Journal of Biomedical Informatics 46 (2013) 75–86

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

Is standard multivariate analysis sufficient in clinical and epidemiological studies?

Tânia F.G.G. Cova, Jorge L.G.F.S.C. Pereira, Alberto A.C.C. Pais *

Chemistry Department, University of Coimbra, Rua Larga, 3004-535 Coimbra, Portugal

ARTICLE INFO

Article history:
Received 10 May 2012
Accepted 14 September 2012
Available online 28 September 2012

Keywords:
Cancer diagnosis
Cancer incidence
Principal component analysis
Hierarchical cluster analysis

ABSTRACT

Clinical tests and epidemiological studies often produce large amounts of data, being multivariate in nature. The respective analysis is, in most cases, of importance comparable to the clinical and sampling tasks. Simple, easily interpretable techniques from chemometrics provide most of the ingredients to carry out this analysis. We have selected available data from different sources pertaining to cancer diagnosis and incidence: (1) cytological diagnosis of breast cancer, (2) classification of breast tissues through parameters obtained from impedance spectra and (3) distribution of new cancer cases in the United States. Hierarchical cluster analysis (HCA) is needed especially in cases where there is no a priori identification of classes, suggesting a structure of the data based on clusters. These clusters or the classes, are then further detailed and rationalized by principal component analysis (PCA). Partial least squares (PLS) and linear discriminant analysis (LDA) provide further insight into the systems. An additional step for understanding the data set is the removal of less characteristic data (NR) using a density-based approach, so as to make it more clearly defined. Results clearly reveal that breast cytology diagnosis relies on variables conveying mostly the same type of information, being thus interchangeable in nature. In the study on tissue characterization by electrical measurements, the distribution of the different types of tissues can be easily constructed. Finally, the distribution of new cancer cases possesses clear, easily unravelled, geographical patterns.

1. Introduction

Due to the profusion of analytical techniques and diagnosis tools building massive sets of data, multivariate data analysis cannot be dissociated from most of the problems found in health studies. The automatic diagnosis of breast cancer is an important, real-world medical problem. Breast cancer is the most common type of cancer in women and there are some studies indicating that the incidence rate of new cases is increasing [1–3]. The diagnosis of breast cancer is essentially based on self-examination, clinical examination, mammography, breast ultrasound (in conjunction with a mammogram) and core biopsy [4–10].

In this work, an automated data analysis procedure based on chemometrics algorithms is presented. This procedure is carried out by employing widely used data sets available from literature. As examples, we select three very different studies on human cancer. Specifically, we combine hierarchical cluster analysis (HCA), principal component analysis (PCA), linear discriminant analysis (LDA), partial least squares (PLS) and an outlier removal approach to understand what can be extracted from these different studies, from exploratory data analysis to variable selection. We believe that a set of standard techniques, duly combined, is in many cases able to replace more complex approaches. HCA is mostly useful for situations in which there is little a priori knowledge on the data structure. However, the dendrogram, in spite of the useful information that conveys, does not show the relative positioning of the groups. For that, PCA allows the two- or three-dimensional representation of the data, and the clusters extracted from HCA can be accurately positioned. Additionally, inspection of the loadings permits a detailed understanding of the reasons behind the formation of the clusters. It is interesting to notice that the synergistic power of these two techniques is often overlooked. The graphical representation of the clusters, which may benefit from the convex-hull technique for establishing adequate borders, also enlightens on the possibility of setting a decision rule if the clusters are turned into classes. In some cases, simple inspection of the data discards the need for a, yet simple, classification technique, such as linear discriminant analysis. The same applies for the use of response evaluation methods, such as partial least squares. In some cases, the dependence of the data on the original variables is so comprehensively interpreted resorting to PCA (or the HCA/PCA combination), that PLS becomes little more than a confirmatory tool.

2. General procedure

The algorithms used in this work were implemented and optimized by the authors using GNU Octave language (version 3.2.4)

* Corresponding author. Fax: +351 239827703.
E-mail address: pais@qui.uc.pt (A.A.C.C. Pais).

1532-0464/$ - see front matter © 2012 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.jbi.2012.09.005

[11]. The respective graphical representations were obtained using Gnuplot (version 4.4). We propose an automated approach consisting in four main steps: (i) description of the data structure based on HCA, (ii) data overview and variable selection by PCA, (iii) resolution of the different classes using LDA and (iv) definition of relationships between the predicted and observable variables based on PLS. An additional step that can help in characterizing the data set is the removal of less characteristic data (NR). Its implementation depends on the problem under study. Note that, in some cases, only a selected part of these steps is used.

2.1. Clustering process

Clustering algorithms are directed to divide data into groups of similar objects using an unsupervised learning method. Many types of unsupervised clustering techniques exist such as partitional, hierarchical, density-based or grid-based with a number of related clustering algorithms [12]. Hierarchical methods are extremely common, and one of the reasons is that they allow the visualization of the data structure, even in complex cases. They proceed by a successive association (or dissociation) of the objects in the data, leading to a final output which consists of a cluster sequence, represented via a dendrogram [13,14]. In this structure, each level of association corresponds to the partitioning of the data set into a specific number of clusters. It is possible to additionally predict the number of clusters, on the basis of the dendrogram, but this task relies more often on common sense than on a definite criterion [15,16]. In HCA there are two important choices when defining a method: the type of similarity measure between objects and/or groups, and the linkage technique [17]. The first task is to determine a numerical value for the similarity between objects, constructing a similarity matrix. The most popular ways to determine the similarity between objects use the Euclidean distance and the correlation coefficient, but there are many alternatives for similarity indicators [17]. The next step is to group or ungroup the objects. A common approach is an agglomerative technique, whereby single objects are gradually connected to each other in groups. The first connection corresponds necessarily to the most similar pair of objects. Once the first group is formed, it is necessary to define the similarity between the new group and the remaining objects. This step requires a new choice among a variety of available techniques. Some of the most used linkage algorithms are single-linkage, complete linkage, average-linkage and Ward's linkage [17,18]. In this work, Ward's linkage is the underlying technique. It finds at each stage those two clusters, C_A and C_B with sizes n_A and n_B, which, after merging, promote the minimum increase in the total within group error sum of squares, i.e., the minimum distance, $d_{A,B}$, between the centroids, μ_A and μ_B, of the merged clusters

$$d_{A,B} = \frac{n_A n_B}{n_A + n_B}(\mu_A - \mu_B)'(\mu_A - \mu_B) \tag{1}$$

In hierarchical clustering, the value of the within group sum of squares starts at zero, because every point is in its own cluster, and then grows as clusters are merged. In summary, Ward's method keeps this growth as small as possible.

Once the similarity measure and the linkage method are defined, the agglomeration of objects and groups in each step of the process follows the order of larger similarity.

2.2. Dimensionality reduction process

Principal component analysis is probably the oldest and best known of the techniques of multivariate analysis [19]. PCA allows compressing the data by reducing the number of dimensions, with a minimized loss of information. The most influential variables in the system are highlighted, and underlying factors may be identified. This analysis is based on the assumption that most of the information about the structure of the data is contained in the directions along which the variations are the largest.

The procedure is carried out by a linear transformation of the m original variables, $\mathbf{x_1} \ldots \mathbf{x_m}$, into a new set, the principal components, $\mathbf{u_1} \ldots \mathbf{u_p}$. The ith principal component is given by

$$\mathbf{u_i} = w_{i1}\mathbf{x}_1 + w_{i1}\mathbf{x}_2 + \ldots + w_{im}\mathbf{x}_m \tag{2}$$

where $w_{i1} \ldots w_{im}$ are the loadings, i.e. the weights of the original variables on the linear combination [20]. The principal components are not correlated with each other and altogether explain the total variance of the data. The transformation matrix $\mathbf{W}$ whose elements are the loadings w_{ij} and the vector $\boldsymbol{\lambda}$, whose components correspond to the recovered variance λ_i in each ith principal component, can be obtained via a singular value decomposition

$$\mathbf{C_x W} = \boldsymbol{\lambda}\mathbf{W} \tag{3}$$

where $\mathbf{C_x}$ corresponds to the variance/covariance matrix of the original data. Also, $\sum_i^m \lambda_i$ gives the total variance of the data. Frequently, $\mathbf{C_x}$ is replaced by the correlation matrix $\boldsymbol{\rho}$, in a normalized approach. In this case, $\sum_i^m \lambda_i = m$.

A major argument for using correlation matrices, rather than covariance matrices, to define the principal components is that the results of analysis for different sets are more directly comparable than those of the analysis based on covariance [19]. The big drawback of PCA based on covariance matrices is the sensitivity of the PCs to the order of magnitude of the elements of $\mathbf{X}$, or more specifically, of the respective variance (these two being usually related). However, it may be justified in situations in which larger variables are, for some reason, more relevant or signal sensitivity is important. In either case, these components are ranked, and the percentage of explained variance λ_i decreases from the first PC to the second and so on. [21,22], suggesting the criteria for the selection of the most relevant first p principal components. The most common one is Pearson's criterion, which can be used in conjunction with both the variance/covariance matrix and the correlation matrix [19]. The value p is selected as the minimum integer that warrants

$$\sum_{i=1}^{p}\lambda_i \Big/ \sum_{i=1}^{m}\lambda_i \geqslant 0.8 \tag{4}$$

If the correlation matrix is used, the most common criteria corresponds to retain the p components for which $\lambda_i \geqslant 1$, although other values have been suggested [19,23].

The definition and computation of principal components are thus straightforward and this apparently simple technique has proved extremely useful in a wide variety of different applications. It provides a very useful exploratory tool to uncover unknown trends in the data [24,25]. Note that preliminary data inspection is especially important when establishing medical diagnosis procedures. In this context, PCA should be the first choice to look for patterns, aggregates, trends and outliers in the data under scrutiny [26,27] and it has been used for identifying significant variables (which are related to the main principal components) and distinguishing patients from healthy subjects in oncologic diseases [28], as an example among various others [29–41].

2.3. Classification process

The classification techniques rely on training or learning data sets, in which the objects are previously divided into classes, which implies the use of external information, turning them into

supervised methods. These data sets allow establishing what is called a decision rule, subsequently used for selecting the class to which a new object belongs. The classical methods for the supervised classification are correlation based methods, distance based methods, linear discriminant analysis (LDA), soft independent modelling by class analogy (SIMCA), and partial least squares discriminant analysis (PLS-DA) [42,43]. The accuracy of LDA and other classification methods such as quadratic discriminant analysis (QDA) and *k*-nearest-neighbour (KNN) methods has been assessed in different clinical studies [44]. In the present study, we focus solely on the LDA approach. LDA is a linear parametric method with discriminating characteristics [28,45]. It focuses on finding the optimal boundaries between classes, by selecting the directions that achieve a maximum separation among the different classes [46]. In other words, it finds the vectors in the variables space that best discriminate among classes. More formally, given a number of independent variables relative to which the data is described, LDA creates a linear combination which yields the largest mean differences between the desired classes. For this, two matrices are defined: the between-class scatter matrix and the within-class scatter matrix. For all samples of all classes, the between-class scatter matrix $\mathbf{C_B}$ and the within-class scatter matrix $\mathbf{C_W}$ are defined by

$$\mathbf{C_B} = \sum_{i=1}^{c} M_i \cdot (\mu_\mathbf{i} - \mu) \cdot (\mu_\mathbf{i} - \mu)^T \tag{5}$$

$$\mathbf{C_W} = \sum_{i=1}^{c} \sum_{x_k \in \mathbf{X_i}} (x_k - \mu_\mathbf{i}) \cdot (x_k - \mu_\mathbf{i})^T \tag{6}$$

where M_i is the number of objects in class i, c is the number of distinct classes, μ_i is the mean vector of samples belonging to class i and $\mathbf{X_i}$ represents the set of samples belonging to class i with x_k being the kth variable of that class. $\mathbf{C_W}$ represents the scatter of objects around the mean of each class and $\mathbf{C_B}$ represents the scatter of objects around the overall mean for all classes. The goal is to maximize $\mathbf{C_B}$ while minimizing $\mathbf{C_W}$, in other words, maximize the ratio $\det|\mathbf{C_B}|/\det|\mathbf{C_W}|$. To classify an unknown object, its coordinates are projected along a line, derived from the decision rule, and it is assigned to the group with the nearest center of mass.

2.4. Prediction process

The partial least squares (PLS) method, also known as projection on latent structures, is a recent technique that combines features from principal component analysis and multiple linear regression, which are generalized. It pertains to a wide class of methods for modeling relations between sets of observed variables by means of latent variables [47–49]. Furthermore, it comprises classification tasks as well as dimension reduction techniques [49].

The underlying assumption of all PLS methods is that the observed data is generated by a system or process which is driven by a small number of latent (not directly observed or measured) variables. In its general form, PLS creates orthogonal score vectors (also called latent vectors or components) by maximizing the covariance between different sets of variables. The data are divided into two blocks, one block containing the predictor variables and the other containing the response variables. PLS models the relation between these two data sets (blocks of variables). Denote by $\mathbf{X} \in R^N$ a N-dimensional space of variables representing the first block and similarly by $\mathbf{Y} \in R^M$ a space representing the second block of variables. The relations between these two blocks are given by means of score vectors. After observing n data samples from each block of variables, PLS decomposes the $(n \times N)$ matrix of zero-mean variables $\mathbf{X}$ and the $(n \times M)$ matrix of zero-mean variables $\mathbf{Y}$ into the form

$$\mathbf{X} = \mathbf{TP^T} + \mathbf{E} \tag{7}$$

$$\mathbf{Y} = \mathbf{UQ^T} + \mathbf{F} \tag{8}$$

where, by analogy with PCA, the $\mathbf{T}(n \times p), \mathbf{U}(n \times p)$ are matrices of the p extracted score vectors (components, latent vectors), and the $\mathbf{P}(N \times p), \mathbf{Q}(M \times p)$ are matrices of loadings. The $\mathbf{E}(n \times N)$, $\mathbf{F}(n \times M)$ are matrices of residuals [50]. Both equations (Eqs. (7) and (8)) represent outer relations in predictor and response subspaces, respectively [50]. The PLS model equation, also known as inner relation, can be defined by looking at the projections of the $\mathbf{Y}$ block scores, $\mathbf{u}(1 \times n)$, against the $\mathbf{X}$ block scores, $\mathbf{t}(1 \times n)$, for every PLS latent factor,

$$\mathbf{u} = \beta \mathbf{t} \tag{9}$$

where $\boldsymbol{\beta}$ are the PLS model sensitivity coefficients, which reflect the relevance of the respective latent factor.

2.5. Removal of outliers

The benefits of additionally introducing a noise/outlier reduction filter (NR) will also be assessed in the present work [15]. The combination of the NR with some of the standard methods may be used in situations in which the data contains some amount of outliers or less characteristic data. Noise makes the data structure less defined and the number of groups that are formed may be either too low or too high, depending on the actual situation and algorithm. Removing what potentially are less characteristic data provides a more defined perspective of the system, and of the underlying patterns. Note that we are not pursuing a more precise description removing these points, we are simply trying to obtain an alternate description, based on the core data. Let us, for example, assume that there is a situation in which two classes overlap. If the noise removal procedure focus mainly on this overlap region, its importance is probably lower for establishing a decision rule, and a higher degree of failure is acceptable here.

This identification and removal of outliers is suggested via a self-consistent technique in which system properties are used to make an automatic specification of the necessary parameters [15]. It is a density-based approach, that removes points found in lower density regions. It can be summarized by the following pseudo-code:

---

cycle for all current objects i
cycle for $j <> i$
$d_{ij} = distance(i,j)$;
cycle for all current objects i
$d_{nn,i} = min(d_{ij}$, *for all* $j <> i)$;
calculate $\bar{d}_{nn} = average(d_{nn,i}$, *for all current objects i)*;
  cycle for all current objects i
$n_{3d,i} = count(d_{ij} < 3\bar{d}_{nn}$, *for all* $j <> i)$;
calculate $\bar{n}_{3d} = average(n_{3d,i}$ *for all current objects i);*
cycle for all current objects i
if $n_{3d,i} < (1/3)\bar{n}_{3d}$, *discard object i;*
$n_{disc} = count(n_{3d,i} < (1/3)\bar{n}_{3d}$, *for all current objects i)*
repeat from top until $n_{disc} = 0$

---

in which d_{ij} is the distance between objects i and j, $d_{nn,i}$ is the nearest neighbour distance for object i, $n_{3d,i}$ the number of objects found around object i within a radius $3\bar{d}_{nn}$, and n_{disc} the number of discarded objects in each iteration.

The identification of outliers can be further summarized in four main steps: (1) calculate the average nearest-neighbour distance, $\bar{d}_{nn}$, (2) determine the number of objects, n_{3d}, around each object, (3) discard all objects for which $n_{3d} < \frac{1}{3\bar{n}_{3d}}$ (4) repeat from (1),

without the discarded objects, until the number of newly discarded ones equals zero.

2.6. Graphical representation

An efficient way to visualize the formed groups in 2D is achieved through the convex hull representation. Computing the convex hull means that a non-ambiguous and efficient representation of the required convex shape is constructed. The complexity of the corresponding algorithms is usually estimated in terms of the number of input points, and the number of points on the convex hull [51–53]. The convex hull of a set **Q** of points in the plane is the smallest convex polygon that surrounds them. Any geometric figure is called convex if it includes all line segments that join the points [54]. Thus, all points of **Q** must be within the polygon or on its boundary.

3. Databases

The data analysis proposed in this work is used in (1) the Wisconsin breast cancer determination problem [55], (2) the breast tissue classification problem [56], and (3) a study of a different type, that focuses on the incidence of new cancer cases for some selected cancer types by US state [57].

The data source of the first two cases under study is the University of California at Irvine (UCI) Machine Learning Repository [58,59]. The data pertaining to the third case were obtained from the American Cancer Society, Surveillance and Health Policy Research [57].

The first example is a study conducted on 699 subjects, with nine attributes in a two-class data set. It refers to the breast cancer diagnosis based on physiological microscopic observations of cells, including the extent to which epithelial cell aggregates are mono- or multilayered (CT), uniformity of cell size (UCSz), uniformity of cell shape (UCSp), cohesion of the peripheral cells of the epithelial cell aggregates (MA), the diameter of the population of the largest epithelial cells relative to erythrocyte (SECS), the proportion of single epithelial nuclei that were devoid of surrounding cytoplasm (BN), blandness of nuclear chromatin (BC), normal nucleoli (NN) and frequency of mitosis (MIT), for details see [55,60]. Inspection of the data set revealed that 16 patients displayed missing values. These cases were removed, leading to a total of 683 patients, being 239 corresponding to malignant and 444 to benignant cases. This database will be denoted as *breast cancer I* in what follows.

The second example concerns the classification of breast cancer tissues, resorting to electrical measurements. Details on the data collection procedure as well as classification of the cases and frequencies used are given in references [1,61,62]. The data set contains information data on 106 impedivity spectra collected from breast tissue samples of 64 patients undergoing breast surgery. Six groups of tissues were defined before the experiments, according to the pathology and morphology of the breast, both of which will be used in our analysis. These six groups are partitioned into normal and pathological. The group of normal breast tissues is formed by the glandular (denoted GLD, 16 cases), connective (denoted CNN, 14 cases) and adipose tissues (denoted ADI, 22 cases). In the pathological group there are included the carcinoma (denoted CAR, 21 cases), fibro-adenoma (denoted FAD, 15 cases) and mastopathy (denoted MAS, 18 cases). The study relies on electrical impedance measurements, with nine variables used as predictors: impedivity at zero frequency (I0), phase angle at 500 kHz (PA500), high-frequency slope of phase angle (HFS), impedance distance between spectral ends (ID), area under spectrum (AS), area normalized by ID (AN), maximum of the spectrum (MAX), distance between impedivity and real part of the maximum frequency point (DIR) and length of the spectral curve (LS). This database will be denoted as *breast cancer II* in what follows.

The data used in the third study are the estimated numbers of new cancer cases in 2010 [57], for 11 selected cancer types presented by residents of all US states. The selected cancer types are female breast cancer (FBC), uterine cervix (UTCx), colon and rectum (CR), uterine corpus (UTCs), leukaemia (LEU), lung and bronchus (LBR), melanoma of the skin (MS), non-Hodgkin lymphoma (NHL), prostate (PR) and urinary bladder (UBL). The 11th variable contains other cases (OT), obtained from the difference to the total reported new cases.

4. Results and discussion

The procedure described in Section 2 was applied to all three different systems. For clarity and simplicity, each case will be here analyzed separately. However, general conclusions will be drawn from the whole set.

4.1. Breast cancer I

Table 1 contains the main results pertaining to a direct PCA analysis of the data set variables, using the correlation matrix. It is seen that the first two and three principal components are able to recover ca. 74.2% and 80.2%, respectively, of the data variability, indicating that a graphical representation based on these two or three components is clearly meaningful. Kaiser's criterion [19] suggests that only one component would be sufficient for a correct description of the data. However, the question arises if other components may convey some additional, relevant, information. This is to be checked below.

Fig. 1 presents the breast cancer cases in the PC_1/PC_2 plane. From this figure it is apparent that malignant cases are related to a large diffuse data cloud, while benign cases correspond to a smaller and very dense group. This figure further explores the PCA results, displaying the representation, on the basis of frequency histograms of the projection of the objects on each two first principal components. It shows that the separation of the two groups is possible along the first component (PC_1), but not in the second since there is a marked overlap of these groups. This observation leads us to conclude that PC_1 also contains the most relevant information for discrimination.

Table 2 presents the dependency of the PCA transformed data on each original variable, for the first two principal components. Since PC constitutes an orthonormal vectorial base, for a m dimensional case we expect an average loading value in each component of $1/\sqrt{m}$. The criterion for selecting a significant load (in bold in the table) is based on the comparison to the average expected value, i.e., the loading is simply considered significant if above the average value, and not significant if otherwise.

Table 1
Eigenvalues and data recovery evolution with respect to the number of principal components for the original data set ($N = 683$).

Principal components	Eigenvalues (λ_i)	Explained variance (%)	Cumulative explained variance (%)
PC_1	5.90	65.5	65.5
PC_2	0.78	8.6	74.2
PC_3	0.54	6.0	80.2
PC_4	0.46	5.1	85.3
PC_5	0.38	4.2	89.5
PC_6	0.30	3.4	92.8
PC_7	0.29	3.3	96.1
PC_8	0.26	2.9	99.0
PC_9	0.09	1.0	100

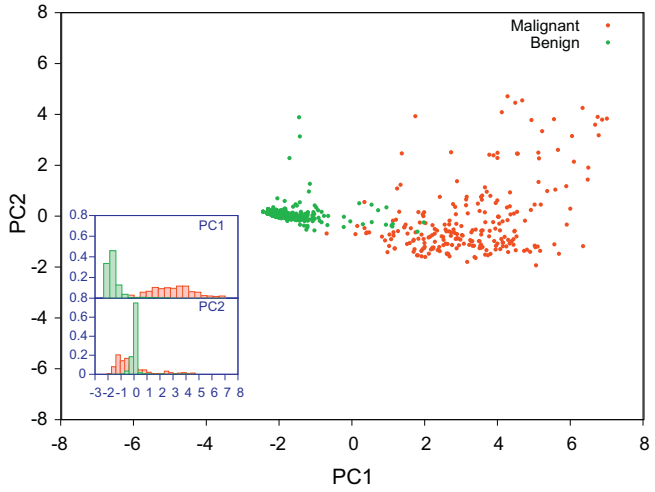

Fig. 1. Representation of the *breast cancer I* cases on the PC_1/PC_2 plane. The normalized frequency histogram for the projection of the objects upon these principal components is shown in the inset. Green refers to benign cases, while red corresponds to malignant ones. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

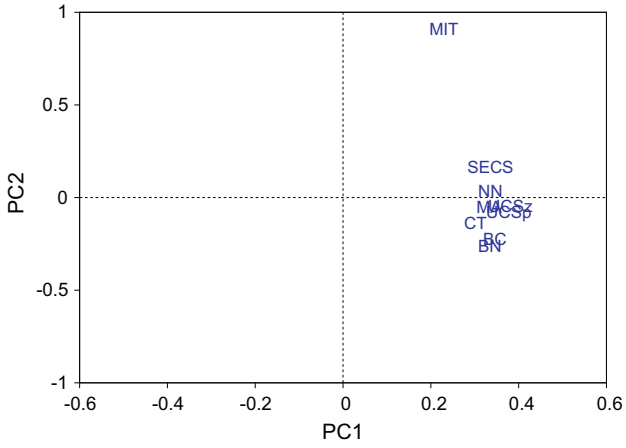

Fig. 2. Representation of the cytological observations on the first two components.

Table 2
Loading values of all original variables on the first two principal components. The most relevant contributions are highlighted in bold.

	PC_1	PC_2
CT	0.302	−0.141
UCSz	**0.381**	−0.047
UCSp	**0.378**	−0.082
MA	**0.333**	−0.052
SECS	**0.336**	0.164
BN	**0.335**	−0.261
BC	**0.346**	−0.228
NN	**0.336**	0.034
MIT	0.230	**0.906**

Fig. 2 depicts the contribution of each variable on the first two components (PC_1 and PC_2). In this figure are visible the variables responsible for the discrimination of the benign and malignant cases presented in Fig. 1. Thus, the first component retains essentially information from UCSz, UCSp, MA, SECS, BN, BC and NN variables. Further conclusions can be drawn from the fact that the values of this first component are all positive, suggesting that it represents, at least partially, a measure of the degree of malignancy.

Correlation coefficients indicate that variables UCSz, UCSp, SECS, BN, BC and NN are related, being the highest correlation between variables UCSz and UCSp (0.907) and the lowest between UCSp and BN (0.714). The second component is mostly related with MIT, which is not directly correlated with the other variables. It displays a highest correlation value with SECS, still a small value (0.481). From a biological point of view, the number of mitoses expresses the activity of cell division. This means that the higher proliferative activity of the tissue, the larger the number of mitoses observed. In benign tumours, mitoses are rare and have a typical appearance, whereas in the malignancies, they are more numerous and atypical. An uncontrolled variation of the shape and size of the cells (UCSp and UCSz respectively) coupled with a high number of mitoses (MIT) is the worst scenario. The discrimination between benign and malignant is found along PC_1, but PC_2 also includes information on the degree of malignancy. In other words, malignancy increases along a direction that comprises both shape and size and mitosis indicators.

In data sets obtained experimentally, the presence of either outliers or less representative objects is a common occurrence. To illustrate the impact of using a filtering algorithm in group boundaries, we consider three different situations depicted in the convex hull representation of Fig. 3. Panel (a) is the representation of the original raw data, and the remaining panels show the impact of NR [15] in group boundaries. This is achieved in two ways, either performing NR on the raw data, panel (b), or on PCA scores, panel (c). In all cases, the groups are treated one by one. It is clear that the group of benign samples was drastically reduced (45.0% removal), leaving a denser zone essentially untouched – this reflects some unhomogeneity in the respective density. The boundaries uncompassing the malignant samples are not as affected (less than 1% removal) by the noise reduction procedure. This class is disperse, but displays an homogeneous density. Panel (c) reveals that the impact of this procedure is higher when applied on the lower dimensionality PCA scores, with 47.3% and 4.6% removal, respectively for the benign and malignant classes. Note, additionally, that most of the discarded points are situated in the class overlap region.

Based on these observations, and recalling that each variable may only take a few values, an effort is now made on the possibility of imposing decision rules directly over each variable, i.e., to introduce a threshold value to distinguish malignant and benign cases. Table 3 contains the degree of misclassification obtained, taking into account the original diagnosis.

As expected from Table 3, the variables present different abilities to be used solely for breast cancer diagnosis, as evaluated from the minimum misclassification levels attained in each case. From Table 3 we can see that, for this most adequate selection criterion, variables UCSz (uniformity of cell size) and UCSp (uniformity of cell shape) are able to predict response with misclassification errors of ca. 7% and 8%, respectively. It should be noted that variable MIT, associated with the degree of mitosis, displays the highest classification error (28.3%). It provides information complementary to that of the other variables, as extracted from the correlation data, and is probably associated to the degree of malignancy, as suggested before.

In conclusion, and concerning class discrimination, the information obtained from PCA is sufficient to pinpoint the most important variables, and establish a single-variable (or a very low dimension) decision rule. Even a low complexity additional technique (such as LDA) is somewhat superfluous in this scenery.

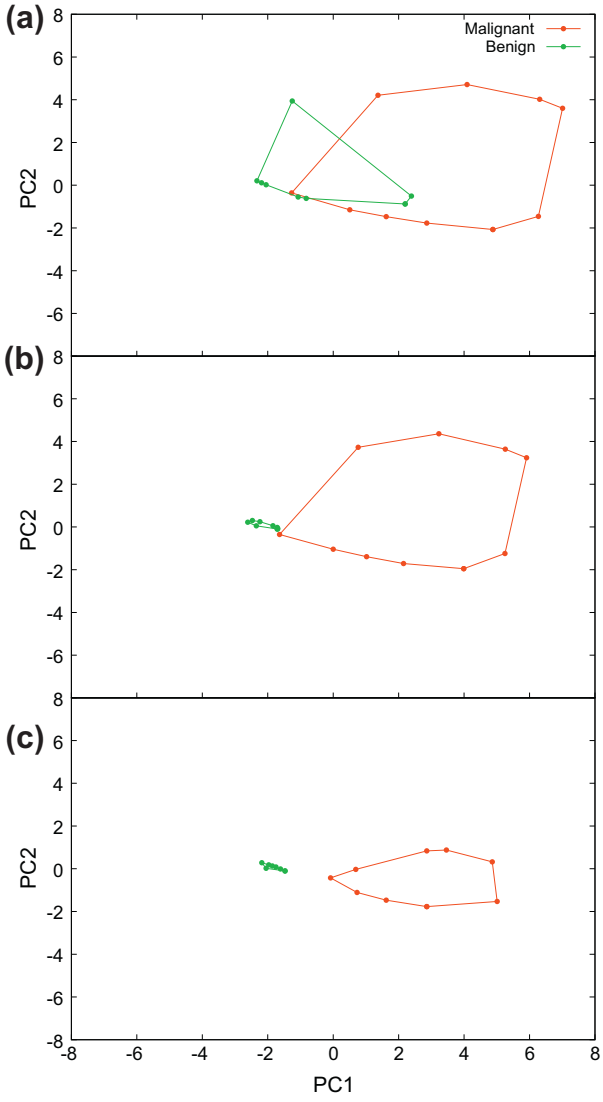

Fig. 3. Representation of the two groups (benign and malignant cases) in 2D convex hull form, showing the impact of NR in the group boundaries: before filtering (a), after performing NR on the raw data, for each group separately (b) and after performing NR over PCA scores, for each group separately (c). The green-coloured group refers to the benign cases, while the red-coloured one to the malignant cases. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Patients misclassification (%) with respect to the medical decision by directly imposing decision rules based on the threshold value ($T : X_i \leqslant T$ benignant, $X_i > T$, malignant) directly on each variable score (X_i).

Threshold value (T)	CT	UCSz	UCSp	MA	SECS	BN	BC	NN	MIT
<1	65.0	65.0	65.0	65.0	65.0	65.0	65.0	65.0	65.0
1	45.8	11.9	15.2	16.5	58.9	10.8	43.6	13.8	21.2
2	39.7	7.6	8.8	14.2	10.2	**9.1**	22.5	**10.2**	24.0
3	28.0	**7.3**	**7.8**	**13.3**	12.4	**9.1**	**9.5**	13.8	**28.3**
4	20.2	10.5	10.5	16.7	16.8	10.1	13.2	16.1	30.0
5	**14.6**	14.6	14.6	18.9	21.1	11.9	17.0	18.3	30.6
6	15.1	18.3	18.3	21.1	26.9	12.4	17.7	20.4	31.0
7	17.9	20.8	21.8	23.0	28.0	13.3	26.6	22.1	32.1
8	23.1	24.6	25.5	26.6	30.5	15.8	30.5	24.6	32.9
9	25.2	25.2	26.5	27.2	30.7	17.1	32.1	26.2	32.9
10	35.0	35.0	35.0	35.0	35.0	35.0	35.0	35.0	35.0

Table 4
PLS results, accounting for the 9 predictors and the description of the response, i.e., the diagnosis.

f[a]	V_X%[b]	UV_X%[c]	TV_X%[d]	V_Y%[b]	UV_Y%[c]	TV_Y%[d]	β[e]
0	100.00	0.00	0.00	100.00	0.00	0.00	0.00
1	34.48	**65.52**	65.52	18.60	**81.40**	**81.40**	**23.56**
2	26.89	**7.58**	73.11	16.03	**2.58**	**83.97**	**4.19**
3	21.57	5.32	78.43	15.70	0.32	84.30	**1.48**
4	16.86	4.72	83.14	15.67	0.03	84.30	0.45
5	13.38	3.48	86.62	15.67	0.00	84.30	0.18
6	10.43	2.95	89.57	15.67	0.00	84.30	0.13
7	7.31	3.12	92.69	15.67	0.00	84.30	0.02
8	3.34	3.97	96.66	15.67	0.00	84.30	0.00
9	0.00	3.34	100.00	15.67	0.00	84.30	0.00

[a] Number of latent factors.
[b] Information contained on the predictors and response subspaces, respectively (residual sum of squares).
[c] Information used for a given latent factor, f.
[d] Cumulative amount of information used for a given latent factor f.
[e] PLS model coefficients.

Let us now consider a more sophisticated approach based on PLS, to select only a few underlying or latent factors that account for most of the variation in the response, i.e., in the diagnosis. Thus, the **X** block consists in all the nine predictors related to the cytological observations. Standard PLS is performed after variable normalization. The respective results may be contrasted with those obtained by PCA. Table 4 presents the response data description, related to the diagnosis, taking into account all the 9 predictors. The first latent factor (LF1) uses 65.5% of the information on the predictors sub-space to describe 81.4% of the response, leading to an efficiency of 1.24. This efficiency index is in fact proportional to the PLS parameters, computed in the inner relation, Eq. (9), and may be used as a measure of the sensitivity with which the response is described. In order to identify the most relevant latent factors, we consider that efficiencies over than 1 may be used as an identifying criterion. The second latent factor (LF2) requires more 7.6% of the predictors information to justify only 2.6% of additional information on the response sub-space, leading to an efficiency of 0.34, much lower than LF1. Using these two latent factors ca. 84.0% of the response is described. Further efforts to describe the response are not successful, since only 84.3% of the total response information can be explained. Similar results can be obtained by direct inspection of the model coefficients, β. The first three values (23.56, 4.19 and 1.48) reveal the dominance of LF1 over both LF2 and LF3 (LF1 $\gg$ LF2 > LF3). However, this evaluating process is less conclusive since there is not a specific criterion to follow.

Table 5 presents the loadings for the two main latent factors, being the first one much more relevant than the second. Considering a criterion similar to that used in PCA with respect to the identification of the most significant loadings, the first latent factor (LF_1) retains essentially information about UCSz, UCSp, BC, BN,

Table 5
Predictor sub-space loadings obtained in the first two latent factors, LF_1 and LF_2.

Predictor	LF_1	LF_2
CT	0.305	0.372
UCSz	0.381	−0.065
UCSp	0.378	0.006
MA	0.332	−0.079
SECS	0.334	−0.300
BN	0.339	0.465
BC	0.347	0.085
NN	0.335	−0.135
MIT	0.225	−0.720

Table 6
PCA results for the first three components, considering the correlation approach.

Principal components	Eigenvalues (λ_i)	Explained variance (%)	Cumulative explained variance (%)
PC_1	5.46	60.7	60.7
PC_2	1.81	20.1	80.8
PC_3	0.78	8.61	89.5

NN and SECS while the second latent factor (LF_2) is based mostly in MIT, which is in accordance with the previous PCA results.

4.2. Breast cancer II

The previous example has shown that PCA replaces most of EDA (Exploratory Data Analysis) tasks, and provides a good data visualization. As such, we proceed directly to this technique and the results are summarized in Table 6. Using the correlation matrix, at least two components are required.

Considering the eigenvalues reported in that table, the first two values (5.46 and 1.81) represent ca. 80.8% of the data variability described with the first two principal components. Other subsequent eigenvalues, such as 0.78, are much less significant, representing only 8.6% of variability recovery. Again, this means that a graphical representation based on the first two components clearly reflects the structure of the data.

Fig. 4 presents a data overview pertaining to each tissue type in the new PCA system. There is a severe group overlap, and a very dense region where at least four groups are indistinguishable. This fact indicates that one may face a difficult task in finding a correct diagnosis. However, it should be noted that, in what concerns the separation between normal and pathological tissues, data pertaining to the glandular tissue are the only that overlap with that of pathological cases.

In order to retrieve some relevant information for the most discriminant variables, Table 7 presents the impact of each variable, in the first two components. As for the previous case, the criterion for selecting a significant loading is based on the comparison to the

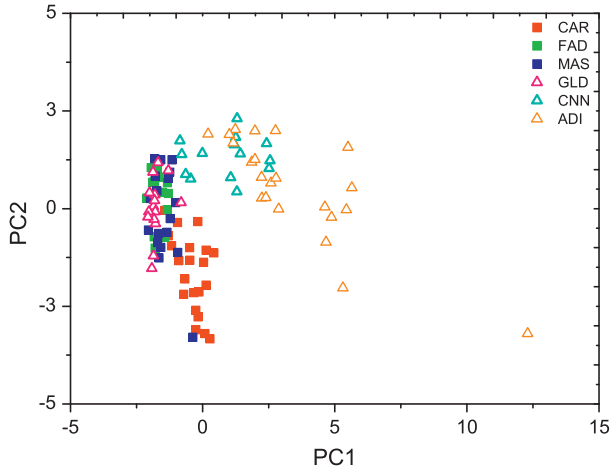

Fig. 4. Representation of the breast tissues on the two main principal components. Colours and symbols are related to each tissue type. The red-coloured group refers to the Carcinoma (CAR), the green-coloured group refers to the Fibro-adenoma (FAD), the blue-coloured one to the Mastopathy (MAS). The remaining three groups, related to the Glandular (GLD), Connective (CNN) and Adipose (ADI) tissues correspond to the tissues with normal characteristics. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 7
Loading values of all variables, on the first two principal components (correlation approach).

	PC_1	PC_2
I0	0.387	0.240
PA500	−0.047	−0.665
HFS	0.094	−0.586
ID	0.395	−0.059
AS	0.352	−0.167
AN	0.355	−0.276
MAX	0.392	−0.098
DIR	0.358	0.092
LS	0.389	0.179

average value, as explained before. The first component retains mainly information over ID, MAX, LS, I0, DIR, AN and AS variables which are, in some cases, inter-related. The second component is mostly related with PA500 and HFS, which are also somewhat inter-related, with a correlation coefficient of 0.509. Selecting the significant loads directly by the loading value criterion, there are seven variables with relevant information in PC_1 (ID, MAX, LS, I0, DIR, AN and AS) and two variables in PC_2 (PA500 and HSF). This case was further studied in order to evidence the most significant variables for data description. Specifically, we use two methods, described in reference [19], for selecting the set of original variables that are more influential in the data structure.

Briefly speaking, the first method starts by removing the variable with the highest loading in absolute value associated with the less significant component. This is the least relevant variable. A new PCA is performed on the remaining variables, and a new variable is removed following the same criterion. This is the second least important variable. The procedure is repeated until all variables are ranked in order of ascending importance. In the second method, variables with loadings exceeding $1/\sqrt{m}$ associated with successive principal components are preserved, while variables with loadings inferior to this value are discarded (this is a slightly modified version of that found in [19]). In this case, the most influential variables are those selected in the first component, with the respective importance decreasing in order of decreasing absolute value of loading, then those selected in the second component, and so on. Naturally, after ranking one needs to establish how many variables are relevant. For this, a straightforward procedure was employed. The scores were represented in PC_1/PC_2 (the two selected components), and variables eliminated one by one from the least relevant until significant distortion of the data was visible upon inspection.

These two methods yield the same result, and suggest that only three (I0, AS and LS) of the nine variables are sufficient to capture the main structure of the data. A representation of the positioning of the different samples in this 3-variable PCA, shows that the separation between pathological and normal tissues is preserved, with the exception noted above. As in the previous case, we now inspect the impact of NR in the PCA analysis. For direct comparison, we consider the correlation approach in two different situations, depicted in the convex hull representation of Fig. 5. Panel (a) represents the original raw data, and panel (b) shows the impact of NR in group boundaries. Each group is treated separately. There is clearly a significant overlap of objects between the groups for each type of tissue, thereby implying that the discrimination is less straightforward than in the *breast cancer I* case. Even after filtering, panel (b), the overlap of the groups is still marked. Both before and after NR, the separation between the normal group, including the connective (CNN) and adipose tissues (ADI), and the pathological one, including carcinoma (CAR), fibro-adenoma (FAD) and mastopathy (MAS), is seen along the first principal component. The normal

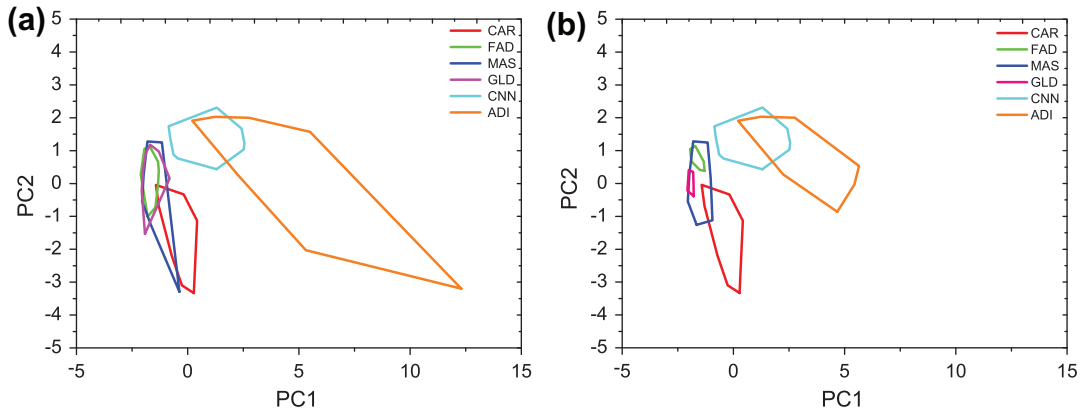

Fig. 5. Representation of the tissue classes in 2D convex hull form, showing the impact of NR in the group boundaries: before filtering data in original PCA system (a), after performing NR on PCA scores (b). For direct comparison, same colours are used in both panels. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

class of breast tissues is found towards positive values of the PC_1 axis, with the exception of the glandular tissue, while the pathological class are found for values of PC_1 close to zero or negative. Although it is possible to separate these two main groups, within each group tissues are strongly interconnected.

The filtering algorithm makes the data structure more defined, i.e., acts primarily on data points in the overlap regions. The separation between the glandular tissue (GLD) and fibro-adenoma (FAD) is complete after applying the NR algorithm. We stress again that, in these cases, noise reduction cannot be seen as a tool to eliminate outliers, but rather as a way to retain only the most relevant characteristic points of each group, although in some cases these concepts cannot be distinguished. Once again, we see that it is only possible to separate the pathological group from the normal one with the exception of the glandular tissue (GLD). This tissue falls in the region of the pathological groups, possibly due to similar morphological features. Another way to visualize this overlap is by displaying the data distribution according to the frequencies of each sample along PC_1 and PC_2, in Fig. 6. This figure represents the tissue distribution for the original PCA system, in panel (a), and after cleaning, in panel (b). It is seen that the overall appearance of the distributions is changed, mainly in fibro-adenoma, mastopathy, glandular and adipose tissues.

The precise classification of tumours in a specific issue is an extremely important task for the correct diagnosis, treatment and clinical follow-up of cancer patients. In this context, the LDA method was performed in two ways, either on the raw data or after NR on each group separately. Table 8 reports the misclassifications obtained by making combinations of pairs of tissues, each one pertaining to a different main group. From this table it is concluded that, in general, there is an increase in the percentage of correctly classified objects after performing NR. As expected, the adipose tissue has an accuracy of 100% in both situations. This means that the adipose tissue has different properties from the other tissues. Furthermore, after NR it is possible to completely separate carcinoma from glandular tissue. The highest misclassifications are obtained for the combination of glandular tissue with mastopathy (16.91%) and with fibro-adenoma (11.67%).

These results lead us to conclude that there are two choices for the most characteristic tissue of the normal class: either the connective tissue or the adipose tissue. However, given the proximity of the connective tissue to the pathological tissues, the best choice would be to use adipose tissue as a reference to the class of normal tissues. Furthermore, we found that the glandular tissue cannot be used as a reference. Finally, we should note that, as observed both in this latter and the former studies, the NR algorithm tends to increase the accuracy of discrimination, which means that it focuses on the 'grey' areas.

4.3. Data on the estimation of new cancer cases

Let us now address a final, different problem, relative to the distribution of cases and types of cancer within the United States. Following our proposed scheme, we will firstly employ hierarchical clustering analysis (HCA), which provides a visual means of estimating relationships among data points. Euclidean distance is used to represent the dissimilarity between states. A fundamental question concerns the normalization of the data. In this case, our option is to autonormalize each data point, corresponding to a specific state, that is, each variable is used in the form of a fraction

$$c_{ij} = n_{ij} \Big/ \sum n_{ij} \tag{10}$$

where c_{ij} is the fraction of cancer cases of type i in state j, n_{ij} the number of new cases of that type of cancer in that state, $\sum n_{ij}$ the total number of new cases for each state. It means that the number of cases for each type of cancer is thus divided by the total number of cases predicted for that state. As such, the data point is described by a set of variables which are the fraction of predicted occurrences for each type. The more similar states are those that have the same profile of cases, irrespective of the magnitude of incidence. Ward's method [18] is a standard for such analysis, and will be the one selected for this study. The dendrogram presented in Fig. 7 provides a very simple two dimensional plot of the data structure indicating the merging objects and the merging distances. It is constructed on the basis of the total existing information for the 51 states. From this figure, it is apparent that the data possesses a super-structure in which four groups of states are visible. These four groups are superimposed in the US map in Fig. 8.

They are located in the northern region (Group 1), in the eastern coast (Group 2), in the central region (Group 3) and in the south-west part (Group 4). The latter two groups have a certain degree of overlap. A clear underlying geographical pattern is, as such, visible in cancer distribution. After establishing the number of clusters, PCA was then directly applied, without any previous NR treatment, in order to reveal the relationship between the states and cancer types in these four groups. As a preliminary PCA result, the data scores representation of Fig. 9 is in direct agreement with the results obtained via HCA. In both cases, the groups are identified by matching colours.

Table 9 summarizes the PCA results, using the covariance approach. This approach, used in the normalized data, emphasizes

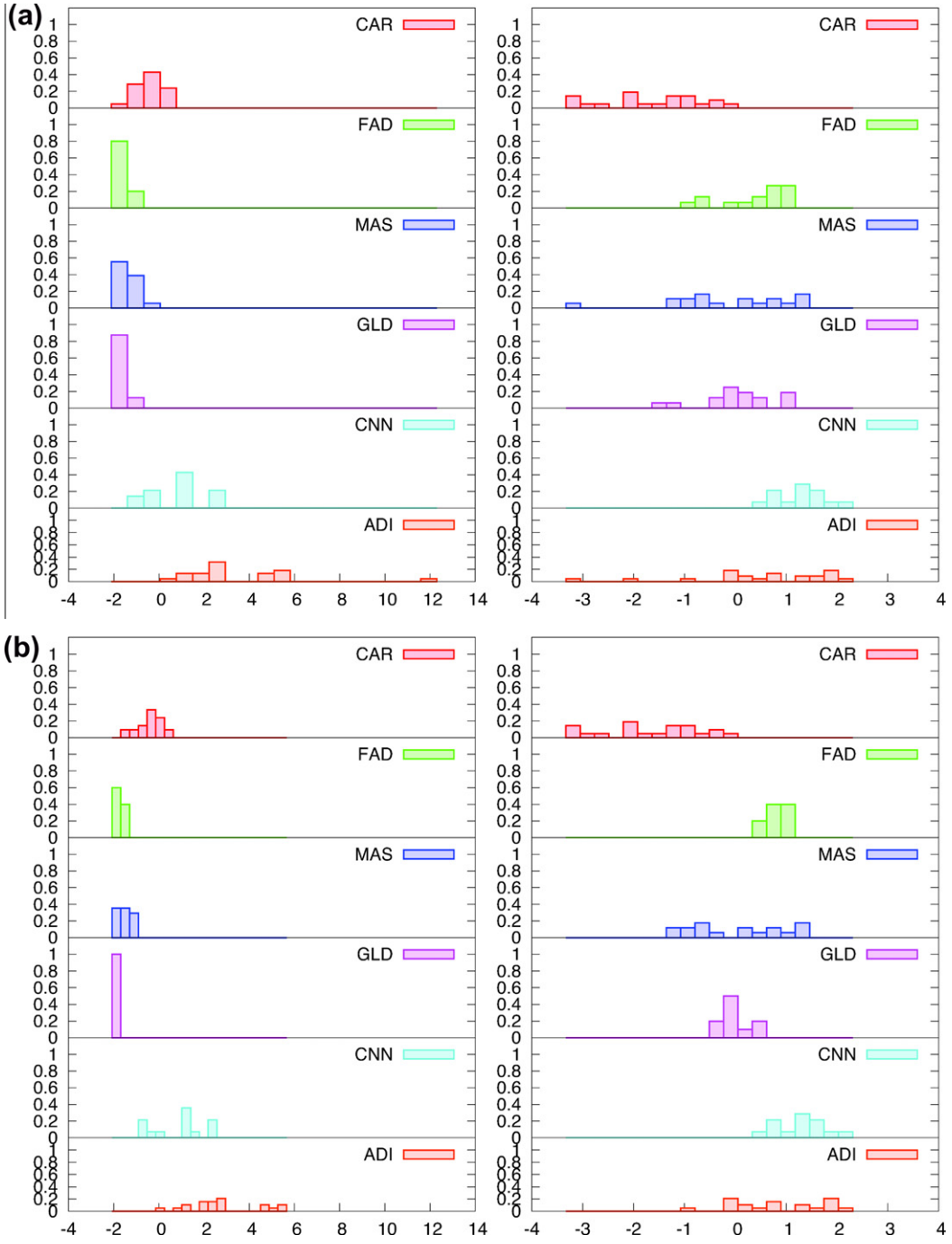

Fig. 6. Breast cancer tissues types in histogram form before (a) and after (b) being subject to the filtering algorithm. The distributions are coloured according to the each tissue type. The bars on the histograms display, for colour matching, the frequencies of the scores pertain to each group along PC_1 and PC_2.

Table 8
Misclassifications obtained by performing LDA on the tissue combinations, before and after NR.

	Misclassifications (%)	Normal tissues					
		Glandular tissue		Connective tissue		Adipose tissue	
		Before NR	After NR	Before NR	After NR	Before NR	After NR
Pathological tissues	Carcinoma	4.00	0.00	0.80	0.31	0.00	0.00
	Fibro-adenoma	11.67	9.38	0.54	0.60	0.00	0.00
	Mastopathy	16.91	5.93	2.56	2.58	0.00	0.00

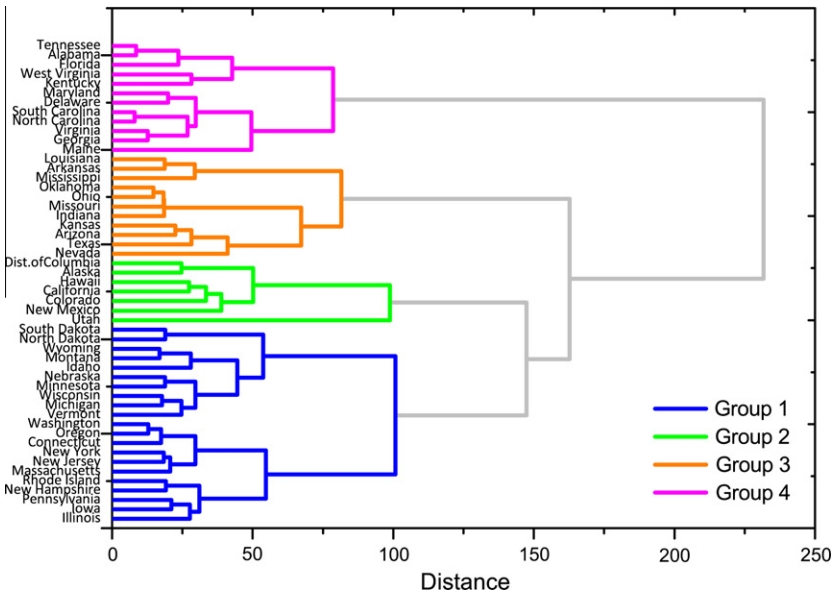

Fig. 7. Similarity among US states in terms of the distribution of new cancer cases by the selected cancer types. Dendrogram constructed resorting to Ward's method with euclidean distances, using the fraction of predicted occurrences for each cancer type as variables defining each state.

Fig. 8. Geographical representation of the groups formed from the HCA and PCA. Colours are related to the groups represented in Fig. 6. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

those fractions with a higher degree of absolute variation, i.e., variables with a large variation but small value are lost in the characterization. The first two principal components selected describe ca. 81% of the total variability. Inspecting the scores in Fig. 9 and the loadings of each variable, in Table 10, for the two relevant principal components, we conclude that lung and bronchus cancer (LBR) is the most important in the characterization of the data (weights of 0.724 and 0.403 to the first and second principal

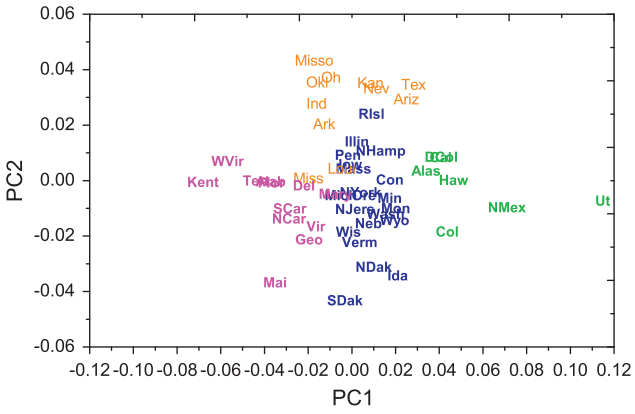

Fig. 9. Scatter plot of covariance scores using PC_1 versus PC_2, with 81% of information recovery. Data set contains 51 different states and 11 cancer types. Colours and symbols are related to each group type. For convenience, we used the abbreviations for the names of states. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 9
PCA results for the first three components using the covariance approach.

Principal components	Eigenvalue (λ_i)	Explained variance (%)	Cumulative explained variance (%)
PC_1	6.37	57.9	57.9
PC_2	2.51	22.8	80.7
PC_3	1.01	9.14	89.9

Table 10
Loading values of all variables, on the first two principal components (correlation approach).

	PC_1	PC_2
FBC	−0.051	−0.142
UTCx	0.002	0.006
CR	−0.109	0.028
UTCs	0.019	−0.032
LEU	0.023	−0.017
LBR	**−0.724**	**0.403**
MS	0.007	−0.131
NHL	0.005	−0.018
PR	0.189	**−0.633**
UBL	−0.02	−0.086
OT	**0.652**	**0.625**

component, respectively). From Fig. 9, it is seen that the four groups spread essentially along PC_1, but PC_2 also contributes for the discrimination of some of them.

Most of the states are placed in the PC_1/PC_2 plane in such a manner that their closest neighbours are those found in the map of the USA. PCA results also reveal that variable OT, in which some other types of cancer are gathered, is also influential in the characterization of the data. However, the low specificity renders it less useful for interpretation purposes. In this last case, the HCA/PCA combination provided some clear cut patterns that may serve as an important basis for subsequent studies.

5. Conclusions

In this work, we have shown that relatively simple and well known techniques from multivariate analysis, are sufficient to provide the tools for an in depth scrutiny of data originating from different screening methods used in the diagnosis of cancer, or epidemiological studies. HCA provides the overall structure of the data, based on all variables: groups of similar objects can be discerned, outliers identified, etc. It is usually not as informative as PCA, which allows this structure to be directly visualized and also provides the relative positioning of the different groups, which suggests the combined use of these two techniques. It is also seen that principal component analysis can be used as an effective method of feature selection, although more specialized alternatives exist in the literature [63,64]. The use of a noise reduction technique preserves only the more characteristic portion of the data, thus emphasizing the underlying patterns. LDA and PLS, in spite of their potential usefulness in other situations, have appeared from the present examples essentially as confirmatory tools. This happens essentially because the information for the discrimination of the data is contained in the directions along which the variation is the largest, which in fact is related to a good feature selection. Thus, the PCA or the HCA/PCA combination is extremely informative and provides a rationale for the observations. Note, however, than one drawback of PCA is that it arrives at linear combinations that capture only the characteristics of the predictive variables. Thus, no relevance is given to the relation of the dependent or target variable with the predictive variables. In cases where the overall structure of the response cannot be directly extracted from that of the descriptive variables, PLS provides an alternative approach to PCA. These cases encompass situations in which, for instance, the response is built from a set of small contributions from multiple predictors.

For the *breast cancer I* case study, the analysis suggests a set of interchangeable variables, that can be used to distinguish benign from malignant cases. In fact, each of these variables can be used on its own with high discriminating power. In the *breast cancer II* data, the distribution of the different types of tissues is readily available from the PCA procedure. Discrimination analysis, with and without noise reduction, also confirms the PCA results. Finally, HCA and PCA, used in conjunction, provide clear patterns in the geographical distribution of new cancer cases. The most discriminating, and specific, types of cancer in a USA state-by-state analysis are identified as lung and bronchus, and prostate. In summary, the use of this set of multivariate analysis techniques facilitates interpretation, allows graphical visualization, and can be used in an almost automated sequence for diagnosis or in epidemiological studies.

References

[1] Jossinet J, Lavandier B. The discrimination of excised cancerous breast tissue samples using impedance spectroscopy. Bioelectrochem Bioenerg 1998;45(2):161–7.
[2] Fabregue M, Bringay S, Poncelet P, Teisseire M, Orsetti B. Mining microarray data to predict the histological grade of a breast cancer. J Biomed Inform 2011;44:S12–6.
[3] Breast Cancer Treatment Information and Pictures. <breastcancer.org> (accessed July 2012).
[4] Mazurowski M, Lo J, Harrawood B, Tourassi G. Mutual information-based template matching scheme for detection of breast masses: from mammography to digital breast tomosynthesis. J Biomed Inform 2011;44:812–23.
[5] Biesheuvel C, Czene K, Orgeás C, Hall P. The role of mammography screening attendance and detection mode in predicting breast cancer survival – is there added prognostic value? Cancer Epidemiol 2011;35:545–50.
[6] Sawarkar S, Ghatol A, Pande A. Neural network aided breast cancer detection and diagnosis. In: Proceedings of the 7th WSEAS international conference on neural networks. World Scientific and Engineering Academy and Society (WSEAS); 2006. p. 158–63.
[7] Khanmohammadi M, Rajabi F, Garmarudi A, Mohammadzadeh R. Chemometrics assisted investigation of variations in infrared spectra of blood samples obtained from women with breast cancer: a new approach for cancer diagnosis. Eur J Cancer Care 2010;19(3):352–9.
[8] Thurfjell E, Lernevall K, Taube A. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994;191(1):241–4.

[9] Mautner B, Schmidt K, Brennan M. New diagnostic techniques and treatments for early breast cancer. Seminars in oncology nursing, vol. 16. Elsevier; 2000. p. 185–96.
[10] Piacenti da Silv M, Zucchi O, Ribeiro-Silva A, Poletti M. Discriminant analysis of trace elements in normal, benign and malignant breast tissues measured by total reflection X-ray fluorescence. Spectrochim Acta Part B: Atom Spectrosc 2009;64(6):587–92.
[11] Eaton JW, Bateman D, Hauberg S. GNU octave manual: a high-level interactive language for numerical computations. Network Theory Ltd.; 2007. Version 3 for Octave version 3.2.4.
[12] Acar E, Bro R, Schmidt B. New exploratory clustering tool. J Chemometrics 2008;22(1):91–100.
[13] Downs G, Barnard J. Clustering methods and their uses in computational chemistry. Rev Comput Chem 2002:1–40.
[14] Daszykowski M, Walczak B, Massart D. Density-based clustering for exploration of analytical data. Anal Bioanal Chem 2004;380(3):370–2.
[15] Almeida J, Barbosa L, Pais A, Formosinho S. Improving hierarchical cluster analysis: a new method with outlier detection and automatic clustering. Chemometrics Intell Lab Syst 2007;87(2):208–17.
[16] Meila M. Comparing clusterings – an information based distance. J Multivariate Anal 2007;98(5):873–95.
[17] Everitt B, Landau S, Leese M, Stahl D. Cluster analysis. 5th ed. John Wiley & Sons Ltd.; 2011.
[18] Ward Jr J. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963:236–44.
[19] Jolliffe I. Principal component analysis. 2nd ed. New York: Springer; 2002.
[20] Wang X, Paliwal K. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. Pattern Recognit 2003:2429–39.
[21] Brereton R. Chemometrics: applications of mathematics and statistics to laboratory systems. E. Horwood; 1990.
[22] Massart D. Chemometrics: a textbook. 2nd ed. Elsevier Science; 1988.
[23] Torokhti A, Friedland S. Towards theory of generic principal component analysis. J Multivariate Anal 2009;100(4):661–9.
[24] Campanella L, De Angelis G, Visco G. Chemometric investigation of the efficiency of different TiO_2-based catalysts as principal components of toc photochemical sensors under development. Anal Bioanal Chem 2003;376(4):467–75.
[25] Kokot S, Grigg M, Panayiotou H, Phuong T. Data interpretation by some common chemometrics methods. Electroanalysis 1998;10(16):1081–8.
[26] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics Intell Lab Syst 1987;2(1–3):37–52.
[27] Jackson J, Wiley J. A user's guide to principal components. Wiley Online, Library; 1991.
[28] Madsen R, Lundstedt T, Trygg J. Chemometrics in metabolomics – a review in human disease diagnosis. Anal Chim Acta 2010;659(1–2):23–33.
[29] Polat K, Günes S. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digit Signal Proc 2007;17(4):702–10.
[30] Latifoglu F, Polat K, Kara S, Günes S. Medical diagnosis of atherosclerosis from carotid artery doppler signals using principal component analysis (PCA), *k*-NN based weighting pre-processing and artificial immune recognition system (AIRS). J Biomed Inform 2008;41(1):15–23.
[31] Polat K, Günes S. Computer aided medical diagnosis system based on principal component analysis and artificial immune recognition system classifier algorithm. Expert Syst Appl 2008;34(1):773–9.
[32] Reich D, Price A, Patterson N. Principal component analysis of genetic data. Nat Genet 2008;40(5):491–2.
[33] Espeland M, Bray G, Neiberg R, Rejeski W, Knowler W, Lang W, et al. Describing patterns of weight changes using principal components analysis: results from the action for health in diabetes research group. Ann Epidemiol 2009;19(10):701–10.
[34] Cheng L, Burns M, Taylor J, He W, Halpern E, McDougal W, et al. Metabolic characterization of human prostate cancer with tissue magnetic resonance spectroscopy. Cancer Res 2005;65(8):3030–4.
[35] Ringnér M. What is principal component analysis? Nat Biotechnol 2008;26(3):303–4.
[36] Spratlin J, Serkova N, Eckhardt S. Clinical applications of metabolomics in oncology: a review. Clin Cancer Res 2009;15(2):431–40.
[37] Edefonti V, Bravi F, Garavello W, La Vecchia C, Parpinel M, Franceschi S, et al. Nutrient-based dietary patterns and laryngeal cancer: evidence from an exploratory factor analysis. Cancer Epidemiol Biomarkers Prev 2010;19(1):18–27.
[38] Xie Y, Xiao G, Coombes K, Behrens C, Solis L, Raso G, et al. Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients. Clin Cancer Res 2011;17(17):5705–14.
[39] Gualberto A, Dolled-Filhart M, Gustavson M, Christiansen J, Wang Y, Hixon M, et al. Molecular analysis of non-small cell lung cancer identifies subsets with different sensitivity to insulin-like growth factor I receptor inhibition. Clin Cancer Res 2010;16(18):4654–65.
[40] Méndez E, Houck J, Doody D, Fan W, Lohavanichbutr P, Rue T, et al. A genetic expression profile associated with oral cancer identifies a group of patients at high risk of poor survival. Clin Cancer Res 2009;15(4):1353–61.
[41] Selaru F, Yin J, Olaru A, Mori Y, Xu Y, Epstein S, et al. An unsupervised approach to identify molecular phenotypic components influencing breast cancer features. Cancer Res 2004;64(5):1584–8.
[42] Woo H, Kim K, Choi M, Jung B, Lee J, Kong G, et al. Mass spectrometry based metabolomic approaches in urinary biomarker study of women's cancers. Clin Chim Acta 2009;400(1–2):63–9.
[43] Liland K, Indahl U. Powered partial least squares discriminant analysis. J Chemometrics 2009;23(1):7–18.
[44] Candolfi A, Wu W, Massart D, Heuerding S. Comparison of classification approaches applied to NIR-spectra of clinical study lots. J Pharm Biomed Anal 1998;16(8):1329–47.
[45] Fisher R. The use of multiple measurements in taxonomic problems. Ann Human Genet 1936;7(2):179–88.
[46] Sharaf M, Illman D, Kowalski B. Chemometrics, vol. 82. Wiley Interscience; 1986.
[47] Vinzi VE, Chin WW, Henseler J. Handbook of partial least squares: concepts, methods and applications. Springer; 2010.
[48] Abdi H. Partial least squares regression and projection on latent structure regression (PLS regression). Wiley interdisciplinary reviews: computational statistics 2010;2(1):97–106.
[49] Rosipal R, Kräme N. Overview and recent advances in partial least squares. Subspace Latent Struct Feature Sel 2006:34–51.
[50] Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Anal Chim Acta 1986;185:1–17.
[51] Devroye L, Toussaint G. A note on linear expected time algorithms for finding convex hulls. Computing 1981;26(4):361–6.
[52] Avis D, Bremner D, Seidel R. How good are convex hull algorithms? Comput Geomet 1997;7(5–6):265–301.
[53] Wolfram MathWorld. mathworld.wolfram.com in ConvexHull (accessed February 2011).
[54] Cormen T. Introduction to algorithms. The MIT Press; 2001.
[55] Mangasarian O, Wolberg W. Cancer diagnosis via linear programming, vol. 23. University of Wisconsin-Madison, Computer Sciences Dept.; 1990.
[56] Estrela da Silva J, Marques de Sá J, Jossinet J. Classification of breast tissue by electrical impedance spectroscopy. Med Biol Eng Comput 2000;38(1):26–30.
[57] American Cancer Society. Cancer Facts & Figs. 2010 Database. <www.cancer.orgin/Research/CancerFactsFigures/CancerFactsFigures/cancer-facts-and-figures-2010> (accessed November 2010).
[58] University of California, Irvine, Machine Learning Repository. ics.uci.edu in directory/pub/machinelearning-databases/wisconsin-breastcancer (accessed November 2010).
[59] University of California, Irvine, Machine Learning Repository. rchive.ics.uci.edu in ml/machine-learning-databases/00192 (accessed November 2010).
[60] Zhang J. Selecting typical instances in instance-based learning. In: Proceedings of the international machine learning conference; 1992. p. 470–9.
[61] Jossinet J. Variability of impedivity in normal and pathological breast tissue. Med Biol Eng Comput 1996;34(5):346–50.
[62] Jossinet J. The impedivity of freshly excised human breast tissue. Physiol Meas 1998;19:61–5.
[63] Ravi V, Reddy P, Zimmermann H. Pattern classification with principal component analysis and fuzzy rule bases. Eur J Oper Res 2000;126(3):526–33.
[64] Hu Y. Genetic algorithm in designing fuzzy information retrieval-based classifier by principal component analysis. Comput Ind Eng 2006;51(1):117–27.