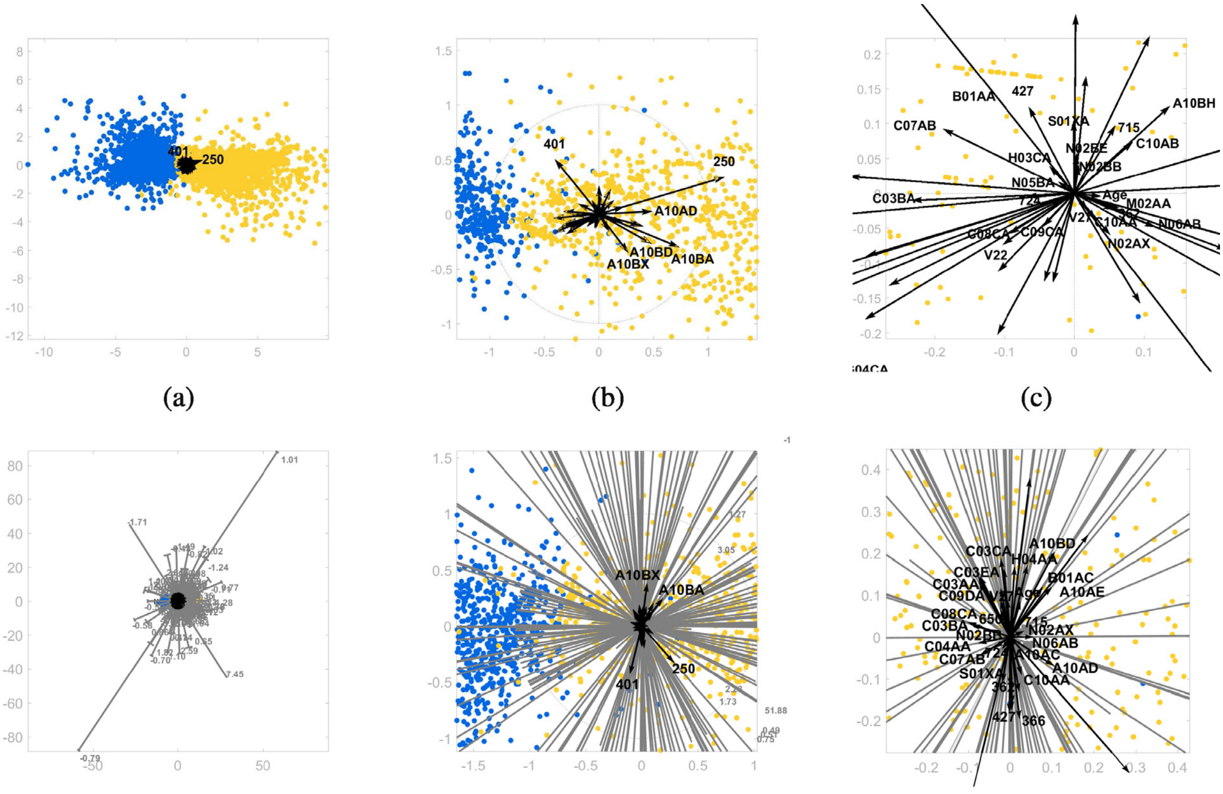
(d)

Fig. 10. SRA plots related to LMNN for patients with hypertension (darker points, crg-5192) and diabetes (lighter points, crg-5424) considering 50 and 16 features, in (a) and (b), respectively. The plot in (c) represents a zoom of (b), and in (d) we show the (minor) effect of removing the feature ‘Age’.

towards the first quadrant. Thus, they are clearly related to diabetes. The feature for the diagnosis code ‘250’ (DM) also appears pointing towards the diabetic group, and has a higher contribution than the ATC codes, since its axis vector is shorter. Clinicians also suggested to retain the drug code ‘C10AA’ (HMG CoA reductase inhibitors) in spite of the long length of its axis vector, since it could have some relation with diabetic patients. Finally, regarding the ‘Age’ feature, the length of its axis vector is similar to that of the remaining ones. However, it does not play a key role in separating the crg-5192 and crg-5424 groups, since its axis vector is roughly perpendicular to the direction that separates the classes. This also occurs for other features like the diagnosis code ‘401’ (essential hypertension). If the ‘Age’ feature is removed (as shown in 10(d)), the classes remain clearly separated, and the quality of class separation is enhanced to 99.01%.

For comparison purposes, in Fig 11 we show SC and ARA plots related to the LMNN mapping, with the initial 50 features. In both cases shorter vectors have a weaker impact on the resulting plots. Thus, in practice it is required to zoom in several times to be able to identify the features to be removed. In the example, the initial SC plot is shown in (a), while (b) and (c) show 4x and 40x zooms, respectively. Similarly, (d) is the initial ARA plot, while (e) and (f) show 20x and 100x zooms, respectively. Observe that the axis vectors (and the axis lines in ARA) overlap considerably, which makes it difficult to visualize and select the shortest axis vectors. In addition, depending on the scale of the data, the projected points may fall outside of the plot. Thus, we can lose the overall picture of the data set, which is necessary for considering the orientations of the vectors (in this case, the direction that separates the classes). In our experiments, all clinicians were able to immediately obtain the longest axis (‘N02BB’) using SRA, and agreed to remove it (see Fig. 10(a)). However, when using SC and ARA they had to zoom in several times, obtaining the plots in (c) and (f), before deciding on the least relevant features. Most importantly, they did not agree on the feature to be removed, as some vectors were of similar size.

In the next study the data set was expanded by including a third health status encompassing both chronic conditions, diabetes and hypertension (crg-6144). In this case, we selected a total of 6498 patients (2166 of each health status), and tested our approach by relying on the NCA mapping of the data set. Fig. 12 shows several SRA plots associated with NCA, where the lighter (yellow), darker (blue), and mid-color (green) points represent patients with DM (crg-5424), HBP (crg-5192) and both chronic conditions (crg-6144), respectively. Similarly to the first study, we gen-



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 11. SC and ARA plots related to the SRA plot in Fig. 10(a) with 50 features. The initial configuration of the SC plot is shown in (a), while (b) and (c) show 4x and 40x zooms, respectively. Analogously, (d) contains the initial ARA plot, while (e) and (f) show 20x and 100x zooms, respectively. On the one hand the axis vectors (and axes lines) overlap considerably. On the other hand, we can lose the distribution of the plotted points when zooming.

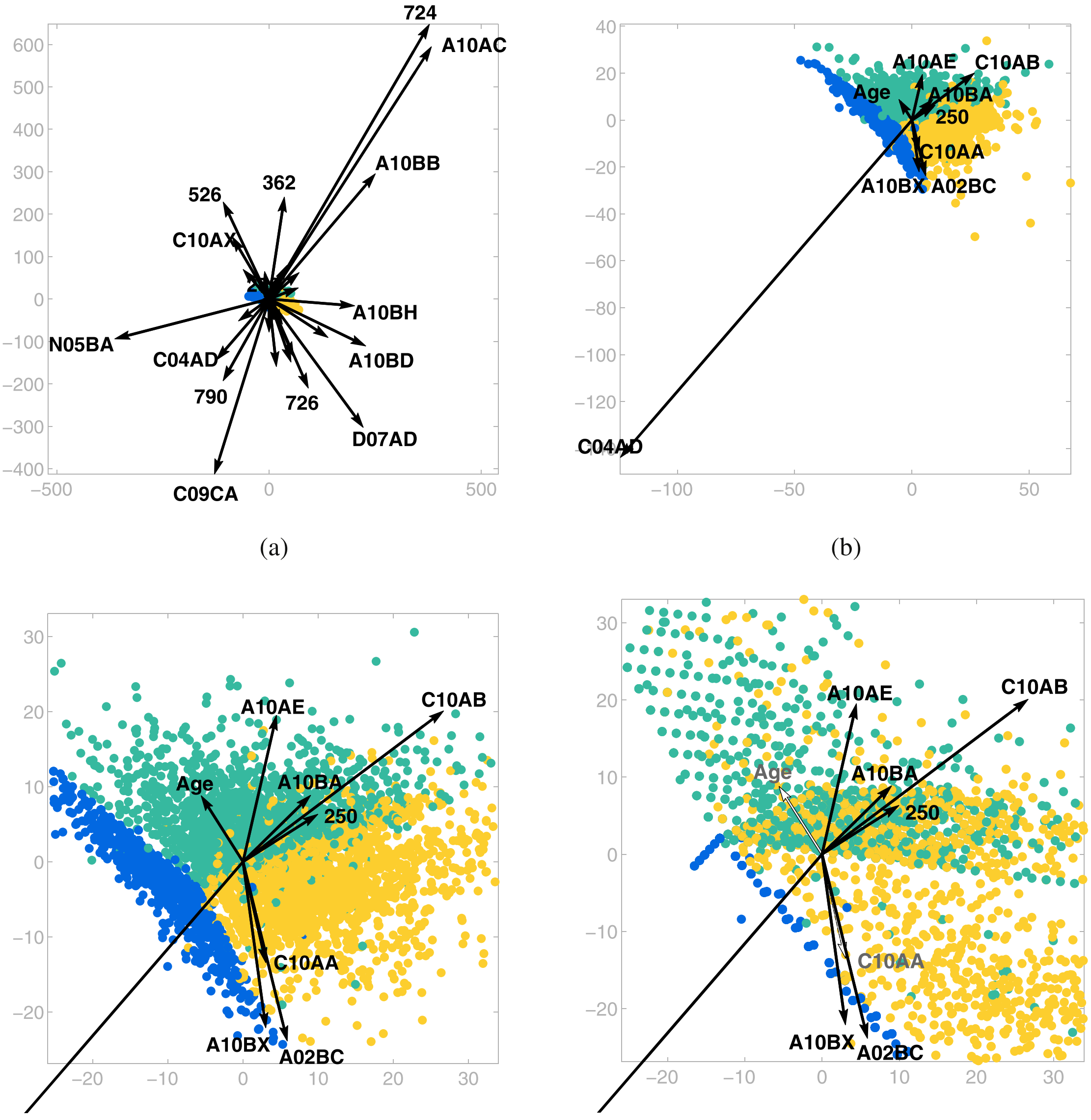
erated an initial plot by using all of the 50 features, as shown in Fig. 12(a). The quality of class separation according to a nearest neighbor classifier was 92.67% (we used $k = 81$, since $N = 6498$). Subsequently, the clinicians progressively eliminated features by relying on the visualization and their domain knowledge until obtaining the plot in Fig. 12(b), which only contains 9 features and provides a quality of class separation of 87.17%.

We can observe the axis vectors (and their contribution) more clearly in Fig. 12(c), which is a zoom of Fig. 12(b). On this occasion, clinicians did not select the diagnosis code '401' because there were other features with more influence for separating both groups. Instead, although in the first study the drug code 'C10AA' (HMG CoA reductase inhibitors) did not contribute much in distinguishing between hypertensive and diabetic patients (according to the layout of vectors obtained when reproducing LMNN), the clinicians suggested to retain it since in their opinion it had a clear relation to diabetes. In this case, it is apparent that the feature 'C10AA' is key for separating the groups (note that its axis vector is one of the shortest ones). This confirms the medical knowledge that reductase inhibitors are related to diabetic patients. Likewise, the feature 'Age' does have a strong impact on class separation, since individuals in CRGs with chronic comorbidities (crg-6144) tend to be older than patients with just one chronic condition (crg-5192 or crg-5424). 'Age' is especially relevant for patients with diabetes, which supports existing knowledge about juvenile diabetes. Finally, in order to visually confirm the importance of both features ('C10AA' and 'Age') we discarded their axis vectors. The resulting plot is shown in Fig. 12(d), where the lighter (crg-5244) and mid-color (crg-6144) classes clearly overlap. In this case, the quality of class separation dropped to 75.45%.

The study carried out, involving clinicians and a real medical data set, shows that SRA can be a valid tool when it is used by domain experts without previous experience in interactive visual data analysis tools. The visualizations have allowed the clinicians at HUF to confirm previous medical knowledge, and to obtain new insight into the area of application.

5. Discussion

In practice, analysts can use radial axes plots for visual feature selection by studying the impact of the features on a plot. However, it is problematic to use these visualizations in a sequential forward selection process, mainly due to the large number of plots that users would have to analyze. Note that having a subset of $m < n$ features, it would be necessary to visualize the $n - m$ additional plots that include one more feature in order to expand the subset. Since this procedure would be carried out multiple times, the number of visualizations would be excessive in a practical setting. In particular, this approach would require $(m + 1)(n - m/2)$ visualizations for obtaining a subset of m features. Alternatively, users in a sequential backwards elimination procedure analyze a single plot to discard one of the features. Thus, this approach requires analyzing $n - m$ visualizations in order to choose a subset of m features, which is much smaller than the number required by the sequential forward selection scheme. Thus, if m is some percentage of n (i.e., $\alpha n = m$, with $\alpha \in (0, 1)$), then the forward selection strategy requires on the order of n^2 visualizations, while the backwards approach needs on the order of n plots. Moreover, when performing a backwards selection it is also possible to identify an entire group (i.e., set) of features to discard by analyzing a single



(a)

(b)

(c)

(d)

Fig. 12. SRA plots related to NCA for patients with just hypertension (darker blue points, crg-5192), just diabetes (lighter orange points, crg-5424), and both comorbidities (mid-range green color, crg-6144) considering 50 and 9 features, in (a) and (b), respectively. The plot in (c) is a zoom of (b), and in (d) we show the (strong) effect of removing the features 'Age' and 'C10AA'. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

plot, which can speed up the selection process notably when the initial number of features is large.

In SRA a backwards feature selection is implemented by removing longer axis vectors, which are easy to spot. In SC and ARA it is possible to perform a similar feature elimination by discarding shorter axis vectors. However, as shown in Fig. 11, it is more difficult to identify these axis vectors. In practice, analysts may need to zoom in on the plots considerably, which is not only time-consuming, but the overall view of the data can be lost in the resulting graphic, since many of the projected points may not appear in the plot. Therefore, in SC and ARA it can be harder to take advantage of the directions of the axis vectors.

Although methods based on radial axes can represent as many variables as desired, in practice n is usually small (see (Chen & Liu, 2004; Gabriel, 1971; Kandogan, 2000; 2001; Sun, Yuan, Hu, & Xiao, 2008; Tsai & Chiu, 2008; Zhang, Orgun, & Zhang, 2006)). Note that if n is large a feature reduction process would be time-consuming and cumbersome, mainly due to the overlap between the axis vectors. In that case one solution consists of carrying out a preliminary feature reduction with an automatic method (in Section 4.2 we have used the entropy gain to reduce the number of features). Another possibility is to generate an SRA plot and eliminate the features related to long axis vectors, according to a length threshold, or to a particular number of features the analysts may wish to retain before applying the proposed feature reduction approach. Another limitation of the approach is related to the type of data it can support. In particular, all of the radial axes methods described in this paper require using numerical data (it is possible to use binary features).

In order to evaluate the method's potential for data analysis, we have developed a data visualization prototype in MATLAB® using the toolbox for dimensionality reduction (Maaten, 2015). In preliminary usability tests, users were able to carry out: i) tasks directly related to the technique like classification, clustering, feature selec-

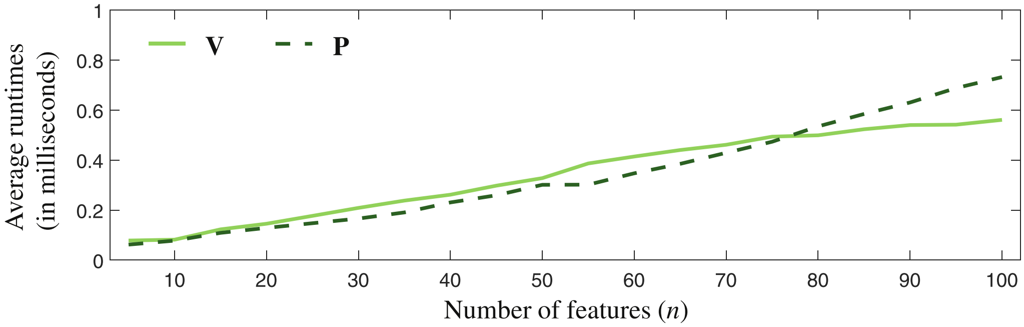


Fig. 13. Average runtimes for computing the axis vectors (**V**) given some initial linear transformation matrix through (9) and (10), and for calculating 10000 embedded points (**P**) through (6).

tion, outlier detection, or attribute value estimation; and ii) other basic data analysis tasks like those described in Amar, Eagan, and Stasko (2005) and Yi, ah Kang, Stasko, and Jacko (2007), such as retrieving values, determining correlations, filtering, etc.

Regarding the efficiency of the approach, it is worth mentioning that the key factor depends on the computational cost of the chosen linear method (e.g., LDA, LMNN, NCA, etc.), which provides a particular $2 \times n$ matrix **A**. The process of determining the axis vectors **V** through (9) and (10), as well as computing the embedded points (**P**) through (6) can be carried out in the order of microseconds, even for a large number of features (n), since these operations can be carried out in linear time with respect to n. Fig. 13 shows average runtimes needed to compute **V** given some random initial matrix **A**, and to project $N = 10000$ random high-dimensional points (**X**), for several values of n. The results were averaged over 1000 trials, and the components of **A** and **X** were drawn from a standard normal distribution. In particular, the simulation was carried out on a personal computer with a fourth generation Intel® Core$^{\text{TM}}$ i7-4712HQ 3.3 GHz processor and 16 GB of RAM. It is apparent that the calculations can be carried out in real time.

Finally, the proposed visualization method is an exploratory data analysis tool that can lead to interesting and possibly unexpected discoveries in an overview phase of a data mining process (Shneiderman, 1996; Witten & Frank, 2005). However, it is worth pointing out that analysts must confirm the findings through appropriate statistical and scientific procedures. In this regard, the insight obtained through the user study with chronic conditions data only provides an initial guidance for a further analysis, which is clearly out of the scope of the paper.

6. Conclusions

This paper has introduced and analyzed a multivariate visualization method called SRA, which is based on a set of radial axis vectors that represent data features, and can generate any linear projection of high-dimensional data points onto a two-dimensional plane. On the one hand, unlike SC, SRA plots allow users to approximate high-dimensional data values. On the other hand, in comparison with ARA, SRA provides less cluttered plots, and allows users to analyze the axis vectors and all of the projected points simultaneously. Moreover, in SRA longer axis vectors generally represent features that have a smaller influence on a projection. Since it is easier to identify these vectors, the technique can be used to carry out an interactive backwards feature selection effectively, where users progressively eliminate vectors from the plots. Additionally, in contrast to other works in the literature, we argue that analysts should consider not only the lengths of the axis vectors, but also their orientations, and expert domain knowledge.

In particular, we have used SRA to carry out visual feature selection procedures with a real-world data set associated with medical chronic conditions of high prevalence in our society. Results show that SRA allows us to visualize groups of chronic patients with one or two chronic conditions (DM and/or HBP), while showing the contribution of different clinical features for discriminating among health statuses. These kinds of visualizations, which in principle are designed for performing exploratory data analyses, can be very valuable for experts in the clinical domain. In particular, the visual identification of drugs and diagnoses somehow related to chronic conditions may be of great value for a better understanding of these conditions, and may even reveal potential new relationships among diagnoses and drugs. Therefore, the method proposed in this work can be of great help to clinicians and health managers for planning care and health resources allocation. This could lead to an improvement of the health care system, both from an economical and social point of view.

Finally, as future research, we plan to work with time series data in order to find chronic patient trajectories. This could allow experts to identify the risk factors associated with the onset or evolution of a chronic condition. As a consequence, health managers could establish prevention programs according to the risk of a patient of suffering certain conditions.

Acknowledgements

This work has been partly funded by the Spanish Ministry of Economy (projects TIN2014-62143-EXP, TIN2015-70799-R, TEC2016-75361-R, TEC2016-75161-C2-1-4, and TIN2015-66731-C2-1-R) and the Institute of Health Carlos III (grant DTS17/00158).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.eswa.2018.01.054.

References

Alcala-Fdez, J., Sanchez, L., Garcia, S., del Jesus, M. J., Ventura, S., Garrell, J. M., et al. (2008). Keel: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing, 13*(3), 307–318. doi:10.1007/s00500-008-0323-y.

Amar, R., Eagan, J., & Stasko, J. (2005). Low-level components of analytic activity in information visualization. In *Proceedings of the proceedings of the 2005 IEEE symposium on information visualization* (pp. 15–21). Washington, DC, USA: IEEE Computer Society.

Averill, R. F., Goldfield, N., Eisenhandler, J., Muldoon, J. H., Hughes, J., Neff, J. M., et al. (1999). Development and evaluation of clinical risk groups (CRGs). *3M Health Information Systems*.

Berlinguet, M., Preyra, C., & Dean, S. (2005). *Comparing the value of three main diagnostic-based risk-adjustment systems (DBRAS)*. Canadian Health Services Research Foundation.

Bertini, E., Tatu, A., & Keim, D. (2011). Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics, 17*(12), 2203–2212. doi:10.1109/TVCG.2011.229.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*(1), 245–271.

Centers for Disease Control and Prevention (2011). International classification of diseases, ninth revision, clinical modification (ICD-9-CM). [Online] http://www.cdc.gov/nchs/icd/icd9cm.htm. Accessed Jan. 2018.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering, 40*(1), 16–28.

Chen, K., & Liu, L. (2004). VISTA: validating and refining clusters via visualization. *Information Visualization, 3*, 257–270.

Cox, T., & Cox, M. (1994). Multidimensional scaling. *Monographs on statistics and applied probability 88*. Chapman & Hall.

Dasarathy, B. V. (1991). *Nearest neighbor (NN) norms: NN pattern classification techniques*. Los Alamitos, CA: IEEE Computer Society Press.

Ding, C., & Li, T. (2007). Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the 24th international conference on machine learning* (pp. 521–528). New York, NY, USA: ACM. ICML'07.

Draper, G. M., Livnat, Y., & Riesenfeld, R. F. (2009). A survey of radial methods for information visualization. *IEEE Transactions on Visualization and Computer Graphics, 15*, 759–776.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. Wiley.

Fernández-Sánchez, J., Soguero-Ruiz, C., de Miguel-Bohoyo, P., Rivas-Flores, F. J., Gómez-Delgado, A., Gutiérrez-Expósito, F. J., & Mora-Jiménez, I. (2017). Clinical risk groups analysis for chronic hypertensive patients in terms of ICD9-cm diagnosis codes. In *Proceedings of the 4th international conference on physiological computing systems - volume 1: Phycs* (pp. 13–22). doi:10.5220/0006218700130022. INSTICC SciTePress.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika, 58*(3), 453–467.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3–42.

Goldberger, J., Roweis, S., Hinton, G., & Salakhutdinov, R. (2005). Neighborhood component analysis. In *Proceedings of the advances in neural information processing systems* (pp. 513–520). 17.

Gower, J., Gardner-Lubbe, S., & le Roux, N. (2011). *Understanding biplots*. John Wiley & Sons.

Guo, D. (2003). Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization, 2*(4), 232–246. doi:10.1057/palgrave.ivs.9500053.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1–3), 389–422. doi:10.1023/A:1012487302797.

Hughes, J. S., Averill, R. F., Eisenhandler, J., Goldfield, N. I., Muldoon, J., Neff, J. M., & Gay, J. C. (2004). Clinical risk groups (CRGs): A classification system for risk-adjusted capitation-based payment and health care management.. *Medical Care, 42*(1), 81–90.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). Independent component analysis. *Adaptive and learning systems for signal processing, communications, and control*. J. Wiley.

Ingram, S., Munzner, T., Irvine, V., Tory, M., Bergner, S., & Möller, T. (2010). Dimstiller: Workflows for dimensional analysis and reduction. In *Proceedings of the IEEE conference on visual analytics science and technology (VAST)* (pp. 3–10). IEEE Computer Society.

Inselberg, A., & Dimsdale, B. (1990). Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on visualization* (pp. 361–378). Los Alamitos, CA, USA: IEEE Computer Society Press. VIS'90.

Johansson, S., & Johansson, J. (2009). Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization & Computer Graphics, 15*, 993–1000.

Jolliffe, I. T. (2010). Principal component analysis. *Springer series in statistics*. Springer-Verlag.

Kandogan, E. (2000). Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In *Proceedings of the 6th IEEE information visualization symposium* (pp. 9–12). Salt Lake City, USA: IEEE Computer Society. Late Breaking Hot Topics.

Kandogan, E. (2001). Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of the 7th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 107–116). New York, NY, USA: ACM. KDD'01.

Krause, J., Perer, A., & Bertini, E. (2014). Infuse: Interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization & Computer Graphics, 20*(12), 1614–1623.

Leban, G., Zupan, B., Vidmar, G., & Bratko, I. (2006). Vizrank: Data visualization guided by machine learning. *Data Mining and Knowledge Discovery, 13*(2), 119–136.

Lichman, M. (2013). UCI machine learning repository.

Maaten, L. v. (2015). Matlab toolbox for dimensionality reduction.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability: 1* (pp. 281–297). University of California Press.

May, T., Bannach, A., Davey, J., Ruppert, T., & Kohlhammer, J. (2011). Guiding feature subset selection with an interactive visualization. In *Proceedings of the 2011 IEEE conference on visual analytics science and technology (VAST)* (pp. 111–120). doi:10.1109/VAST.2011.6102448.

McLachlan, G. J. (2004). Discriminant analysis and statistical pattern recognition. *Probability and mathematical statistics*. Wiley-Interscience.

McNicoll, G. (2002). World population ageing 1950–2050. *Population and Development Review, 28*(4), 814–816.

Norwegian Institute of Public Health (2017). *WHO Collaborating centre for drug statistics methodology, guidelines for ATC classification and DDD assignment 2018*. Oslo.

Paulovich, F. V., Nonato, L. G., Minghim, R., & Levkowitz, H. (2008). Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics, 14*(3), 564–575.

Rauber, P. E., Silva, R. R. O. d., Feringa, S., Celebi, M. E., Falcão, A. X., & Telea, A. C. (2015). Interactive image feature selection aided by dimensionality reduction. In *Proceedings of the EuroVis workshop on visual analytics (EuroVA)* (pp. 67–74). The Eurographics Association.

Rauber, T., & Steiger-Garção, A. (1993). Feature selection of categorical attributes based on contingency table analysis. In *Proceedings of the 5th Portuguese conference on pattern recognition*. Porto, Portugal: The Portuguese Association for Pattern Recognition.

Rubio-Sánchez, M., Raya, L., Díaz, F., & Sanchez, A. (2016). A comparative study between Radviz and star coordinates. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 619–628.

Rubio-Sánchez, M., & Sanchez, A. (2014). Axis calibration for improving data attribute estimation in star coordinates plots. *IEEE Transactions on Visualization and Computer Graphics, 20*(12), 2013–2022.

Rubio-Sánchez, M., Sanchez, A., & Lehmann, D. J. (2017). Adaptable radial axes plots for improved multivariate data visualization. *Computer Graphics Forum, 36*(3), 389–399. doi:10.1111/cgf.13196.

Seo, J., & Shneiderman, B. (2005). A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization, 4*(2), 96–113. doi:10.1057/palgrave.ivs.9500091.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE symposium on visual languages* (pp. 336–343). Washington, DC, USA: IEEE Computer Society.

Soguero-Ruiz, C., Hindberg, K., Mora-Jiménez, I., Rojo-Álvarez, J. L., & et al. (2016). Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *Journal of Biomedical Informatics, 61*, 87–96.

Sun, Y., Yuan, J., Hu, Y., & Xiao, W. (2008). An improved multivariate data visualization technique. In *International conference on information and automation (ICIA'08)* (pp. 1525–1530).

Tatu, A., Maaß, F., Färber, I., Bertini, E., Schreck, T., Seidl, T., & Keim, D. A. (2012). Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *Proceedings of the IEEE symposium on visual analytics science and technology* (pp. 63–72). IEEE Computer Society.

Tsai, C.-Y., & Chiu, C.-C. (2008). A clustering-oriented star coordinate translation method for reliable clustering parameterization. In *Proceedings of the 12th Pacific-Asia conference on advances in knowledge discovery and data mining* (pp. 749–758). Berlin, Heidelberg: Springer-Verlag. PAKDD'08.

Wang, Y., Li, J., Nie, F., Theisel, H., Gong, M., & Lehmann, D. J. (2017). Linear discriminative star coordinates for exploring class and cluster separation of high dimensional data. *Computer Graphics Forum (Proc. EuroVis)*.

Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research, 10*, 207–244.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

World Health Organization (1999). Hypertension guidelines. *Journal of Hypertension, 17*, 151–183.

World Health Organization (2006). Preventing chronic diseases. a vital investment: WHO global report. *International Journal of Epidemiology, 35*(4).

Yang, J., Peng, W., Ward, M. O., & Rundensteiner, E. A. (2003a). Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proceedings of the 9th annual IEEE conference on information visualization* (pp. 105–112). Washington, DC, USA: IEEE Computer Society. INFOVIS'03.

Yang, J., Ward, M. O., & Rundensteiner, E. A. (2002). Interring: An interactive tool for visually navigating and manipulating hierarchical structures. In *Proceedings of the IEEE symposium on information visualization (INFOVIS'02)* (pp. 77–84). IEEE Computer Society. INFOVIS'02.

Yang, J., Ward, M. O., Rundensteiner, E. A., & Huang, S. (2003b). Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proceedings of the symposium on data visualisation 2003* (pp. 19–28). Eurographics Association. VISSYM'03.

Yi, J. S., ah Kang, Y., Stasko, J., & Jacko, J. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics, 13*, 1224–1231.

Zhang, K.-B., Orgun, M., & Zhang, K. (2006). HOV^3: An approach to visual cluster analysis. In *Advanced data mining and applications: 4093* (pp. 316–327). Springer Berlin / Heidelberg. Lecture Notes in Computer Science.