SOLID BIOFUELS – IMAGE ANALYSIS AND MULTIVARIATE MODELING
FOR THE AUTOMATED QUALITY ASSESSMENT OF WOOD CHIPS

(s.s.d. AGR/09)

Tesi di dottorato di
Ing. Paolo Felli

Coordinatore del corso
Prof. Massimiliano Cecchini

Firma

Tutore
Prof. Danilo Monacra
Firma

Co-tutore
Dott. Luigi Pari
Firma

23 giugno 2014
Per fare Scienza c’è bisogno di rigore matematico e riproducibilità sperimentale
(Antonino Zichichi, *Tra Fede e Scienza*)
**ABSTRACT**

The European technical standards EN 14961-1:2010 and EN 14961-4:2011 on solid biofuels determine the fuel quality classes and specifications for wood chips. One of the standard methods for size classification of particulate biofuels is the oscillating screen method (EN 15149-1:2010); it allows the determination of the size distribution of wood chips.

Five 8-liter samples of wood chips deriving from the harvesting of short-rotation forestry (SRF) poplar have been analyzed. The wood chips, after sieving through horizontally oscillating sieves, were sorted in decreasing size classes (<3.15, 3.15÷8, 8÷16, 16÷45, 45÷63, 63÷100, >100 mm) and the mass of each fraction was measured to determine the size distribution of the particles. Typically, the results of sieving are presented in cumulative size distribution curves, which refer to the cumulative percentage mass of all fractions.

Image analysis can provide accurate measurement of the dimensions defined for each individual particle and could provide a new method to determine the wood chip size integrated with shape. Therefore, images of 7583 wood chips (corresponding to three size fractions of the entire sample: 8÷16, 16÷45, 45÷63 mm) were analyzed to extract size and shape descriptors (area, major and minor axis lengths, perimeter, eccentricity, equivalent diameter, fractal dimension index, Feret diameters and Fourier descriptors). However, traditional image analysis allows size distribution curves to be constructed with respect to cumulative percentage area. Thus, a Partial Least Squares Regression (PLS-R) model was developed for mass (or weight) prediction of poplar chips. These predicted weights were used to construct, for the first time within the biofuel sector, a cumulative distribution curve completely based on two-dimensional image analysis.

Both supervised and unsupervised pattern recognition methods were considered for wood chips classification, in accordance with the sieving fractions. In the manuscript *Determining wood chip size: image analysis and clustering methods*, a cluster analysis on the shape and size descriptors was applied to observe the wood chips’ distribution. The obtained dendrogram shows a group separation according to the original three sieving fractions considered in the analysis. In the submitted version of the manuscript *Automated determination of wood chip size distribution based on combined image and multivariate analyses*, two Partial Least Squares Discriminant Analysis (PLS-DA) models, with and without Fourier descriptors (shape descriptors), were developed for estimation of wood chips size classification, realizing a type of sieving software.
simulation. The classification of the three size fractions showed high values of correct classification for both models in the independent test.

Considering the size fraction assignment obtained by the PLS-DA model and the weight predicted by the PLS-R model to each of the 7583 particles, a conversion from image-analysis morphometries to sieving result was realized. The combination of image analysis with multivariate modeling approaches allowed cumulative size distribution curves to be constructed based on chip esteemed mass (or weight) and assigned fraction, which show a positive agreement with the sieve results.

PLS-R and PLS-DA methods are explained in detail in the respective chapters. Other chapters are used to outline the covered topics (e.g., principal component analysis, regression, discriminant analysis) and to allow the peculiarities of the selected models (PLS-R and PLS-DA) to be highlighted.

The proposed method could help in fixing standard methodologies which would be highly replicable, objective and reliable. An improvement of the results is expected by introducing a thickness measurement, which could contribute to better express particle form (in 3-D) and mass prediction, and to achieve an improved modeling.
CONTENTS

INTRODUCTION ...................................................................................................................... 7

1 SHAPE DESCRIPTION ........................................................................................................ 8
   1.1 Wang’s shape theory .................................................................................................. 8
   1.2 Fourier descriptors ................................................................................................. 10

2 UNSUPERVISED PATTERN RECOGNITION METHODS .................................................. 14
   2.1 The basic for resemblance ..................................................................................... 14
   2.2 Cluster analysis ...................................................................................................... 15
   2.3 Projecting data sets in a few dimensions ............................................................... 17
   2.4 Principal axes ......................................................................................................... 18
   2.5 Principal component analysis .............................................................................. 19
   2.6 Clustering and ordination ..................................................................................... 22

3 REGRESSION ..................................................................................................................... 24
   3.1 Forecasting and prediction .................................................................................... 24
   3.2 Linear regression ................................................................................................. 25
   3.3 Principal component regression (PCR) ................................................................. 27
   3.4 Correlation coefficient .......................................................................................... 30
   3.5 Coefficient of determination .............................................................................. 31

4 PARTIAL LEAST SQUARES (PLS) REGRESSION .......................................................... 32
   4.1 General considerations .......................................................................................... 32
   4.2 Partial least squares (PLS) algorithm ................................................................. 33
   4.3 PLS regression model ......................................................................................... 35
   4.4 SIMPLS .................................................................................................................. 39
   4.5 Interpretation of the PLS-R model ....................................................................... 39
   4.6 Latent variables ..................................................................................................... 41
   4.7 The number of PLS components ......................................................................... 41
   4.8 Mean square error .................................................................................................. 46

5 DISCRIMINANT ANALYSIS ............................................................................................. 48
   5.1 Principles of canonical analysis ............................................................................ 48
   5.2 Discriminant analysis ........................................................................................... 49
   5.3 Multivariate classification methods ........................................................................ 54

6 PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS ............................................. 57
   6.1 PLS-DA modeling method ..................................................................................... 57
   6.2 Prediction probability and classification threshold for PLS-DA ......................... 61
   6.3 Variable importance in the projection (VIP) ......................................................... 65

REFERENCES ...................................................................................................................... 67

POSTERS AND MANUSCRIPTS ......................................................................................... 73
POSTERS .............................................................................................................................................. 74
DETERMINING WOOD CHIP SIZE: IMAGE ANALYSIS AND CLUSTERING METHODS ...... 76
AUTOMATED DETERMINATION OF WOOD CHIP SIZE DISTRIBUTION BASED ON COMBINED
IMAGE AND MULTIVARIATE ANALYSES .......................................................................................... 84
FURTHER ANALYSES DEVELOPED APART THE MANUSCRIPTS .................................................. 104

GENERAL CONCLUSIONS .................................................................................................................. 110
INTRODUCTION

The European standard EN 14961-1:2010 (Solid biofuels - Fuel specifications and classes - Part 1: General requirements) determines the fuel quality classes and specifications for solid biofuels. Table 5 of this technical standard states the properties of dimensions, moisture and ash of wood chips; the specification of dimensions refers to documents CEN/TS 15149-1 and -2. Sieving methods are currently used for the determination of particle size distribution. The European standard EN 15149-1:2010 (Solid biofuels - Determination of particle size distribution – Part 1: Oscillating screen method using sieve apertures of 1 mm and above) describes the reference method for size classification of particulate biofuels by the horizontally oscillating screen method.

The aim of this PhD Thesis is to analyze how image analysis combined with multivariate modeling methods can be used to construct cumulative size distribution curves, which can be compared with the sieving results required by the mentioned EN standards. This has been done through a Partial Least Squares Regression model for the mass prediction of poplar chips and Partial Least Squares Discriminant Analysis models for estimation of wood chips size classification.

The following six chapters, which have been elaborated from books and scientific journals, represent an overview of the covered topics and are intended to illustrate how the considered methods work. The final part is the core of the thesis; it contains a selection of manuscripts, posters and further analyses developed apart.
Chapter 1

SHAPE DESCRIPTION

1.1 Wang’s shape theory

The most common measurements that are made on objects are those that describe their form. Form features are physical dimensional measures that characterize the appearance of an object (Costa et al., 2011). Since the wood chip shape and size influence the sieving results, Wang’s shape theory (Wang, 1994), which defined shapes to classify the particles, has been analyzed.

Image analysis, which allows the shape of objects to be characterized and recognized, can work with Feret diameters (Howarth and Rowlands, 1987), the length between two tangential lines on opposite sides of a particle. To determine the length and width of a particle, the algorithm measures Feret diameters in 32 (typical value) azimuth directions and takes the longest distance as $D_{\text{max}}$ (length) and the shortest distance as $D_{\text{min}}$ (width). Wang (Wang, 1994) propounded a theory suggesting that geometric forms (rectangles, triangles, trapezoids, diamonds) should fall into different areas of a plot of the shape factor ($Sh$, the ratio between the particles perimeter and area) against the ratio between $D_{\text{max}} / D_{\text{min}}$ (Fernlund, 1998).

Figure 1 shows a parallelepiped, where $CH = D_{\text{min}} = h$, $AC = D_{\text{max}} = n^*h$, $BH = x$. When the distance $x$ equals 0, the shape is a rectangle. When $x = [h(n^2-2)]/[2\sqrt{(n^2-1)}]$, $AB$ equals $BC$, the shape becomes a diamond; if $n = \sqrt{2}$ the shape becomes a square. Since $D_{\text{max}} / D_{\text{min}} = n$, it is easy to verify that the ratio between perimeter and area ($Sh$) is:

$$P/A = \frac{2\left(h\sqrt{n^2-1} - x + \sqrt{x^2 + h^2}\right)}{h^2 \sqrt{n^2-1} - hx}$$

Figure 1: generic parallelepiped

Figure 2 shows a triangle where $CH = D_{\text{min}} = h$, $AB = D_{\text{max}} = n^*h$, $AH = x$. When $AC=BC$, the $x$ distance equals $n^*h/2$; when $AC=AB$, then $x = h\sqrt{(n^2-1)}$. By varying $x$ between these two values, a generic triangle is obtained. Because $D_{\text{max}} / D_{\text{min}} = n$, it is easy to verify that the ratio between perimeter and area ($Sh$) is:

$$P/A = \frac{\sqrt{x^2+h^2} + \sqrt{(nh-x)^2+h^2} + nh}{\frac{nh^2}{2}}$$

Figure 2: generic triangle
In order for the forms to be invariant to the scale of the contour, each of them has been normalized by the length of $D_{\text{min}} = h = 1$; then, the shape classification becomes size independent.

In figure 3, diamonds, rectangles and triangles are plotted for $n$ varying in the range $\sqrt{2} \div 10$.

![Figure 3: plots of typical shapes (diamonds, rectangles and triangles)](image)

A plot of $Sh$ against the ratio between length and width was made (Fernlund, 1998) for rectangular, triangular and diamond shaped particles. Mathematically, the different shapes fall into different areas in the diagram. However, the geometric forms were not easily differentiated using the 32 azimuth directions to compute the Feret diameters. The computed values did not agree with the theoretical ones and so there was a great deal of overlap. In a two dimensional plot, it is quite difficult to distinguish sufficiently the different shapes.

Wang considered four descriptors (perimeter, area, $D_{\text{max}}$, $D_{\text{min}}$) to represent forms in a two-dimensional space; the relative positions of point-objects in that space cannot always be distinguished. This method contracts the reference space, as if the distances between objects were smaller in that region of space.

The distances between shapes in a two-dimensional space are smaller than in a multidimensional space. To discriminate the real shapes of objects, it is helpful to represent the point-objects in a multidimensional diagram, with many axes, which is equivalent to dilating the reference space.
1.2 Fourier descriptors

Global descriptors such as area, circularity, eccentricity and axis orientation can only discriminate shapes with large dissimilarities (Zhang and Lu, 2003). Instead, methods based on Fourier transformation have found many applications in image processing and shape description.

The Fourier series representation of a continuous-time periodic signal generally requires infinitely many harmonically related complex exponentials, whereas the Fourier series for any discrete-time signal with period \( N \) requires only \( N \) harmonically related complex exponentials. The sequence of discrete Fourier series coefficients of a periodic sequence is itself a periodic sequence (with period \( N \)). The finite-length sequence corresponding to one period of the discrete Fourier series coefficients is the Discrete Fourier Transform (DFT). Since DFT is equal to samples of the Fourier transform at equally spaced frequencies, consequently computation of the \( N \)-point DFT corresponds to the computation of \( N \) samples of the Fourier transform at \( N \) equally spaced frequencies. (Oppenheim and Schafer, 1989).

Spectral descriptors include Fourier descriptors (FDs), which have been successfully used for the characterization of closed contours. The considered outline shapes, obtained from real world objects (e.g. wood chips), can be described as single plane closed curves. The binarized shape images can be filtered to eliminate noise, then the shape boundaries are obtained.

The coordinates of each boundary, \( x(t) \) and \( y(t) \) \( (t = 0, 1,.., N-1) \), can be extracted from each object shape. The boundary is represented as the sequence of successive pixel coordinates in the complex plane, \( x(t) + j \ y(t) \); it is a one-dimensional representation of a two-dimensional image contour. In order to eliminate the effect of bias, shifted coordinates can be used to centre the contour at the origin point

\[
u(t) = (x(t) - x_c) + j \ (y(t) - y_c)\]

where \((x_c, y_c)\) are the coordinates of the centroid of the shape, which is the average of the boundary coordinates:

\[
x_c = \frac{1}{N} \sum_{t=0}^{N-1} x(t) \quad y_c = \frac{1}{N} \sum_{t=0}^{N-1} y(t)
\]

This shift makes the shape representation invariant to translation (Zhang and Lu, 2003).

In that way, a closed curve, which defines a given shape, can be represented by a complex coordinate function \( u(t), \ t = 0, 1,.., N-1 \), often called shape signature, where \( N \)
is the number of boundary points (number of pixels along the boundary). Since the curve is closed, \(u(t)\) can be considered periodic. The Fourier series representation of a periodic sequence corresponds to the DFT of a finite-length sequence. The discrete Fourier transform is given by

\[
F[n] = \frac{1}{N} \sum_{t=0}^{N-1} u(t) e^{-j2\pi nt/N}
\]

The complex coefficients \(F[n]\) \((n = 0, 1, \ldots, N-1)\) are the non-normalized Fourier descriptors of the contour; they are used to derive the normalized, invariant Fourier descriptors of a shape, denoted as \(FD_n\) \((n = 0, 1, \ldots, N-1)\).

An important property of Fourier methods is the potential to invert the Fourier transform and reconstruct an outline shape from a set of Fourier coefficients. The inverse transform of the \(F[n]\) restores the boundary \(u(t)\). Since the Fourier transform is a complex transform, the frequency spectrum has negative frequencies as well as positive frequencies, with the constant component \(F[0]\) in the middle. The range of \(n\) can be restricted to \([-N/2+1, N/2]\) because the shape is described by a set of \(N\) points on the contour. According to Shannon’s theorem, the highest frequency of the discrete representation of the outline is obtained for \(n = N/2\) and any \(F[n]\) with \(n\) greater than \(N/2\) would be redundant. The coefficients \(F[n]\) describe the frequency contents of the closed curve. Thereby, contours can be represented by vectors whose elements are the Fourier coefficients.

When only \(M\) \((< N)\) frequency components (descriptors) are used to reconstruct the contour, the inverse transform can be given by

\[
\hat{u}[l] = \sum_{n=-M/2}^{M/2} F[n] e^{j2\pi nl/L}
\]

where \(\hat{u}[l]\) denotes the reconstructed point \((l = 0, 1, \ldots, L-1)\). \(L\) parameter is given by the ratio of the curve length to the distance between the points to be reconstructed. The approximated boundary consists of \(L \leq N\) points. Less terms are used for the reconstruction \((M < N)\); the smaller \(M\) becomes, the more detail is lost on the boundary.

Fourier transformation of a boundary signature function generates a complete set of complex numbers (the Fourier descriptors), which represent the shape of the object in a frequency domain (Kauppinen et al., 1995). Lower frequency components contain information about the general shape features, while higher frequency components contain information about finer features of the shape (details). For classification
purposes, a subset of the components is often enough to capture the overall features of the contour and to discriminate different shapes (Kauppinen et al., 1995, Zhang and Lu, 2003).

In order to compare shape representations, these must be invariant to translation, starting point, rotation and scale (Zhang and Lu, 2003). The described shape representation is invariant to translation, i.e. insensitive to the placement of the object in the picture plane (Kauppinen et al., 1995), therefore the corresponding $F[n]$s are also translation invariant. The constant component $F[0]$ depends only on the position of the shape; it is not useful in describing shape, thus it is not needed (Kauppinen et al., 1995).

Orientation invariance is normally achieved by phase normalization (Persoon and Fu, 1977; Mitchell and Grogan, 1984). Different starting points affect the order of elements in $u(t)$ and the $F[n]$ obtained, so changes in starting point affect only the phase of the descriptor. Starting point and rotation invariance of the Fourier coefficients $F[n]$ are achieved by ignoring the phase information and by using only the absolute values of the descriptors (Mitchell and Grogan, 1984). Scale invariance, which standardizes the size of the contour, is achieved by dividing the absolute values of the descriptors by the absolute value of the first non-zero frequency component. The invariant feature vector of the contour Fourier method, used for indexing the shape, is (Kauppinen et al., 1995)

$$FD_s = \left[ \begin{array}{c} \frac{|F_{-(N/2-1)}|}{|F_1|}, \frac{|F_{-1}|}{|F_1|}, \frac{|F_2|}{|F_1|}, \ldots, \frac{|F_{N/2}|}{|F_1|} \end{array} \right]^T$$

A common approach to shape classification is to use only a subset of the descriptors (Kunttu et al. 2004). $FD_s$ provides a framework for shape description and comparisons. Complex shapes can be represented with a limited number of invariant coefficients, which can be viewed as features extracted from the original shapes. In a multidimensional space, similarity between shapes can be measured by existing Euclidean distance between the two feature vectors of the shapes (Zhang and Lu, 2002). Shapes that are similar will “live” near each other in the feature space. The sum of squared differences between shapes can give a measure of similarity between shapes.

If $M$ is the truncated number of components needed to index the shape, for an $m$ shape indexed by $FD^m = [FD^m_{-M/2}, \ldots, F D^m_0, F D^m_1, \ldots, F D^m_{M/2}]^T$ and a $d$ shape indexed by $FD^d = [FD^d_{-M/2}, \ldots, F D^d_{-1}, F D^d_0, \ldots, F D^d_{M/2}]^T$, the Euclidean distance between the two feature vectors can be used as the similarity measurement (Zhang and Lu, 2003)
\[ D = \left( \sum_{i=\frac{-M}{2}}^{\frac{M}{2}} |FD_i^m - FD_i^d|^2 \right)^{1/2} \]

If \( M/2 = 50 \) is selected, 99 \( FD_i \) are in the invariant feature vector.

The Euclidean distances between point-objects becomes greater as the harmonic content increases. In considering \( M \) components, with \( M < N \), some information is lost. Higher \( M \) values define fewer errors, but it is also desirable to limit the number of components used in a subsequent statistical analysis. The Fourier descriptors give unique and separable classifications as long as sufficient components are included in the truncated Fourier series. In (Schnabel, 1995), various levels of \( M \) were used for the reconstruction of a medical image contour (\( M = 3, M = 453, \) and \( M = N = 953 \)). A visual estimate of the number of components required to describe an outline with a given degree of precision can be made from a series of inverse Fourier reconstructions using incremental numbers of components.
Chapter 2

UNSUPERVISED PATTERN RECOGNITION METHODS

2.1 The basic for resemblance (Legendre, 2012)

Measures of resemblance among objects (or descriptors) are used to cluster the objects (or descriptors) or to produce ordination diagrams in spaces of reduced dimensionality. The clustering of objects is an operation by which the set of objects is partitioned in two or more subsets (clusters), using pre-established rules of agglomeration or division. Ordination in reduced space is an operation by which the objects are represented in a space that contains fewer dimensions that in the original data set; the positions of the objects with respect to one another may also be used to cluster them.

In the space of descriptors, the reference axes correspond to quantitative, metric descriptors. This space is called metric, or also Euclidean because Euclide’s geometry holds in that space. The objects may be represented along these axes (which correspond to the descriptors).

The resemblance or difference between objects may be quantified using similarity or distance coefficients. Similarities are maximum ($S=1$) when the two objects are identical and minimum when the two objects are completely different; distances follow the opposite rule. A similarity can be transformed into a distance by computing its one-complement. For a similarity measure varying between 0 and 1 the corresponding distance may be computed as $D = 1 - S$.

The most common metric measure is the Euclidean distance, which is computed using Pythagoras’ formula from points positioned in a K-dimensional space. The Euclidean distance does not have an upper limit, its value increasing indefinitely with the number of descriptors. The value also depends on the scale of each descriptor. This latter problem may be avoided by using standardized variables instead of the original data. Standardized descriptors have the same scale (i.e. unit standard deviation) and are dimensionless. Standardization is achieved by subtracting the mean (translation) and dividing by the standard deviation of the variable (expansion). A standardized variable has three interesting proprieties: its mean is zero, its variance is 1, it is a dimensionless variable.

Distances, which in some cases are not bound by a pre-determined upper value, may be normalized using equation $D_{\text{norm}} = D / D_{\text{Max}}$, where $D_{\text{norm}}$ is the distance normalized.
in the interval $0 \div 1$, and $D_{Max}$ is the maximum value taken by the distance coefficient. Normalized distances can be used to compute similarities: $S = 1 - D_{norm}$. Distance coefficients are functions which take their maximum values (often 1) for two objects that are completely different, and 0 for two objects that are identical over all descriptors. Distances, like similarities, are used to measure the association among objects.

2.2 Cluster analysis (Legendre, 2012)

Classifying objects into collective categories requires the recognition of discontinuous subsets in an environment which is sometimes discrete, but most often continuous. To cluster is to recognize that objects are sufficiently similar to be put in the same group and to also identify distinctions or separations among groups of objects. Clustering is an operation of multidimensional analysis that consists in partitioning a collection of objects (or descriptors). A partition is a division of a set of objects (collection) into subsets, or clusters, such that the objects within each cluster are more similar to one another than to objects in the other clusters. In hard or crisp clustering, the groups are mutually exclusive and each object belongs to a single group of a partition; in fuzzy clustering, on the contrary, an object may simultaneously belong, to different degrees, to two or more groups of a partition (overlapping clusters). In a partition resulting from a hard clustering method, each object is characterized by a state (its cluster) of the classification and it belongs to only one of the clusters. A clustering of objects defined in this way imposes a discontinuous structure onto the data set, even if the objects have originally been sampled from a continuum. This structure results from the grouping into subsets of objects that are recognized as sufficiently similar given the variables considered. It is then possible to look for characteristics that differentiate the clusters from one another.

There are two major categories of clustering methods. In a descriptive clustering misclassifying objects is to be avoided, even at the expense of creating single object clusters. In a synoptic clustering all objects are forced into one of the main clusters; the objective is to construct a general conceptual model that encompasses a reality wider than the data under study. Both approaches have their usefulness.

Partitioning considers hierarchical and non-hierarchical methods. In hierarchical methods, the members of inferior-ranking clusters become members of larger, higher-ranking clusters. Most of the time, hierarchical methods produce non-overlapping clusters. There are two approaches to hierarchical clustering: “bottom up”, grouping
small clusters into larger ones, and “top down”, splitting big clusters into small ones. These are called agglomerative and divisive clusterings, respectively. *Non hierarchical* methods produce a single partition which optimizes within-group homogeneity, instead of a hierarchical series of partitions optimizing the hierarchical attribution of objects to clusters. Non hierarchical methods should be used in cases where the aim is to obtain a direct representation of the relationship among objects instead of a summary of their hierarchy. Lance & Williams (1967) restrict the term *clustering* to the non-hierarchical methods and call the hierarchical methods *classification*. Non-hierarchical methods include *K*-means partitioning and the ordination techniques (e.g. PCA) used as clustering methods.

Hierarchical clustering results are commonly represented, in most cases, as *dendrograms* (trees). They are made of branches that meet at nodes which are drawn at the similarity value where fusion of branches takes place. The branches may point upwards, downwards or sideways. Dendrograms are graduated in distances or similarities; the branching pattern indicates the distances or similarities of bifurcating branches. Dendrograms clearly illustrate the clusters formed at each partition level. They only display the clustering topology and object names, or their code numbers, which are written at the tips of the branches, not the links between objects. The branches bifurcating from a node are not fixed but may be swiveled as required by the presentation of results without altering the nature of the information contained in the dendrogram.

The similarity (or distance) level at which two objects become members of the same cluster is called cophenetic similarity (or distance). Any dendrogram can be uniquely represented by a matrix in which the similarity (or distance) for a pair of objects is their cophenetic similarity (or distance). Any ordering of objects in the *cophenetic matrix* is acceptable. For a partition of the data set (as in the *K*-means method), the resulting groups of objects are not related through a dendrogram. A cophenetic matrix may nevertheless be obtained. Any classification or partition can be fully described by a cophenetic matrix. This matrix is used for comparing different classifications of the same objects.

The degree of *isolation* of clusters in metric space of descriptors can be measured as the distance between the two closest objects in different clusters. It may also be measured as the mean distance between all the objects in one cluster and all objects in another, or else as the ratio of the distance between the two closest objects to the distance between
the centroids of the two clusters. These measures are ways of quantifying the distances among clusters; a clustering or ordination of clusters can be computed using these distances.

A simple clustering method is based upon the Euclidean distances among points representing the objects. Different clustering algorithms may give different results on the same data. The same clustering algorithm may give different results on the same data, if, like K-means, it involves some arbitrary initial condition.

The presence of errors in the results of analysis (objects assigned to the wrong groups) can be derived from the confusion matrix, or classification table, which is a square matrix, $G \times G$, where $G$ is the number of groups. It is a contingency table comparing the original assignment of objects to groups, usually in rows, to the group assignments made by the classification functions, in columns; each entry $n_{jg}$ represents the number of objects belonging to group $j$ and assigned to group $g$. Consequently, the diagonal elements ($n_{gg}$) represent the correctly classified objects, while the off-diagonal elements ($n_{jg}$) represent the objects erroneously classified. The confusion matrix is generally asymmetric, i.e. the number of objects belonging to group $j$ and assigned to group $g$ is not usually equal to the number of objects belonging to group $g$ and assigned to group $j$ (Ballabio and Todeschini, 2008). From this table, it is possible to determine the number and percentage of the objects correctly classified and to assess how the classification model is performing.

Clustering models make it possible to link the points representing the objects without requiring prior positioning in a graph (i.e. a metric space), which would be impractical in more than three dimensions.

2.3 Projecting data sets in a few dimensions (Legendre, 2012)

In multivariate statistics (multivariate as abbreviation for multidimensional random variable), ordination refers to the representation of objects as points along several reference axes, representing ordered relationships. The ordered relationships are usually quantitative, but it would suffice for them to be semi-quantitative relations to serve as the basis for ordinations. Several descriptors are observed for each object under study. In the multivariate approach, the scatter of objects is represented in a multidimensional diagram, with as many axes as there are descriptors in the study. For the purpose of analysis, the multidimensional scatter diagram may be projected onto bivariate graphs whose axes are chosen to represent a large fraction of the variability of the
multidimensional data matrix, in a space with reduced dimensionality relative to the
original data set. Methods for ordination in reduced space also allow one to derive
quantitative information on the quality of the projections and to study the relationships
among descriptors as well as among objects.
Several methods of data analysis try to reduce a large number of inter-correlated
variables to a smaller number of composite, but linearly independent variables, each
explaining a different fraction of the observed variation. The goal of numerical data
analysis is to generate a small number of variables, each explaining a large portion of
the variation, and to ascertain that these new variables explain different aspects of the
phenomena under study.
Ordination methods may sometimes be used to delineate clusters of objects. Ordinations
can also be used as complements to cluster analyses, which investigates pairwise
distances among objects, looking for fine relationships, whereas ordination in reduced
space brings out general gradients.
Generally, an ordination methods allow to study the relative positions of objects in
reduced space. An important aspect to consider is the representativeness of the
representation in reduced space, which usually has two or three dimensions. When the
projection in reduced space accounts for a high fraction of the variance, the distances
among projections of the objects in reduced space are quite similar to the original
distances in multidimensional space. When the projection is less efficient, the distances
among objects in reduced space are much smaller than in the original space; in this case,
the projection may be useless. However, even if the reduced space only accounts for a
small fraction of the variance, a projection in that space may still be informative (e.g.
when the objects are at proportionally similar distances in the two spaces).
Principal component analysis (PCA) orders objects in the multidimensional space of
descriptors using the Euclidean distance. With quantitative descriptors, PCA is a
powerful technique for ordination even in reduced space.

2.4 Principal axes (Legendre, 2012)
When analyzing sets of multidimensional data (N observations and K variables), only
methods that take into account the dependence among descriptors will generate proper
results. In order to characterize the proprieties of the observations, it is possible to
measure the variables. K descriptors can make up a K-dimensional random variable X
(x₁,..,xₖ). Each element xₖ (a vector) of the multidimensional variable is an
unidimensional random variable. Every descriptor $x_k$ is observed (its element $x_{nk}$) in each of the $N$ vectors “object”, each sampling unit $n$ ($x_{n1},..., x_{nk}$) providing a realization of the $K$-dimensional random variable.

When a $K$-dimensional random variable, whose $K$ dimensions (descriptors) are not independent, can be described by a multidimensional normal distribution, this multinormal distribution is specified by three sets of parameters: the vector of means, the diagonal matrix of standard deviations and the correlation matrix. The correlation matrix is the dispersion matrix of standardized variables. The dispersion matrix of $X$ contains the variances and covariances of the $K$ descriptors. In the unidimensional normal distribution, mean and standard deviation are the only parameters because there is no correlation for a single independent variable.

Bivariate normal distributions have a typical “bell” shapes. Moving down along the sides of the distribution, a series of “contour lines” are formed; this family of ellipsoids, called concentration ellipses for the two-dimensional normal distribution, have the multidimensional mean point as their common centre. When correlation is equal 0, the “bell” of probability densities is perfectly circular (in overhead view); as correlation increases, the “bell” of the probability densities flattens out, until it becomes unidimensional (e.g. a bisecting line) when correlation equals $\pm 1$. When the number of dimensions is higher than three, the family of concentration ellipses becomes a family of hyper-ellipsoids. The concentration hyper-ellipsoids of a distribution of objects pass through points of equal probabilities around the multinormal distribution. If a standardized variable is used, the family of hyper-ellipsoids is centered on the origin.

In a multinormal distribution, the first principal axis is the line that passes through the greatest dimension of the concentration hyper-ellipsoid describing the distribution; the next principal axes go through the next greatest dimensions of the $K$-dimensional ellipsoid. The $K$ consecutive axes are all perpendicular to one another in the hyperspace, and successively shorter. A maximum of $K$ principal axes may be derived from a data table containing $K$ variables. These principal axes are used as the basis for principal component analysis (PCA).

2.5 Principal component analysis (Legendre, 2012)

The principal axes of a dispersion (covariance) matrix $S$ are found by solving the equation $(S-\lambda_k I)\cdot u_k = 0$, where $u_k$ are the eigenvectors associated with the eigenvalues $\lambda_k$. There are as many eigenvalues as there are descriptors. The successive eigenvalues
account for progressively smaller fractions of the variance (the sum of variance in $S$ is equal to the sum of the eigenvalues). The eigenvectors are the principal axes of dispersion matrix $S$. The eigenvectors are normalized (i.e. scaled to unit length) before computing the principal components (which give the coordinates of the objects on the successive principal axes). The values of the $K$ elements $u_{jk}$ of each normalized eigenvector $u_k$ determine the rotation of the system of axes; they are *direction cosines*. Each direction cosine specifies the angle between an original Cartesian axis $j$ and a principal axis $k$. They indicate how the original variables “load” into (contribute to) principal components. The normalized eigenvector $u_k$ constitute the $K$ columns of square matrix $U$.

Since any dispersion matrix $S$ is symmetric, its principal axes $u_k$ are orthogonal to one another; they correspond to linearly independent directions in the concentration ellipsoid of the distribution of objects. The eigenvalues $\lambda_k$ of a dispersion matrix $S$ give the amount of *variance* corresponding to the successive principal axes. Because of these two properties, principal component analysis (PCA) can often summarize, in a few dimensions, most of the variability of a dispersion matrix of a large number of descriptors. It also provides a measure of the amount of variance explained by these few independent principal axes.

The elements of the eigenvectors $u_{jk}$ are also weights, or *loadings* of the original descriptors, in the linear combination of descriptors from which the principal components are computed. The *principal components* give the positions of the objects with respect to the new system of principal axes. The position of an object $x_i$ on the first principal axis is given by the following linear combination

$$f_{11} = (x_{i1} - \bar{x}_1)u_{11} + \ldots + (x_{ik} - \bar{x}_k)u_{k1} = [x - \bar{x}]_1^T u_1$$

where the values $(x_{ik} - \bar{x}_k)$ are the coordinates of object $x_i$ on the various centered descriptors $k$ and the values $u_{j1}$ are the loadings of the descriptors on the first eigenvector. The matrix $F$ of the transformed variables gives the positions of all objects with respect to the system of principal axes. It is also called the *matrix of principal component*, $F = X_c U$, where $X_c$ is the matrix of centred observations. The system of principal axes is centered with respect to the scatter of point-objects. Since the columns of the *matrix of component scores* $F$ are the coordinates of the objects with respect to the principal axes, they can be used to plot the objects with respect to principal axes.

Algebraically, the loadings inform how the variables are linearly combined to form the scores. The loadings unravel the *magnitude* (large or small correlation) and the *manner*
(positive or negative correlation) in which the measured variables contribute to the scores. (Ericsson et al., 2006a).

PCA finds lines, planes and hyperplanes in the $K$-dimensional space that approximate the data as well as possible in the least squares sense. A line or a plane that is the least squares approximation of a set of data points maximizes the variance of the coordinates on the line or plane (projection). The first principal component is the line in the $K$-dimensional space that best approximates the data in the least squares sense; it represents the maximum variance direction in the data. Each observation may be projected onto this line in order to get a score. The second principal component is also represented by a line in the $K$-dimensional space and is oriented such that it reflects the second largest source of variation in the data, while being orthogonal to the first principal component; it improves the approximation of $X$ data as much as possible. The third principal component is oriented in the direction perpendicular to the already existing model plane. Frequently, the descriptive ability of the PCA model improves by using more principal components. They are useful to adequately summarize the information in the data set (Ericsson et al., 2006a).

Principal component axes, also called principal axes, constitute a new system of coordinates. The principal components are the new variates (random variables) specified by the axes of a rigid rotation of the original system of coordinates, and corresponding to the successive directions of maximum variance of the scatter of points. The objects are positioned by the principal components in the same way as in the original system of descriptor-axes. PCA simply rotates the axes in such a way that the new axes correspond to the main components of variability. The rotation of the system of descriptor-axes is performed, in multidimensional space, using the centroid of the objects as pivot. Therefore, after the analysis (rotation), the original descriptor-axes are still orthogonal. The relative positions of the objects in the rotated $K$-dimensional space of principal components are the same as in the $K$-dimensional space of the original descriptors. This means that the Euclidean distances among objects have been preserved through the rotation of axes. The angles between descriptor-axes and principal axes are projections of the rotation angles corresponding to the elements of orthonormal matrix $U$. The principal components can define a reduced hyperspace. The first two principal components define the plane allowing the representation of the largest amount of variance. The objects are projected on that plane in such a way as to preserve, as much
as possible, the relative Euclidean distances they have in the multidimensional space of the original descriptors.

The plots of PCA scores show an overview of the investigated data set, which may reveal groups of observations, how the observations are related and if there are outliers. Outliers are observations that are extreme or that do not fit the model. Strong outliers have strong power to pull the PCA model toward themselves and may “consume” one principal component just because of their existence (Ericsson et al., 2006a).

When the descriptors are all of the same type and order of magnitude, and have the same units, a dispersion matrix is used to compute PCA. When the descriptors are of different types or orders of magnitude, or have different units, it is necessary to conduct PCA on the correlation matrix instead of the dispersion matrix. Principal components extracted from correlation matrices are not the same as those computed from dispersion matrices, because the distances among objects are not the same in the two analyses. In the case of correlations, the descriptors are standardized. It follows that the distances among objects are independent of the measurement units, whereas those in the space of the original descriptors vary according to measurement scales.

The PCA model is calculated to approximate the observations as well as possible. It finds the directions in multivariate space that represent the largest sources of variation. However, it is not necessarily the case that these maximum variation directions coincide with the maximum separation directions among the groups. Knowledge related to group membership is not used to find the location of the principal components. It may be that other directions are more pertinent for discriminating among groups of observations (Ericsson et al., 2006a).

### 2.6 Clustering and ordination (Legendre, 2012)

Cluster analysis differs from PCA in that the goal is to detect similarities between objects and find groups in the data on the basis of calculated distances, whereas PCA does not focus on how many groups will be found. Both PCA and cluster analysis do not use information related to predefined classes (or groups) of objects (Ballabio and Todeschini, 2008). A commonly-used non-statistical method to validate the results of cluster analyses is to plot the clusters onto an ordination diagram and look for separation of the clusters.

Ordination is used when gradients are sought, and clustering when one is looking for a partition of the objects into subsets. When the gradient results from the combined action
of several descriptors, the ordination must be carried out in a reduced space, using f. e. PCA. It may also happen that an ordination is used as a basis for visual clustering. The combined use of clustering and ordination may optimize the partition of the objects. In the full multidimensional ordination space of PCA, Euclidean distances among the main clusters of objects are the same as in the original space of descriptors. However, when only the first two or three dimensions are considered, ordinations in reduced space may misrepresent the structure by projecting together clusters of objects which are distinct in higher dimensions. Clustering methods allow one to separate clusters whose projections in reduced space may sometimes obscure the relationships among them. It is recommended to simultaneously carry out clustering and ordination on a set of objects and to represent clustering results onto ordination diagrams. Description of the data structure is clearer when the clustering results are drawn onto the ordination.
Chapter 3

**REGRESSION**

### 3.1 Forecasting and prediction (Legendre, 1998)

A distinction has to be made between forecasting and prediction in ecology. *Forecasting models* (or correlative models) extend, into the future or to different situations, structural relationships among descriptors that have been quantified for a given data set. A set of relationships among variables, which simply describe the changes in one or several descriptors in response to changes in others as computed from a ‘training set’, make up a *forecasting* model. In contrast, when the relationships are considered causal and to describe a mechanistic process, the model is *predictive*. A condition to successful forecasting is that the values of all important variables that have not been observed be about the same in the new situation as they were during the survey (or experiment). In addition, forecasting does not allow extrapolation beyond the observed range of the explanatory variables. In contrast, predictive models describe known or assumed causal relationships. They allow one to estimate the effects, on some variables, of changes in other variables. Multiple linear regression, regression on principal components, identification functions in discriminant analysis and direct comparison (canonical analysis) are numerical methods to forecast one or several descriptors (response or dependent variables) using other descriptors (explanatory or independent variables).

Contrary to the forecasting models, *predictive models* allow one to foresee how some variables of interest would be affected by changes in other variables. Prediction is possible when the model is based on causal relationships among descriptors (*i.e.* not only correlative evidence). Causal relationships are stated as hypotheses (theory) for modeling; they may also be validated through experiments in the laboratory or in the field. One may hypothesize that there exist causal relationships among the observed descriptors or, alternatively, that the observed descriptors are caused by *underlying hidden variables*. Depending on the hypothesis, the methods for analyzing causal relationships are not the same.

Techniques of forecasting may be used for predictive purposes when there are reasons to believe that the relationships between explanatory and response variables are of causal nature.
3.2 Linear regression (Legendre, 1998)
Regression analysis is a type of modeling method. The mathematical model is an algebraic formulation of a set of relationships among variables, whose parameters have to be estimated. The purpose of regression analysis is to find the best functional model relating a response random variable \(y\) to several explanatory variables \(x_1, x_2, \ldots, x_k\), in order to forecast or predict the values of the response variable for given values of the independent variables. Forecasting (or prediction) consists in calculating values of the response variable using a regression equation. The objective is achieved by using the equation that minimizes the residual mean square error, or maximizes the coefficient of determination \(r^2\) in simple regression, \(R^2\) in multiple regression.
Linear regression is used to compute the parameters of a linear equation relating a response variable \(y\) and a single explanatory variable \(x\). The model for simple linear regression has the form:
\[
\hat{y} = b_0 + b_1 x
\]
This corresponds to the equation of a straight line that crosses the scatter of points in some optimal way and allows the computation of an estimated value \(\hat{y}\) for any value of \(x\). Parameter \(b_1\) gives the slope of the regression line and it is the regression coefficient, while parameter \(b_0\) is the \(y\)-intercept.
The difference between the observed \((y_i)\) and estimated \((\hat{y}_i)\) values along \(y\), \(e_i = y_i - \hat{y}_i\) for every observation \(i\), is the residual value of observation \(y_i\) after fitting the regression line. \(y_i\) is equal to the value \(\hat{y}_i\) predicted by the regression equation plus the residual \(e_i\)
\[
y_i = \hat{y}_i + e_i = b_0 + b_1 x + e_i
\]
In simple linear regression, one is looking for the straight line with equation \(\hat{y} = b_0 + b_1 x\) that minimizes the sum of squares of the vertical differences residuals, \(e_i\), between the observed values \((y)\) and values calculated using the regression equation \((\hat{y})\). This is the principle of least squares, first proposed by the mathematician Adrien Marie Le Gendre (in 1805) and later by Karl Friedrich Gauss (in 1809). Least-squares estimation provides the line of best fit for parameter estimation and forecasting when the explanatory variable is controlled.
Ordinary least squares (OLS) method produces fitted values with the smallest error (sum of squared residuals, \(\sum(y_i - \hat{y}_i)^2\)). It should be used when the purpose of regression analysis is simply to forecast or predict values of \(y\) for given \(x\)’s, or when \(x\) is
a random variable measured without error or if the magnitude of the error on \( y \) is known to be much larger than that on \( x \) (more than three times) (McArdle, 1988).

When there are several explanatory variables \( x_1, x_2, \ldots, x_K \), it is possible to compute a regression equation where the response variable \( y \) is a linear function of all explanatory variables \( x_j \). The *multiple linear regression* (MLR) model is a direct extension of simple linear regression:

\[
y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_K x_{iK} + \varepsilon_i
\]

for object \( i \). That equation leads to the known formula for the fitted values

\[
\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_K x_{iK}
\]

When there are more samples than variables; this does not allow an exact solution for the vector of regression parameters \( \mathbf{b} (= [b_j]) \), but it is possible to get a solution by minimizing the length of the residual vector \( (\varepsilon = y - X\mathbf{b}) \) (Geladi and Kowalski, 1986), where \( X \) (\( K \) column vectors: \( x_1, \ldots, x_K \)) is the matrix of explanatory variables and \( y \) is the column vector of response variable. The vector \( \mathbf{b} \) is computed using ordinary least squares (OLS) method (Draper and Smith, 1981). In order to obtain a least-squares best fit, each member (left and right) of matrix equation \( y = X\mathbf{b} \) is multiplied by the transpose of matrix \( X \), i.e. \( X^T y = X^T X \mathbf{b} \). By doing so, the rectangular matrix \( X \) produces a square matrix \( X^T X \), which can be inverted. The values of coefficients \( b_0, b_1, \ldots, b_K \) of a multiple linear regression are computed directly after inverting the square matrix \( X^T X \):

\[
\mathbf{b} = (X^T X)^{-1} X^T y
\]

If an intercept \( (b_0) \) must be estimated, a column of 1’s is added to matrix \( X \) of the explanatory variables.

When the same multiple regression model is to be computed for several response variables in turn \( (y_1, \ldots, y_M) \), regression coefficients can be estimated by ordinary least squares for all the response variables simultaneously, using a single matrix expression:

\[
\mathbf{B} = (X^T X)^{-1} X^T \mathbf{Y}
\]

In this expression, which is the multivariate equivalent of \( \mathbf{b} = (X^T X)^{-1} X^T y \), \( \mathbf{Y} \) (\( M \) column vectors: \( y_1, \ldots, y_M \)) is the matrix of response variables and \( \mathbf{B} \) is the matrix of regression coefficients of all response variables \( \mathbf{Y} \) on the regressors \( \mathbf{X} \). The coefficients found using this equation are the same as those obtained from multiple regressions computed in separate runs for each response variable. The procedure of computing all linear regressions simultaneously has been called *multivariate linear regression* (Finn, 1974).

In multiple regression, the fitted values \( \hat{\mathbf{y}} \) are computed as

\[
\hat{\mathbf{y}} = \mathbf{X} \mathbf{B} = \mathbf{X} (X^T X)^{-1} X^T \mathbf{Y}
\]
If the number of objects \((N)\) and the number of independent variables \((K)\) are equals, \(\mathbf{X}\) is square; in that case, the multiple regressions always explain the variables in matrix \(\mathbf{Y}\) entirely, so that \(\hat{\mathbf{Y}} = \mathbf{Y}\). This condition is not encountered often in practical situations.

Ordinary regression coefficients \((b)\) are useful when the regression equation is to be used to compute estimated values of \(y\) for objects that have not been used to estimate the parameters of the regression equation and for which observed \(x\) values are available. This is the case when a regression model is validated using a new set of observations: estimates \(\hat{y}\) are computed from the regression equation to be validated, using the observed values of the explanatory variables \(x_j\), and they are compared to the observed \(y\)’s to assess how efficient the regression model is at calculating \(y\) for new data.

Ordinary regression coefficients in multiple linear regression are partial regression coefficients. The term ‘partial’ means that each regression coefficient is a measure of the rate of change variable \(y\) would have per unit of variable \(x_j\), if all the other explanatory variables were held constant.

When the explanatory variables \(x_j\) of the model are uncorrelated, multiple regression is a straightforward extension of simple linear regression. With observational data, however, the explanatory variables used in multiple regression models are most often collinear \((i.e.\) correlated to one another) and strong collinearity may affect the ability to correctly estimate model parameters.

In forecasting, the objective is to maximize the coefficient of multiple determination \((R^2\) in multiple regression); collinearity of the explanatory variables is not a concern. For description, the interest is to correctly estimate the parameters of the model; the effect of multicollinearity on the estimates of the model parameters must be minimized.

The matrix to be inverted, \(\mathbf{X}^\top\mathbf{X}\), may be either singular, and hence not invertible, or ill-conditioned and close to singular, which results in a numerically unstable inverse matrix. In this case, the regressor variables cannot be independent (Krzanowski, 1988). Collinearity, zero determinant and singularity are all names for the same problem: the inverse of matrix \(\mathbf{X}^\top\mathbf{X}\) may not exist.

### 3.3 Principal Component Regression (PCR) (Legendre, 1998)

When the explanatory variables of the model are orthogonal to one another (no collinearity), it is easy to eliminate any parameter that does not significantly contribute to the model. Any variable whose contribution is not statistically significant can be removed from the model. However, the task is not that simple with observational data,
which often display various degrees of collinearity. The problem is that significance may get diluted among collinear variables contributing in the same way to the explanation of a response variable \( y \).

Some statistical programs offer procedures that allow one to test all possible models with \( K \) explanatory variables and select the one where \( R^2 \) is the highest and where all partial regression coefficients are significant. When such a procedure is not available, one of the methods that have been developed for selecting the best subset of explanatory variables may be used.

There are some ways to counter the effects of multicollinearity in multiple regression. Collinearity has the effect of inflating the variance of regression coefficients (which leads to better predictions of the response variable), with the exception of the intercept \( b_0 \). The price to pay for reducing the inflation of variance is some bias in the estimates of the regression coefficients. Regression on principal components is helpful when the objective is forecasting or prediction.

**Major axis regression** (MA) models primarily describe the first principal component of a bivariate normal distribution. In major axis regression, the estimated regression line is the first principal component of the scatter of points. The quantity which is minimized is the sum, over all points, of the squared Euclidean distances between the points and the regression line, instead of vertical distances as in OLS (Fig. 3.1).

![Figure 3.1: (a) Vertical deviations (OLS) and (b) Euclidean distances to the regression line (MA) (Legendre, 1998).](image)

The slope of the major axis is then estimated; also the slope of the minor axis of the bivariate scatter of points (the second principal component) can be estimate. Just as with principal component analysis, this method is mostly useful in situations where both variables are expressed in the same physical units or are dimensionless (naturally or after standardization).

**Regression on principal components** consists of a few steps: 1) Perform a principal component analysis on the matrix of the explanatory variables \( X \). The data matrix \( X \) can
be represented by its matrix of component scores, \( \mathbf{T} = \mathbf{X} \mathbf{P} \) (where dimensions having small eigenvalues are excluded). \( \mathbf{P} \) matrix is made up of the \( \mathbf{p}_a \) vectors and \( \mathbf{T} \) matrix of the \( \mathbf{t}_a \) vectors as columns, where \( a=1,...,A \). The elements of the \textit{loading} vector \( \mathbf{p}_a \) are the direction cosines; the elements of the \textit{scores} vector \( \mathbf{t}_a \) are the coordinates of the respective points on the principal component line. The scores and loadings can be calculated pair-by-pair by an iterative procedure. \textit{Nonlinear Iterative PARTial Least Squares} (NIPALS) algorithm does not calculate all the principal components at once. It calculates \( \mathbf{t}_1 \) and \( \mathbf{p}_1^T \) from the \( \mathbf{X} \) matrix (\( \mathbf{X} = \mathbf{TP}^T \)). Then the outer product, \( \mathbf{t}_1 \mathbf{p}_1^T \), is subtracted from \( \mathbf{X} \) and the residual \( \mathbf{E}_1 \) is calculated (\( \mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1 \mathbf{p}_1^T \)). This residual can be used to calculate \( \mathbf{t}_2 \) and \( \mathbf{p}_2^T \), then \( \mathbf{E}_2 = \mathbf{E}_1 - \mathbf{t}_2 \mathbf{p}_2^T \). In general, \( \mathbf{E}_a = \mathbf{E}_{a-1} - \mathbf{t}_a \mathbf{p}_a^T \) (peeling off) and \( \mathbf{E}_A = \mathbf{X} - \mathbf{t}_1 \mathbf{p}_1^T - \mathbf{t}_2 \mathbf{p}_2^T - \cdots - \mathbf{t}_A \mathbf{p}_A^T \). It has been shown that on convergence the NIPALS solution is the same as that calculated by the eigenvector \textit{formulae}. The NIPALS method is convenient for microcomputers (Geladi and Kowalski, 1986). 2) Compute the multiple regression of \( \mathbf{y} \) on the principal components instead of the original explanatory variables. The MLR formula can be written as \( \mathbf{Y} = \mathbf{TB} + \mathbf{E} \) (solution: \( \mathbf{B} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{Y} \)). The variables of \( \mathbf{X} \) are replaced by new ones that are orthogonal and also span the multidimensional space of \( \mathbf{X} \). The inversion of \( \mathbf{T}^T\mathbf{T} \) should give no problem because of mutual orthogonality of the scores (Geladi and Kowalski, 1986). 3) Find back the contributions of the explanatory variables by multiplying matrix of the eigenvectors with the matrix of regression coefficients \( \mathbf{B} \) of \( \mathbf{Y} \) on the principal components. One obtains a new matrix \( \mathbf{B}' \) of contributions of the original variables to the regression equation: \( \mathbf{P}_{(KxA)} \mathbf{B}_{(AxM)} = \mathbf{B}'_{(KxM)} \), where \( A \) is the number of explanatory variables in the analysis, \( M \) is the number of response variables, \( A \) is the number of principal components retained for step (3).

This procedure does not necessarily solve the problem of multicollinearity, although it is true that the regression is performed on principal components which are not collinear by definition. There is gain in stability of the regression coefficients only by eliminating some of the principal components from the previous equation. One may either eliminate the eigenvectors with the smallest eigenvalues or, better, use only the principal components that significantly contribute to explain \( \mathbf{y} \). By doing so, the regression coefficient estimates become biased. Belsley (Belsley \textit{et al.}, 1980) states that score vectors corresponding to small eigenvalues can be left out in order to avoid collinearity problems from influencing the solution.
PCR solves the collinearity problem (by guaranteeing an invertible matrix in the calculation of \( \mathbf{B} \)) and the ability to eliminate the lesser principal components allows some noise (random error) reduction. However, PCR is a two-step method and thereby has the risk that useful (predictive) information will end up in discarded principal components and that some noise will remain in the components used for regression (Geladi and Kowalski, 1986).

3.4 Correlation coefficient (Legendre, 1998)

The covariance provides information on the orientation of the cloud of data points in the space defined by the descriptors. The covariance measures the joint dispersion of two random variables around their bivariate means.

In order to calculate covariances on different descriptors, they must be brought to some common scale. The most widely used method for making descriptors compatible is to standardize the data (z-scores). Standardization is achieved by subtracting the mean (translation, used to bring to zero the mean) and dividing by the standard deviation of the variable (expansion, a change of scale):

\[
    z_i = \frac{x_i - \bar{x}}{s_x}
\]

where \( \bar{x} \) is the mean of the observations from each value \( x_i \), and \( s_x \) is the estimate of standard deviation. The position of each object on the transformed variable \( z_i \) is expressed in standard deviation units; as a consequence, it refers to the group of objects from which the standard deviation has been estimated. The standardized variable \( z_i \) has some interesting proprieties: its mean is zero; its variance is 1; it is a dimensionless variable. Thus, the standardized descriptors have the same scale and are dimensionless.

The covariance of two standardized descriptors is called the coefficient of linear correlation (Pearson r). The correlation is a measure of the dependence between two random variables.

The coefficient of linear correlation between two variables, \( x \) and \( y \), is

\[
    r = \frac{s_{xy}}{s_x s_y}
\]

where \( s_{xy} \) is the estimate of the covariance of \( x \) and \( y \), \( s_x \) and \( s_y \) are the estimates of standard deviation of \( x \) and \( y \) respectively.

In a multidimensional contest, the multiple linear correlation measures the intensity of the relationship between a response variable and a linear combination of several explanatory variables.
3.5 Coefficient of determination (Legendre, 1998)

With two random variables \( (x \text{ and } y) \), for simplicity, the coefficient of determination \( r^2 \) measures how much of the variance of each variable is explained by the other. The amount of explained variance for \( y \) is the variance of the fitted values \( \hat{y}_i \). It is calculated as

\[
s_y^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}
\]

whereas the total amount of variation in variable \( y \) is

\[
s_y^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}
\]

\( r^2 \) is the ratio of these two values. It can be shown that the coefficient of determination is equal to the square of the Pearson correlation coefficient \( r \).

The coefficient of multiple determination \( R^2 \) is the extension, to the multidimensional case, of the coefficient of determination between two variables. Even the coefficient of multiple determination is the square of the multiple correlation coefficient \( R \); it varies between 0 and 1. \( R^2 \) is the most commonly used statistic to measure the forecasting potential of a multiple regression equation. As in simple linear regression, it is the ratio of the sum of squares of distances of the estimates values \( \hat{y} \) to the mean \( \bar{y} \), to the sum of squares of distances of the original response variable values \( y \) to the mean \( \bar{y} \):

\[
R^2 = \frac{\sum(y_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}
\]

The coefficient of multiple determination \( R^2 \) measures the fraction of the variance of the response variable \( y \) that is explained by a linear combination of the explanatory variables. It detects the proportion of the variation of \( y \) about its mean which is explained by the regression equation.
Chapter 4

PARTIAL LEAST SQUARES (PLS) REGRESSION

4.1 General considerations

Partial Least Squares (PLS) is a wide class of methods for modeling relations between sets of observed variables by means of latent variables. It comprises of regression and classification tasks as well as dimension reduction techniques and modeling tools. 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. Projections of the observed data to its latent structure by means of PLS was developed by Herman Wold (Wold H., 1982, 1985; Wold S. et al., 1984).

In its general form, PLS creates orthogonal score vectors (also called latent vectors, LV, or factors) by maximizing the covariance between different sets of variables. There are different PLS techniques to extract latent vectors, and each of them gives rise to a variant of PLS.

PLS can be naturally extended to regression problems. The predictor and predicted (response) variables are each considered as a block of variables. PLS extracts the score vectors which serve as a new predictor representation and regresses the response variables on these new predictors. The natural asymmetry between predictor and response variables is reflected in the way in which score vectors are computed. This variant is known under the names of PLS1 (1 response variable) and PLS2 (at least 2 response variables) (Rosipal and Krämer, 2006). PLS2 is a version of the PLS method in which several Y-variables are modeled simultaneously, thus taking advantage of possible correlations or collinearity between Y-variables.

PLS Regression (PLS-R) (Wold S. et al., 1984; Geladi and Kowalski, 1986; Höskuldsson, 1996) is related to both PCR and MLR. PCR finds factors that capture the greatest amount of variance in the predictor variables (X). MLR seeks to find a single factor that best correlates predictor variables (X) with predicted variables (Y). MLR works well as long as the X-variables are fairly few and X matrix has full rank. PLS attempts to find factors which both capture variance and achieve correlation. PLS attempts to maximize covariance (McLennan and Kowalski, 1995) and this is the explicit objective of the SIMPLS algorithm for PLS (de Jong, 1993).
There are several ways to calculate PLS model parameters. The most intuitive is known as NIPALS (Non-linear Iterative PArtilial Least Squares); this led to the acronym PLS for these models. NIPALS calculates scores \( T \) and loadings \( P \) (similar to those used in PCR), and an additional set of vectors known as weights, \( W \) (with the same dimensionality as the loadings \( P \)). The addition of weights in PLS is required to maintain orthogonal scores (McLennan and Kowalski, 1995).

Both the NIPALS and SIMPLS algorithms for PLS also work when there is more than one predicted variable (\( Y \)). In such cases of multiple \( Y \)-variables (PLS2), scores \( U \) and loadings \( Q \) matrices are also calculated for the \( Y \)-block. A vector of ‘inner-relationship’ coefficients, \( d \), which relate the \( X \)- and \( Y \)-block scores, must also be calculated. Using NIPALS, the scores, weights, loadings and inner-coefficients are calculated sequentially (McLennan and Kowalski, 1995).

### 4.2 Partial Least Squares (PLS) Algorithm

PLS models the relations between two blocks of variables (\( X \) and \( Y \)) by means of score vectors (Fig. 4.1). The PLS model is developed from a training set of \( N \) statistical units (objects) with \( K \) \( X \)-variables, denoted by \( x_k \) \((k=1,...,K)\), and \( M \) \( Y \)-variables, \( y_m \) \((m=1,...,M)\). These training data form the two matrices \( X \) \((N\times K)\) and \( Y \) \((N\times M)\). PLS decomposes these two matrices into the form (outer relations)

\[
X = TP^T + E \\
Y = UQ^T + F
\]

where \( T \) and \( U \) are \((N\times A)\) matrices of the \( A \) extracted score vectors (latent vectors), \( P \) \((K\times A)\) and \( Q \) \((M\times A)\) represent matrices of loadings, \( E \) \((N\times K)\) and \( F \) \((N\times M)\) are the matrices of residuals.

The PLS method, which in its classical form is based on the NIPALS algorithm (Wold H., 1975), finds weight vectors \( w \), \( c \) such that

\[
[cov(t, u)]^2 = [cov(Xw, Yc)]^2 = max_{\|r\|=\|s\|=1}[cov(Xr, Ys)]^2
\]

where \( cov(t, u) = t^T u / N \) denotes the sample covariance between the score vectors \( t \) and \( u \) (Rosipal and Krämer, 2006). The NIPALS algorithm starts with random initialization of the \( Y \)-score vector \( u \), usually one of the \( Y \) columns, and repeats a sequence of steps until convergence:

\[
w = X^T u / u^T u, \quad \|w\| = 1, \quad t = X w \\
c = Y^T t / t^T t, \quad \|c\| = 1, \quad u = Y c
\]
With a single y-variable, $M = 1$, $u = y$ and the NIPALS procedure converges in a single iteration.

In general, PLS is an iterative process. Each model parameter is iteratively estimated as the slope of a simple bivariate regression (least squares) between a matrix column or row as the y-variable, and another parameter vector as the x-variable. For instance, the PLS weights, $w$, are iteratively re-estimated as $X^T u / u^T u$. The ‘partial’ in PLS indicates that this is a partial regression, since the x-vector ($u$ above) is considered as fixed in the estimation (Wold S. et al. 2001).

After the extraction of the score vectors, $t$ and $u$, the matrices $X$ and $Y$ are deflated by subtracting their rank-one approximations based on $t$ and $u$. Different forms of deflation define several variants of PLS. Using equations $X = TP^T + E$ and $Y = UQ^T + F$, the vectors of loadings $p$ and $q$ are computed as coefficients of regressing $X$ on $t$ and $Y$ on $u$, respectively:

$$p = X^T u / (t^T t) \text{ and } q = Y^T u / (u^T u).$$

The loading vectors $p$ are not orthogonal to each other and neither are the Y-scores, $u$ (Wold S., 2001).

PLS1 and PLS2 (variants of PLS) are the most frequently used PLS approaches. The relationship between $X$ and $Y$ is asymmetric. The X-extracted score vectors $t_a$ ($a=1,...,A$) are “few” and orthogonal; $A$ denotes the number of extracted score vectors / PLS iterations. Some assumptions are made:

- the X-scores are, multiplied by the loadings, good “summaries” of $X$, so that the X-residual, $E$, is ‘small’;
- the X-scores are good predictors of $Y$:
  $$Y = TC^T + F^*$$
  where the elements of the Y-residual matrix, $F^*$, express the deviations between the observed and modeled responses;
- a linear inner relation between the scores vectors $t$ and $u$ exists:
  $$U = TD + H$$
where $D$ is a diagonal $(A \times A)$ matrix ($\hat{u} = d \cdot t$ and $d = \frac{u^T t}{t^T t}$) and $H$ denotes the matrix of residuals.

Both $Y$ and $X$ are assumed to be, at least partly, modeled by the same LV’s. The asymmetric assumption of the predictor–predicted variables relation is transformed into a deflation scheme where the $X$ score vectors are good predictors of $Y$. The extracted score vectors are then used to deflate $Y$, that is, a component of the regression of $Y$ on $t$ is removed from $Y$ at each iteration of PLS (Rosipal and Krämer, 2006)

$$X = X - t \cdot p^T \quad \text{and} \quad Y = Y - t \cdot c^T$$

where $c = Y^T t / t^T t$ are not scaled to unit norm weight vectors. This deflation scheme guarantees mutual orthogonality of the extracted score vectors $t_a$ ($a = 1, \ldots, A$) (Höskuldsson, 1988). The deflated matrices (the residuals) are used as $X$ and $Y$ in the next component. The procedure continues with next component until cross-validation indicates that there is no more significant information in $X$ about $Y$ (Wold S., 2001). The first $A$ dominant factors capture most of the variance in $X$ and have maximum ability for predictive models. The NIPALS algorithm is numerically and statistically stable.

In the iterative algorithm, the $X$ and $Y$ blocks get each other’s scores, this gives a better inner relation (linking both blocks) (Geladi and Kowalski, 1986). Approximately, the algorithm is the following (Garson, 2010):

‘The components of $X$ are used to predict the scores on the $Y$ components, and the predicted $Y$ component scores are used to predict the actual values of the $Y$-variables. In constructing the principal components of $X$, the PLS algorithm iteratively maximizes the strength of the relation of successive pairs of $X$ and $Y$ component scores by maximizing the covariance of each $X$-score with the $Y$-variables. This strategy means that while the original $X$-variables may be multicollinear, the $X$ components used to predict $Y$ will be orthogonal’.

PLS is sometimes called Projection to Latent Structures because of its general strategy.

4.3 PLS Regression model

PLS-regression (PLS-R) provides a method for relating two data matrices, $X$ and $Y$, by a linear multivariate model, but it models also the structure of $X$ and $Y$. PLS-R can analyze data with many, noisy, collinear (correlated), and even incomplete variables in both $X$ and $Y$. PLS-R has the desirable property that the precision of the model
parameters improves with the increasing number of relevant variables and observations (Wold S., 2001).

Before the analysis, the X- and Y-variables are often transformed to make their distributions fairly symmetrical. Results of projection methods, such as PLS-R, depend on the scaling of the data. With an appropriate scaling, one can focus the model on more important Y-variables, and use experience to increase the weights of more informative X-variables. In the absence of knowledge about the relative importance of the variables, the standard multivariate approach is to scale each variable to unit variance (by dividing them by their SD’s) and to center them (by subtracting their averages), so-called auto-scaling. This corresponds to giving each variable (column) the same weight, the same prior importance in the analysis. Auto-scaling is often not the best scaling of X, but non-scaled X-data or some intermediate between auto-scaled and non-scaled can be appropriate (Wold S. et al., 1993). In process data, the acceptable interval of variation of each variable can form the basis for the scaling. For ease of interpretation and for numerical stability, it is recommendable to center the data before the analysis. This is done, either before or after scaling, by subtracting the averages from all variables both in X and Y. In process data other reference values such as set point values may be subtracted instead of averages. Hence, the analysis concerns the deviations from the reference points and how they are related. The centering leaves the interpretation of the results unchanged (Wold S., 2001).

PLS1 and PLS2 can be used to solve linear regression problems. Combining assumption of a linear relation between the scores vectors t and u (U = TD +H) with the decomposition of the Y matrix, equation \( Y = UQ^T + F \) can be written as \( Y = TDQ^T + HQ^T + F \) (mixed relation). This defines the equation

\[
Y = TC^T + F^*
\]

where \( C^T=HQ^T \) denotes the \((A \times M)\) matrix of regression coefficients and \( F^* = HQ^T + F \) is the residual matrix. Equation \( Y = TC^T + F^* \) is the decomposition of Y using ordinary least squares regression with orthogonal predictors T.

It is appropriate to consider orthonormalised score vectors t, that is \( T^TT = I \), and the matrix \( C = Y^T T (Y = TC^T \rightarrow Y^T = CT^T \rightarrow Y^T T = CT^T T = CI) \) of the not scaled to length one weight vectors c (Rosipal and Krämer, 2006). To redefine equation \( Y = TC^T + F^* \) in terms of the original predictors X, it is helpful to use the relationship (Manne, 1987)

\[
T = XW^* = XW(P^TW)^{-1}
\]
where $\mathbf{P}$ is the matrix of loading vectors and $\mathbf{W}$ is the matrix of weight vectors ($\mathbf{t}_a = \mathbf{E}_{a} \mathbf{w}_a$ and $\mathbf{t}_1 = \mathbf{X} \mathbf{w}_1$). The vectors $\mathbf{w}$ form an orthonormal set (weights are necessary in order to obtain orthogonal $\mathbf{X}$-scores). Plugging this relation into $\mathbf{Y} = \mathbf{TC}^\mathsf{T} + \mathbf{F}^*$, it is

$$
\mathbf{Y} = \mathbf{XW}(\mathbf{P}^\mathsf{T}\mathbf{W})^{-1}\mathbf{C}^\mathsf{T} + \mathbf{F}^* = \mathbf{XB} + \mathbf{F}^*
$$

where $\mathbf{B}$ represents the matrix of regression coefficients

$$
\mathbf{B} = \mathbf{W}(\mathbf{P}^\mathsf{T}\mathbf{W})^{-1}\mathbf{C}^\mathsf{T} = \mathbf{X}^\mathsf{T}\mathbf{U}(\mathbf{T}^\mathsf{T}\mathbf{XX}^\mathsf{T}\mathbf{U})^{-1}\mathbf{T}^\mathsf{T}\mathbf{Y}
$$

For the last equality, the relations among $\mathbf{T}$, $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{P}$ are used (Rännar et al., 1994). Different scalings of the individual score vectors $\mathbf{t}$ and $\mathbf{u}$ do not influence the $\mathbf{B}$ matrix. The weight matrix $\mathbf{W}$ is obtained to make $\|\mathbf{F}^*\|$ as small as possible and, at the same time, to derive an useful relation between $\mathbf{X}$ and $\mathbf{Y}$ (Chong and Jun, 2005). Only square matrices have unique inverses. $\mathbf{T}$ is a $(N\times A)$ matrix with $N > A$ and its columns are orthonormal ($\mathbf{T}$ is a full rank matrix); then $\mathbf{T}$ has a left inverse (left inverse is given by $\mathbf{A}^{-1}_{\text{left}} = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}$, i.e. $\mathbf{A}^{-1}_{\text{left}} \mathbf{A} = \mathbf{I}$) and $\mathbf{T}^\mathsf{T}\mathbf{T} = \mathbf{I}$, but $\mathbf{T}\mathbf{T}^\mathsf{T} \neq \mathbf{I}$ (Kolter, 2012).

For training data, the estimate of PLS regression is

$$
\hat{\mathbf{Y}} = \mathbf{X} \cdot \mathbf{B} = \mathbf{X} \cdot \mathbf{X}^\mathsf{T}\mathbf{U}(\mathbf{T}^\mathsf{T}\mathbf{XX}^\mathsf{T}\mathbf{U})^{-1}\mathbf{T}^\mathsf{T}\mathbf{Y} = \mathbf{T} \cdot \mathbf{T}^\mathsf{T}\mathbf{Y} = \mathbf{T} \cdot \mathbf{C}^\mathsf{T}
$$

(to test the equality: $\mathbf{T}^\mathsf{T}\mathbf{X} \cdot \mathbf{X}^\mathsf{T}\mathbf{U}(\mathbf{T}^\mathsf{T}\mathbf{XX}^\mathsf{T}\mathbf{U})^{-1} = \mathbf{I}$ and $\mathbf{T}^\mathsf{T}\mathbf{T} = \mathbf{I}$, then $\mathbf{XX}^\mathsf{T}\mathbf{U}(\mathbf{T}^\mathsf{T}\mathbf{XX}^\mathsf{T}\mathbf{U})^{-1} = \mathbf{I}$).

For testing data, it is

$$
\hat{\mathbf{Y}}_t = \mathbf{X}_t \cdot \mathbf{B} = \mathbf{X}_t \cdot \mathbf{X}^\mathsf{T}\mathbf{U}(\mathbf{T}^\mathsf{T}\mathbf{XX}^\mathsf{T}\mathbf{U})^{-1}\mathbf{T}^\mathsf{T}\mathbf{Y} = \mathbf{T}_r \cdot \mathbf{C}^\mathsf{T}
$$

where $\mathbf{X}_t$ and $\mathbf{T}_r = \mathbf{X}_t \mathbf{X}^\mathsf{T}\mathbf{U}(\mathbf{T}^\mathsf{T}\mathbf{XX}^\mathsf{T}\mathbf{U})^{-1}$ represent the matrices of testing data and score vectors, respectively (Rosipal and Krämer, 2006).

PLS regression coefficients offer a compact representation of the $\mathbf{X}$-$\mathbf{Y}$ relations and simplify the model interpretation when there are several components (> 4÷5). Their advantage is to obtain only one vector of concise model information per response, rather than several vectors of weights. The disadvantage of the coefficients is that information regarding the correlation structure among the responses is lost. This information is preserved by the PLS weights (Ericsson et al., 2006a).

The coefficients $c_{m,a}$’s are estimated by multiple regression of $\mathbf{y}_m$ on the PLS components $\mathbf{t}_a$’s ($\mathbf{t}_1, ..., \mathbf{t}_A$). If $\mathbf{t}_a$ is strongly correlated with $\mathbf{y}_m$, and $\mathbf{x}_k$ is important when building up $\mathbf{t}_a$, then the $\mathbf{x}_k$ will be important to $\mathbf{y}_m$ (Zhang et al., 2006). The idea is reflected in the concept of variable importance in projection for the $k$-th variable $\mathbf{x}_k$, VIP$_k$.

$X$-variables ($\mathbf{x}_k$) which are important for the modeling of $\mathbf{Y}$ are identified by large PLS-R coefficients, $b_{km}$; variables which are important for the modeling of $\mathbf{X}$ are identified
by large loadings, \( p_{ki} \). A summary of the importance of an \( X \)-variable for both \( Y \) and \( X \) is given by \( \text{VIP}_k \) (Wold S., 2001).

In PLS algorithm, the optimization criterion used to define projection directions is (Rosipal and Krämer, 2006)

\[
\text{max}_{\|r\|=1, \|s\|=1} \left[ \text{cov}(X_r, Y_s) \right]^2 = \text{max}_{\|r\|=1, \|s\|=1} \text{var}(X_r) \left[ \text{corr}(X_r, Y_s) \right]^2 \text{var}(Y_s)
\]

The maximal correlation is balanced with the requirement to explain as much variance as possible in both \( X \)- and \( Y \)-spaces. In the case of a one-dimensional \( y \)-space only the \( X \)-space variance is involved. PLS attempts to find factors or Latent Variables that maximize the amount of variation explained in \( X \) that is relevant for predicting \( Y \). This is in contrast to PCR, where the factors (Principal Components) are selected solely based on the amount of variation that they explain in \( X \) (McLennan and Kowalski, 1995).

PCA projects the original variables onto a direction of maximal variance called principal direction; its optimization criterion to define projection directions can be written as \( \text{max}_{\|r\|=1} \text{var}(X_r) \) where \( \text{var}(t) = \frac{t^T t}{N} \) denotes the sample variance (Rosipal and Krämer, 2006).

Because of the unique relationship between weights and loadings in a PLS model, the calculation of scores for new data does not simply involve a projection onto the loadings, as it does with PCA (i.e., \( T_{\text{new}} = X_{\text{new}} P \)); the scores for new data are instead calculated using \( T_{\text{new}} = X_{\text{new}} W (P^T W)^{-1} \).

Like PCR and MLR, prediction of the \( y \)-values for new data is accomplished using the regression equation \( (y = x b, \) with \( x \) a row vector and \( b \) a regression vector).

The scores and loadings calculated in PLS are not the same as those calculated in PCA and PCR. They can be thought of, however, as PCA scores and loadings that have been rotated in a manner that makes them more relevant for predicting \( y \) (McLennan and Kowalski, 1995). As in PCR, the PLS model converges to the MLR solution if the maximum possible number of latent variables is retained in the model.

The main advantage of PLS is often computational speed. In addition, PLS models generally require fewer factors than PCR models for the same set of data, which has advantages in both model implementation and model interpretation.
4.4 SIMPLS

It can be shown (Höskuldsson, 1988) that, in the PLS algorithm, the first weight vector \( w_1 \) is given by the eigenvector of the combined variance–covariance matrix, \( X^TYY^TX \), that is associated with the largest eigenvalue, \( i.e., \)

\[ \lambda w_1 = X^TYY^TXw_1 \]

and the following weight vectors (component \( a \)) are eigenvectors to the deflated versions of the same matrix. Based on this equation, another algorithm for calculating PLS models, called SIMPLS, was developed by Sijmen de Jong (de Jong, 1993). It was introduced to avoid deflation steps at each iteration of PLS1 and PLS2. The SIMPLS approach directly finds the weight vectors \( w_a \) (\( a = 1, \ldots, A \)) which are applied to the original not deflated matrix \( X \) (SIMPLS deflates the cross-covariance matrix). The criterion of the mutually orthogonal score vectors \( t_a \) (\( a = 1, \ldots, A \)) is kept. This method relies on orthogonalization of a Krylov sequence to calculate the PLS weights, which is non-intuitive but very fast and accurate. For univariate responses (\( y \)) SIMPLS gives the exact same result as PLS1; for multivariate \( Y \) there is a slight difference between the SIMPLS approach and NIPALS-PLS2 (a little different solution). There is some evidence that this solution is actually slightly superior to the NIPALS PLS solution because it maximizes the covariance criterion, whereas NIPALS does not. Primarily because of its speed and less memory requirements, SIMPLS is the default PLS algorithm used in PLS_Toolbox.

4.5 Interpretation of the PLS-R model (Wold S. et al. 2001)

The linear PLS-R model finds ‘new X-variables’, \( t \), which are estimates of the LV’s or their rotations. PLS-R forms the \( t \)’s as linear combinations of the original X-variables, and thereafter uses these new \( t \)’s as predictors of \( Y \). Only as many new \( t \)’s are formed as are needed, as are predictively significant. For each component, the parameters \( t, u, w, p \), and \( c \) are determined by a PLS-R algorithm. For the interpretation of the PLS-R model, the scores, \( t \) and \( u \), contain the information about the objects and their similarities/dissimilarities with respect to the given problem and model. The weights \( w \) and \( c \) give information about how the variables combine to form the quantitative relation between \( X \) and \( Y \), thus providing an interpretation of the scores, \( t \) and \( u \). Hence, these weights are essential for the understanding of which X-variables are important (numerically large \( w \)–values) and which X-variables that provide the same information (similar profiles of \( w \)–values).
The PLS weights $w$ express both the ‘positive’ correlations between $X$ and $Y$, and the ‘compensation correlations’ needed to predict $Y$ from $X$ clear from the secondary variation in $X$. The latter is everything varying in $X$ that is not primarily related to $Y$.

The residuals, the part of the data that are not explained by the model, are of diagnostic interest for the quality of the model. Large $Y$-residuals indicate that the model is poor, or rather that there is not a strong relationship between $X$ and $Y$, and a normal probability plot of the residuals of a single $Y$-variable is useful for identifying outliers in the relationship between $T$ and $Y$, analogously to MLR. The residuals for $X$ are the part not used in the modeling of $Y$. Large residuals in $X$ imply that a substantial amount of variation in $X$ is not related to $Y$. These $X$-residuals are useful for identifying outliers in the $X$-space, i.e. objects with structures that do not fit the model.

In a geometric representation of PLS-R, the $X$ matrix can be represented as a swarm of $N$ points in the $K$ dimensional space, where each column of $X$ ($x_k$) defines one coordinate axis. The PLS-R model defines an $A$-dimensional hyper-plane ($A < K$) in $X$-space such that this plane well approximates $X$ (the $N$ points, row vectors of $X$). That hyper-plane, in turn, is defined by one line, one direction, per component (Fig. 4.2). The direction coefficients of these lines (each PLS direction, each component) are $p_{ak}$; this slope is the cosine of the angle between the PLS direction and the coordinate axis, $x_k$. The coordinates of each object, $i$, when its data (row $i$ in $X$) are projected down on the hyper-plane, are the scores $t_{ia}$. These positions of the projected data points are related to the values of the responses, $Y_{im}$. The coordinates of the projection ($t_a$, $a = 1, ..., A$) are good predictors of $Y$.

Figure 4.2: $K$-dimensional space ($x_1, ..., x_K$) and projection of scores (on the left) and geometric representation of PLS-R (on the right, Wold S. 2001). The model defines an $A$-dimensional hyper-plane.
X-scores (\(t\)) can show object similarities and dissimilarities. The plot of the X-scores (e.g., \(t_1\) vs. \(t_2\)) can show the objects grouped according to size and shape.

4.6 Latent Variables (Wold S. et al. 2001)

In PLS modeling, the investigated system is assumed to be actually influenced by just a few underlying variables (LVs). The number of these latent variables is usually not known. Also, the PLS X-scores \(t\) are usually not direct estimates of the LV’s, but rather they span the same space as the LV’s. Thus, the latter (denoted by \(V\)) are related to the former (\(T\)) by a, usually unknown, rotation matrix, \(R\) (with the property \(R^TR=I\)):

\[ T=RV. \]

Both the \(X\) and the \(Y\)-variables are assumed to be realizations of these underlying LV’s, and are hence not assumed to be independent.

If the number of LV’s actually equals the number of \(X\)-variables (\(K\)) and the \(X\)-variables are independent, then PLS-R and MLR give identical results. Hence, PLS-R can be seen as a generalization of MLR, containing the latter as a special case in situations when the MLR solution exists, i.e. when the number of \(X\) and \(Y\)-variables is fairly small in comparison to the number of observations, \(N\). In most practical cases, however, the \(X\)-variables are not independent and \(X\) is rank deficient. Then, PLS-R gives a ‘shrunk’ solution which is statistically more robust than the MLR solution, and hence gives better predictions than MLR (Frank and Friedman, 1993).

PLS-R gives a model of \(X\) in terms of a bilinear projection, plus residuals. PLS-R assumes that there may be parts of \(X\) that are unrelated to \(Y\). These parts can include noise and/or regularities non-related to \(Y\). Thus, unlike MLR, PLS-R tolerates noise in \(X\).

4.7 The number of PLS components

The intension of PLS is to form components that capture most of the information in the \(X\)-variables, that is useful for predicting \(y_m\) (\(m = 1, \ldots, M\)), while reducing the dimensionality of the regression problem by using fewer components than the number of \(X\)-variables (Garthwaite, 1994). Although it is possible to calculate as many PLS components as the rank of the \(X\) block matrix, not all of them are normally used. The main reasons for this are that the measured data are never noise-free and some of the smaller components only describe noise. It is common to leave out small components, score vectors corresponding to small eigenvalues, in order to avoid collinearity.
problems from influencing the solution (Geladi and Kowalski, 1986). Hence, a strict test of the predictive significance of each PLS component is necessary, and then stopping when components start to be non-significant (Wold S., 2001).

In the regression modeling process, *calibration* is the fitting stage (finding the new component). The available objects may be divided in a *calibration set*, used to compute the model parameters (LVs, regression coefficients), and a *test set* (or prediction or *validation set*), used to validate the model, in an effort to assess how a model built using all of the calibration data would perform when applied to new, unknown data. In turn, the calibration set is split into a *training set*, used to develop the model, and an *evaluation set* (or optimization set), used for performance testing. Any model needs to be validated before its use, to estimate the uncertainty of future predictions. The best *validation* (i.e. testing) of a model is its ability to precisely predict the Y-values of observations with new X-values (validation set) (Wold S., 2001). A *validation set* consists of samples with known response values. Only the X-values are fed into the model, from which response values are predicted and compared to the known, true response values. The model is validated if the prediction residuals are low. The model should be calibrated and validated on samples covering the region new samples belong to, *i.e.* on similar samples (similarity being determined by the X-values).

The selection of the LVs number is important when building a PLS model: if more variables are selected, the model will easily result in over-fitting, while the selection of fewer variables will cause under-fitting. Over-fitting means the selected variables work well on the calibration data but with little or no predictive power on new data. On the other hand, a model that does not fit the data well will not predict well either.

In the absence of an independent and representative validation set, a reasonable way of model validation is given by Cross-Validation (CV), which simulates how well the model predicts new data; some samples (*evaluation set*) are picked out for calculating Predictive Residual Error Sum of Squares (PRESS) (Wold S., 1978). *Cross-validation* is a practical and reliable way to test the predictive significance (Wold H., 1982; Tenenhaus, 1998); this has become the standard in PLS-R analysis and it helps in establishing the number of components (*a*) needed to describe the linear model.

Predictive ability of a regression model measures the error on objects that are not used to build the model (Forina et al., 2004). To compare the predictive ability of models, it is useful to introduce several measures of a model's fit to the data and of predictive power. It is common to estimate the average deviation of the model from the data. The
Root-Mean-Square Error of Calibration (RMSEC) tells the fit of the model to the calibration data (McLennan and Kowalski, 1995):

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$

where the $\hat{y}_i$ are the values of the predicted variable when all samples are included in the model formation, and $n$ is the number of calibration samples. The use of the word ‘error’ to indicate the estimate of a standard deviation is frequent in multivariate calibration. This equation can be used to define the Root-Mean-Square Error of Cross-Validation (RMSECV) as well, except that now the $\hat{y}_i$ are predictions for samples not included in the model formulation. RMSEC is a measure of how well the model fits the data; RMSECV is a measure of a model’s ability to predict new samples that were not used to build the model. RMSECV is related to the PRESS (Prediction Residual Error Sum of Squares) value for the number of LVs included in the model (McLennan and Kowalski, 1995):

$$\text{RMSECV}_a = \sqrt{\frac{\text{PRESS}_a}{n}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

where $a$ refers to the number of factors (LVs) used in the model, $n$ is the number of calibration samples, $\hat{y}_i$ is the predicted response when the model is built without sample $i$ (Forina et al., 2004). PRESS is a commonly used criterion for LVs number selection. It is common to calculate PRESS, and thus RMSECV, for a Leave-One-Out (LOO) Cross-Validation, i.e., where each sample is left out of the model formulation and predicted once. PRESS$_a$ is the sum of squared prediction errors for the model which includes $a$ factors:

$$\text{PRESS}_a = \sum_{i=1}^{n} (y_i - \hat{y}_{-i,a})^2$$

with $\hat{y}_{-i,a}$ the predicted $y_i$ response obtained when $a$ latent variables are fitted (in the model) and the $i$-th sample is omitted.

LOO Cross-Validation method is of most application to select the number of variables, however it often causes over-fitting because only one sample is left out (Martens and Dardenne, 1998). Thus, when the number of samples in the data set used is large enough, more samples are suggested to leave out (Shao, 1993; Xu and Liang, 2001).

Basically, CV is performed by dividing the data (calibration set) in a number of $G$ groups (typically, between 3 and 10 cancellation groups) and then developing a number of parallel models from reduced data with one of the groups deleted (Wold S. et al. 2001). Having $g$ equal to the number of objects (LOO approach) is not recommendable.
(Shao, 1993). Each time, the objects in one of the cancellation groups constitute the *evaluation set* and the other objects constitute the *training set*; all the objects are one time in the evaluation set. (Forina et al., 2008). After developing a model, differences between actual and predicted $Y$-values are calculated for the deleted data. The sum of squares of these differences is computed and collected from all the parallel models to form the PRESS$_a$, which estimates the predictive ability of the model (Wold S. *et al.* 2001). Finally, the final model is developed with all the objects. RMSEC$_a$ refers to this model. CV is performed on each component ($a = 1,\ldots, A$). A PRESS is calculated for the final model with the estimated number of significant components. The RMSECV is a ‘composite’ prediction error obtained over all repetitions of the CV procedure.

There are several different cross-validation methods and these vary with respect to how the samples of the $G$ groups are selected. In *Venetian Blinds*, each group is determined by selecting every $g$-th object in the calibration set, starting at objects numbered 1 through $g$ (Fig. 4.3). In *Contiguous Blocks*, each group is determined by selecting contiguous blocks of $n/g$ objects in the calibration set, starting at object number 1. In *Leave-One-Out*, each single object in the calibration set is used as a test set.

Cross-Validation enables an assessment of the correct complexity of a model (the number of LVs in a PLS model) and allows an estimation of the performance of a model when applied to unknown data.
Because fit and prediction are different aspects of a model’s performance, a plot of RMSEC and RMSECV as a function of the number of latent variables retained in the PLS model (A) is useful for determining the optimal number of LVs to retain in a model that is built using the full set of data. In principle, the RMSEC should always decrease as the number of LVs retained in the model increases. However, because the RMSECV is determined from the CV experiment, in which the evaluation samples were not used to build the model that was used to test them, this value can actually increase as too many LVs are added to the model. The optimal number of LVs is typically the number at which the addition of another latent variable does not greatly improve the performance of the model (Using Cross-Validation, 2013) (at least 2%). When prediction is the goal, the model that gives the minimum RMSECV value amongst the prediction models should be selected (Yeniay and Goktas, 2002). The RMSECV curve may have several local minima and a global minimum. When the location of the minimum is not well defined, the models with the smallest values of RMSEC are considered as the optimized.

Alternatively, one could designate a completely independent prediction set of samples that have known Y-values, so that such an assessment can be done directly. In this case, it is possible to calculate a Root-Mean-Square Error of Prediction (RMSEP) when the model is applied to new data (an external validation set) provided that the reference values for the new data are known. RMSEP is calculated exactly as in RMSEC equation, except that the estimates \( \hat{y}_i \) refer to samples in the independent prediction set, which were not involved in either model building or CV.

Generally, a good predictive model should have high values of Pearson correlation coefficient (r), low values for RMSE and maximum Ratio of Percentage Deviation (RPD). RPD is the ratio of the standard deviation of the laboratory measured (reference) data to the RMSE of the cross-validation (Williams, 1987). It is the factor by which the prediction accuracy has been increased compared with using the mean of the original data. RPD values were classified (Viscarra et al., 2007) as follows: RPD < 1 indicates very poor model and/or predictions and their use is not recommended; RPD between 1 and 1.4 indicates poor model and/or predictions where only high and low values are distinguishable; RPD between 1.4 and 1.8 indicates fair model and/or predictions that may be used for assessment and correlation; RPD values between 1.8 and 2 indicates good model and/or predictions where quantitative predictions are possible; RPD
between 2 and 2.5 indicates very good, quantitative model and/or predictions, and RPD > 2.5 indicates excellent model and/or predictions.

### 4.8 Mean square error

If \( x \) is a random (or aleatory, or stochastic) variable, the expected value (or expectation, or first moment) of a measurable function of \( x \), \( f(x) \), is given by

\[
E[f(x)] = \int_{-\infty}^{+\infty} f(x)p_X(x)dx,
\]

where \( p_X(x) \) is the probability density function of \( x \). If \( f(x) = x \),

\[
\int_{-\infty}^{+\infty} xp_X(x)dx = E[x]
\]

is the expected value of \( x \). If \( f(x) = (x - E[x])^2 \), then

\[
E[(x - E[x])^2] = \int_{-\infty}^{+\infty} (x - E[x])^2 p_X(x)dx = \int_{-\infty}^{+\infty} [x^2 - 2xE[x] + (E[x])^2]p_X(x)dx
\]

\[
= \int_{-\infty}^{+\infty} x^2p_X(x)dx - 2E[x] \int_{-\infty}^{+\infty} xp_X(x)dx + (E[x])^2 \int_{-\infty}^{+\infty} p_X(x)dx = E[x^2] - 2E[x]E[x] + (E[x])^2 = E[x^2] - (E[x])^2.
\]

So, the variance for a stochastic variable \( x \) is given by

\[
\text{var}(x) = E[(x - E[x])^2] = E[x^2] - (E[x])^2.
\]

If \( x = y - \hat{y} \), the formula can be written as

\[
\text{var}(y - \hat{y}) = E[(y - \hat{y})^2] - (E[(y - \hat{y})])^2.
\]

In the bidimensional case, since

\[
E[f(x_1, x_2)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x_1, x_2)p_X(x_1, x_2)dx_1 \, dx_2
\]

\[
p_{X_1}(x_1) = \int_{-\infty}^{+\infty} p_X(x_1, x_2)dx_2 \quad p_{X_2}(x_2) = \int_{-\infty}^{+\infty} p_X(x_1, x_2)dx_1
\]

then

\[
E\{g(x_1) + h(x_2)\} = E\{g(x_1)\} + E\{h(x_2)\}.
\]

If \( g(x_1) = x_1 \) and \( h(x_2) = x_2 \), then \( E[x_1 + x_2] = E[x_1] + E[x_2] \).

Therefore,

\[
\text{var}(y - \hat{y}) = E[(y - \hat{y})^2] - (E[y] - E[\hat{y}])^2 \quad \text{and}
\]

\[
E[(y - \hat{y})^2] = (E[y] - E[\hat{y}])^2 + \text{var}(y - \hat{y}).
\]

This formula can be described as: mean square error = squared bias + residual variance (Höskuldsson, 1988). There is a lower bound on the mean square error. In linear modeling of data, after having included a certain number of variables or components, the mean square error does not get any smaller when new variables or components are introduced into the model. Introducing more variables generally reduces the bias, but
after a certain number of variables the residual variance increases or the mean square error may even increase (Höskuldsson, 1988).

The estimate of the bias can be achieved by carrying out the PLS estimation until the regression coefficients become close to zero. The cross-validation tells one when to stop extracting components. The difference in estimated $Y$-values from where CV tells one to stop and all the PLS components is an estimate of the bias. Normally, the CV tells one to stop before all components have been extracted. The last components may have regression coefficients different from zero, but do not contribute to the predictive power of the model (Höskuldsson, 1988).
Chapter 5

DISCRIMINANT ANALYSIS

5.1 Principles of canonical analysis (Legendre, 2012)

Discriminant analysis is a canonical method of analysis. Canonical analysis is the simultaneous analysis of two, or possibly several data tables. Canonical analysis allows one to model a data table from the descriptors of a second data table; these two tables of quantitative data form the training set. Using the resulting model, it is possible to forecast the position of any new observation among those of the training set, e.g. along environmental gradients. The new observation may represent some condition which may occur in the future, or for a different but comparable data set.

Canonical analyses allow to perform a direct comparison or an indirect comparison of two data matrices. In indirect comparison (or indirect gradient analysis), the matrix of explanatory variables $X$ does not intervene in the calculation producing the ordination of $Y$. Correlation or regression of the ordination vectors on $X$ are computed a posteriori. Indirect comparison proceeds in two steps: the structure (ordination axes) is first identified from a set of descriptors of prime interest in the study, and then the structure is interpreted using either the descriptors that were analyzed in the first step to identify the structure or another set of descriptors assumed to help explain the structure.

In direct comparison analysis (canonical analysis), matrix of explanatory variables $X$ (descriptors) intervenes in the calculation producing the ordination of $Y$, forcing the ordination vectors to be maximally related to combinations of the variables in $X$. This description applies to all forms of canonical analysis and in particular to the asymmetric forms. In direct comparison, one simultaneously analyses the response and explanatory data tables in order to identify what they have in common. It allows to bring out the ordination structure common to two data sets (direct gradient analysis). Direct comparison analysis allows one to directly test a priori working hypotheses by bringing out all the variance of $Y$ that is related to $X$, and allowing formal tests of these hypotheses to be performed.

In mathematics, a canonical form is the simplest and most comprehensive form to which certain functions, relations, or expressions can be reduced without loss of generality. The canonical form of a covariance matrix is its matrix of eigenvalues.

There are two main families of canonical ordination methods: asymmetric and symmetric. In the asymmetric forms of canonical analysis there is a response data set
(Y) and an explanatory data set (X). In contrast, symmetric methods are used in cases where the two data sets (called Y1 by Y2 to mark the symmetry) play the same role in the study; two different ordinations of the objects are produced, one for each data set. Asymmetric canonical analysis combines the concepts of ordination and regression. As it is the case with the simple ordination methods (e.g., PCA), the asymmetric methods of canonical analysis produce a single ordination of the objects, which may be plotted in a scatter diagram. The asymmetric methods include linear discriminant analysis (LDA). Canonical analysis is related to multiple regression analysis. Multiple linear regression (MLR) has been described as a method for modeling a response variable y using a set of explanatory variables assembled into a data table X. In regression analysis, while the original response variable y provides, by itself, an ordination of the objects in one dimension, the vector of fitted values ŷi= b0+b1xi1+b2xi2+...+bKxIK creates a new one-dimensional ordination of the same objects. The ordinations corresponding to y and ŷ differ; the square of their correlation is the coefficient of determination of the multiple regression model, \( R^2 = [r(y, ŷ)]^2 \). So, multiple regression creates a correspondence between ordinations y and ŷ, because ordination ŷ is constrained to be optimally (in the least-square sense) and linearly related to the variables in X. The constraint implemented in multiple regression maximizes \( R^2 \). The asymmetric methods of canonical analysis shared this property.

Asymmetric canonical analysis produces ordinations of Y that are constrained to be linearly related to a second set of variables X, and the results are plotted in a reduced space. The way in which the relationship between X and Y is established differs among methods of asymmetric canonical analysis.

The asymmetric forms of canonical analysis should be used when one of the data set (Y) is to be explained by another (X). Linear discriminant analysis is applicable when the response data set (Y) contains a classification of the objects; it is used mostly to discriminate among groups of objects using descriptors of the physical environment.

5.2 Discriminant Analysis (Legendre, 2012)

In discriminant analysis, the grouping of objects is already known at the start of the analysis. The division of the objects into groups constitutes the classification criterion, which is considered to be a qualitative response variable y. The problem consists in interpreting the groups, in trying to determine to what extent a set of quantitative descriptors, seen as the explanatory variables X, can actually explain this grouping.
As in multiple regression, discriminant analysis estimates the parameters of a linear model of the explanatory variables that may be used to forecast the response variable (states of the classification criterion) and can handle several groups. The analysis is based upon an explanatory data matrix $X$ (N×K), where N objects are described by K quantitative descriptors. The quantitative variables (X) are predictors of the classification. If the differences in the $X$, among the predefined groups, are significant then the analysis finds the linear combinations (called discriminant functions or identification functions) of the predictors that best discriminate among the groups.

**Discriminant functions** (also called standardized discriminant functions) are computed from standardized descriptors. The coefficients of these functions are used to assess the relative contributions of the explanatory descriptors in the discrimination among groups. **Identification functions** (also called unstandardized discriminant functions) are computed from the original descriptors (not standardized). **Identification** is to assign objects to pre-established groups. The identification functions allow the assignment of any object to one of the states of the qualitative descriptor, using the values taken by several quantitative variables. The distributions of the explanatory descriptors must not be too far from normality, and their within-group dispersion matrices must be reasonably homogeneous (*i.e.* about the same among groups).

The approach is to search for a linear combination of descriptors that provides the most efficient discrimination among groups. Figure 5.1 shows an idealized example of two groups of objects described by two descriptors only. The groups cannot be separated on either of the two axes taken alone, but the two groups are perfectly separated along a new discriminant descriptor, which is a linear combination of the two original descriptors. That discriminant axis is parallel to the direction of greatest variability between groups. This suggests that the weights used in the discriminant function could be the elements of the eigenvectors of a between-group dispersion matrix. The method can be generalized to several groups and many descriptors.

![Figure 5.1: Two groups of objects on descriptor axes (on the left) and on discriminant axis (on the right).](image-url)
The method was originally proposed by Fisher (1936) for the two-group case ($G=2$); it is called Fisher’s or simple discriminant analysis. A single function is needed to discriminate among the groups. The solution can be entirely derived from the output of a multiple regression using a dummy variable defining the two groups (used as the dependent variable $y$) against the table of explanatory variables $X$. A dummy variable is an artificial variable that assumes a discrete numerical value in the group (or class) description.

In general, multistate qualitative descriptors may be binary-coded as dummy variables. This coding allows the use of qualitative descriptors in procedures such as multiple regression and discriminant analysis, that have been developed for quantitative variables and in which binary variables may also be used. A multistate qualitative descriptor, with $S$ states, can be decomposed into $S–1$, or $S$, dummy variables (Tab. 5.1).

<table>
<thead>
<tr>
<th>States</th>
<th>Dummy variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 0 0 0</td>
</tr>
<tr>
<td>2</td>
<td>0 1 0 0</td>
</tr>
<tr>
<td>3</td>
<td>0 0 1 0</td>
</tr>
<tr>
<td>4</td>
<td>0 0 0 1</td>
</tr>
</tbody>
</table>

In this example, the objects are coded by three ($S–1$) or four ($S$) dummy variables instead of a single multistate qualitative descriptor.

The case with several groups of objects is called multiple discriminant analysis or canonical variate analysis (CVA). CVA separates objects into classes (or groups) by minimizing the within-class variance and maximizing the between-class variance. So, with respect to PCA, the aim of CVA is to find directions (i.e. linear combinations of the original variables) in the data space that maximize the ratio of the between-class to within-class variance, rather than maximizing the between-object variance without taking into account any information on the classes, as PCA does. These directions are the discriminant functions (or canonical variates) (Ballabio and Todeschini, 2008). Quadratic Discriminant Analysis (QDA) is a probabilistic parametric classification technique; it separates the class regions by quadratic boundaries and makes the assumption that each class (or group) has a multivariate normal distribution, while the dispersion (represented by the class covariance matrices) is different in the classes. A special case, referred to as Linear Discriminant Analysis (LDA), occurs if all the class
covariance matrices are assumed to be identical (the dispersion is the same for all the classes) (Ballabio and Todeschini, 2008). In LDA, the objects are divided into $G$ groups, described by a qualitative descriptor forming the response matrix $Y$. The method seeks linear combinations of explanatory variables (matrix $X$) that explain the classification in $Y$ by maximizing the dispersion of the centroids of the $G$ groups. This is obtained by maximizing the ratio of the among-object-group dispersion over the pooled within-object-group dispersion.

The problem consists in finding linear combinations of the predictors in matrix $X$ that maximize the differences among groups while minimizing the variation within the groups. Each descriptor may have already been transformed to meet the condition of multinormality or at least to reduce the asymmetry of its distribution. The discriminant analysis model is robust to departures from this condition, but the parametric statistical tests assume within-group normality of the descriptors.

Computations are carried out using either dispersion matrices ($V, A$) or matrices of sums of squares and cross-products of centered descriptors ($W, B$). Matrix $T$ is the matrix of scalar products of the centered descriptors, $[x - \bar{x}]$, for all objects irrespective of the groups: $T = [x - \bar{x}]^T[x - \bar{x}]$ (*Total sums of squares and cross-products*). When divided by the total number of degrees of freedom $N-1$, it becomes the *total dispersion matrix* $S$ used in PCA. *Degrees of freedom* describe the number of values in the final calculation of a statistic (this is a measure on the items in a random sample, e.g. the mean) that are free to vary (*Glossary of Statistical Terms*, 2008). In this case, there are $N$ values, corresponding to the objects, minus one (their mean).

Matrix $W$, which pools the sums of squares within all groups, is computed by adding up matrices $W_1$ to $W_G$ of the sums of squares and cross-products for each of the $G$ groups. Each matrix $W_g$ is computed from descriptors that have been centered for the objects of that group only. Matrix $W_g$ is the product $[x - \bar{x}]^T[x - \bar{x}]$ (of the centered descriptors) for the objects that belong to group $g$ only. Dividing the *pooled within-group matrix* $W$ by the within-group number of degree of freedom, $N - G$, produces the *pooled within-group dispersion matrix* $V$.

Matrix $B$ of the sums of squares among groups is computed by subtracting the pooled within-group matrix $W$ from the total matrix of sums of squares $T$: $B = T - W$. The *among-group dispersion matrix* $A$ is obtained dividing $B$ by the number of degrees of freedom: $(N-1) - (N - G) = G - 1$. 

52
The solution to the problem of maximizing the variation among groups while minimizing that within groups calls for the eigenvalues and eigenvectors of a matrix corresponding to the ratio of the among-group dispersion (A) to the pooled within-group dispersion (V). The maximization problem is stated by the matrix equation \((V^{-1}A - \lambda_a I)u_a = 0\) or \((A - \lambda_a V)u_a = 0\).

The number of discriminant axes needed for the ordination of \(G\) groups is \(G - 1\).

The eigenvectors \(u_a\), can be normalized in an appropriate manner. If \(U\) is the matrix of the eigenvectors, the matrix \(C = U(U^TVU)^{-1/2}\) contains the normalized eigenvectors defining the canonical space of the discriminant analysis. The principal axes describe the dispersion among groups. The first principal axis indicates the direction of largest variation among group centroids, and so on for the successive canonical axes (after the reference space has been straightened up).

The matrix of discriminant scores \(F\) is obtained by multiplying the matrix of centred data with the matrix of normalized eigenvectors: \(F = (X - \bar{X})C\). This matrix contains the positions (or scores) of the objects, \(i\), on each canonical axis \(a\) \((f_{ia} = (x_{i1} - \bar{x}_1)c_{1a} + \ldots + (x_{iK} - \bar{x}_K)c_{Ka})\).

The origin of the discriminant axes is located at the centroid of all objects. The positions of the centroids of the \(G\) groups of objects in canonical space is computed by multiplying the matrix of the original group centroids (computed from data centered over all objects in the analysis) with the matrix of normalized eigenvectors \(C\). The centroid of a group is a point whose coordinates are made of the mean values of the objects of that group for all descriptors. The matrix of group centroids therefore has \(G\) rows and \(K\) columns.

The columns of matrix \(F\) are the canonical variates. The position of the group centroids in discriminant space can be found by computing these same functions for the mean values of the groups along the \(X\) variables. The distances among objects in discriminant space are Mahalanobis distances. Mahalanobis generalized distance takes into account the covariances among descriptors; it produces identical results for variables that are standardized or not. This measure computes the distance between two points in a space whose axes are not necessarily orthogonal, in order to take into account the correlations among descriptors.

Discrimination among \(G\) groups requires a maximum of \(G - 1\) discriminant functions. It is possible (Bartlett, 1948) to estimate the significance of the discriminant power of the \(G - 1 - A\) axes remaining after accepting the first \(A\) eigenvalues as significant. When the
centroids of the groups do not differ on the remaining $G - A - 1$ discriminant functions, the discriminant power is limited to the first $A$ functions.

If the analysis is carried out on the non-standardized descriptors, the columns of matrix $C$ are the identification functions. Identification functions are used to place new objects in the canonical space. This provides the positions of these objects on the canonical axes. By plotting the point representing a new object in the canonical ordination space together with the original set of objects, it is possible to identify the group to which the new object is most likely to belong.

When applying LDA, the number of objects ($N$) must be significantly greater than the number of variables ($K$), while QDA requires a larger number of objects than LDA, since covariance matrices are calculated for each class (Ballabio and Todeschini, 2008). Moreover, when variables are highly correlated among them, *i.e.* in presence of multicollinearity, discriminant analysis runs the risk of over-fitting (Hand, 1997). In order to overcome these problems, a first approach is simply the reduction of the number of variables.

If the analysis requires that poorly discriminating variables be eliminated in order to identify a subset of good discriminators, one should select variables carefully because any single variable may not discriminate groups well, although it may have high discriminating power in combination with other variables. Variable selection techniques are needed. This is equally true in regression analysis and discriminant analysis.

### 5.3 Multivariate classification methods (Ballabio and Todeschini, 2008)

Methods that attempt to identify classes without using pre-established class memberships are known as *unsupervised* pattern recognition (*e.g.*, cluster analysis and PCA). These techniques are based on functional relationships among the elements (*e.g.*, distances, variances). Methods that require *a priori* information on the set of samples that is used for classification purposes are generally called *classification* or *supervised* pattern recognition. Classification is useful when the response is a category variable that can be interpreted in terms of several classes to which an object may belong. A category variable is a class variable, *i.e.* each of its levels is a category (or class, or group), without any possible quantitative equivalent.

*Classification methods* are multivariate techniques designed to find mathematical models able to recognize the membership of each object to its proper class, or group, on the basis of a set of measurements. Once a classification model has been obtained, the
membership of new objects to one of the defined classes can be reliably predicted. Contrary to regression methods, which model quantitative responses on the base of a set of explanatory variables, classification techniques (classifiers) are quantitative methods for the modeling of qualitative responses. Classification methods find mathematical relationships between a set of descriptive variables and a qualitative variable (i.e. the membership to a defined category).

Several classifiers have been proposed, with different characteristics and properties. Distinctions can be made among the different classification techniques on the basis of the mathematical form of the decision boundary, i.e. on the basis of the ability of the method to detect linear or non-linear boundaries between classes. In linear classification methods, the model calculates the best linear boundary for class discrimination, while non-linear classification methods find the best curve (non-linear boundary) for separating the classes.

Moreover, classification techniques can be probabilistic, if they are based on estimates of probability distributions, i.e. a specific underlying probability distribution in the data is assumed. Among probabilistic techniques, parametric and non-parametric methods can be distinguished, when probability distributions are characterized by location and dispersion parameters (e.g., mean, variance, covariance).

Classification methods can also be defined as distance-based, if they require the calculation of distances among objects or among objects and models.

Another important distinction can be made between pure classification and class-modeling methods. Pure classification techniques separate the hyperspace in as many regions as the number of available classes. Each object is classified as belonging to the category corresponding to the region of hyperspace where the object is placed. In this way, objects are always assigned to one of the defined classes. When pure classification techniques are applied, it is important to assure that the unknown objects to be predicted belong to one of the classes used in the model calculation. Class-modeling techniques represent a different approach to classification, since they focus on modeling the analogies among the objects of a class, defining a boundary to separate a specific class from the rest of the hyperspace. Each class is modeled separately and objects fitting the class model are considered element of the class, while objects that do not fit are recognized as non-members of that class. As a consequence, a particular portion of the data hyperspace can be enclosed within the boundaries of more than one class or of
none of the classes and three different situations can be encountered: objects can be assigned to a class, to more than one class or to none of the considered classes. Classification results can be a diagnostic tool to find outliers (i.e. samples that are not typical of the population). An outlier looks so different from the others that it either is not well described by the model or influences the model too much.

With respect to pure classification techniques, class-modeling methods have some advantages: it is possible to recognize objects that do not fall in any of the considered class spaces and consequently identify members of new classes not considered during the model calculation. Furthermore, as each class is modeled separately, any additional class can be added without recalculating the existing class models. Some classification methods produce probability values representing the degree to which the object belongs to a certain class.
Chapter 6

PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS

6.1 PLS-DA modeling method

Discriminant analysis (DA) is a statistical technique that was developed to identify variables explaining the differences between two or more groups of objects, and that uses measured descriptors to put the observed objects into the groups. The relative contributions of various explanatory variables to the distinction among these groups can be also determined. DA can be considered the first multivariate classification technique. Its approach assumes that each object has to be a member of one of the groups included in the analysis. Since DA classification rules always assign objects to groups or classes, it is not a class modeling technique (Ballabio and Todeschini, 2008).

One of the most known statistical method of discrimination is Linear Discriminant Analysis (LDA) (Fisher, 1936). The objective of the analysis is to find linear combinations of the $X$-variables that discriminate among the classes (discriminant functions), \textit{i.e.} have very different values for the classes (Jackson, 1991). However, LDA method is likely to lead to unstable models and poor predictions in the presence of quasi-collinearity among variables or in situations where the number of variables is large with respect to the samples (Naes and Indahl, 1998). With many and collinear $X$-variables, a PLS version of LDA (PLS-DA) is useful (Sjöström \textit{et al.}, 1986; Ståhle and Wold, 1987).

PLS is a routinely used methodology for both classification and discrimination problems with multicollinear data (correlated predictors), even if it was not designed for those purposes. The usual \textit{Partial Least Squares Discriminant Analysis} (PLS-DA) was proposed to overcome the multicollinearity problem of LDA, and can be defined as a straightforward extension of the PLS regression. PLS-DA method is based on modeling the differences among several classes with PLS; it identifies directions in the data space that discriminate classes directly. The objective of PLS-DA is to find a model that separates classes of objects on the basis of their $X$-variables. This model is developed from a training set of observations of known class membership (Ericsson \textit{et al.}, 2006a). The goal of PLS-DA is very similar to the one of LDA; in fact, PLS-DA is essentially the inverse-least squares approach to LDA and produces fundamentally the same result but with the noise reduction and variable selection advantages of PLS (Barker and
Rayens, 2003). As pointed out (Barker and Rayens, 2003; Sabatier et al., 2003), it is impossible to interpret PLS-DA with respect to the between-groups sums-of-squares and cross-products variance matrix, like LDA, because PLS-DA corresponds to extracting the unit eigenvector associated to the dominant eigenvalue of an altered version of this matrix.

PLS-DA consists in a classical PLS regression where the response variable is a categorical one expressing the class membership of the statistical units. Matrix \( \mathbf{X} (N \times K) \) collects a set of \( K \) quantitative explanatory variables observed on \( N \) statistical units. Each row of \( \mathbf{X} \) represents one observation and each column of \( \mathbf{X} \) represents one variable, or predictor. The categorical response variable \( \mathbf{y} \), qualitative and multistate, may be split into a set of dummy variables (or indicator variables). The number of dummy variables in the set is equal to the number of categories, or groups (\( G \)). The dummy matrix \( \mathbf{Y} \) is a set of binary variables describing the categories of a categorical variable on a set \( \mathbf{X} \) of predictor variables (Perez-Enciso and Tenenhaus, 2003).

PLS-DA is essentially based on the PLS2 algorithm that searches for latent variables (LVs) with a maximum covariance with the \( \mathbf{Y} \)-variables (Ballabio and Todeschini, 2008). In the PLS2 variant, each class is represented by a binary discriminant variable with value 1, logical one, for members of that class and 0 (or -1), logical zero, for non-members. The \( \mathbf{Y} \) block describes which objects are in the classes of interest; it explicitly gives the class memberships. \( \mathbf{Y} \) has \( G \) columns (for \( G \) classes) with 1 and 0, such that the entry in the \( g \)-th column is 1 and the entries in other columns are 0 for observations of class \( g \) (Ericsson et al., 2006a). Each row of \( \mathbf{Y} \) represents the classification of the corresponding row of \( \mathbf{X} \). PLS regression operates on this dataset, with the discriminant variable coded 0/1 (or -1/+1, which is equivalent) as \( \mathbf{y} \) variable in the model.

For each response variable, \( \mathbf{y}_g \) (\( g = 1, \ldots, G \)), a regression model on the \( \mathbf{X} \) components is written:

\[
\mathbf{y}_g = \mathbf{XW}_g \mathbf{c}_g^T + \mathbf{f}_g^* = \mathbf{XW}(\mathbf{P}^T \mathbf{W})^{-1}\mathbf{c}_g^T + \mathbf{f}_g^* = \mathbf{Xb}_g + \mathbf{f}_g^*
\]

where \( \mathbf{c}_g^T \) is the \( g \)-th column of \( \mathbf{C}^T \).

The results of the PLS2 model, built with all the indicator variables as \( \mathbf{Y} \) matrix, are used to predict class membership from the \( \mathbf{X} \) variables describing the objects. For each object, PLS-DA returns the prediction as a vector of size \( G \), with real values in the range between logical zero and logical one in a continuous manner. The model considers the actual logical classes, the real-value predictions, \( \hat{\mathbf{y}} \), and the prior
probabilities for each class, representing the probability of observing each class in the entire population (the probability of observing a 1 for each column of $Y$); then, it calculates, for each predicted $\hat{y}$ value, the probability that the given value belongs to each of the original classes (Wise et al., 2008). A classification rule has to be applied: an object may be assigned to the class with the maximum value in the vector of size $G$ with real values or, alternatively, a classification threshold between logical zero and logical one may be determined for each class. In the latter case, objects with predicted value $\hat{y}_{n,g}$ ($g$-th value of $n$-th object) above the threshold value (logical one) are predicted members, whereas objects with predicted value $\hat{y}_{n,g}$ below the threshold value (logical zero) are predicted non-members.

The threshold may be estimated using Bayes’ theorem and the available data in order to minimize total errors. Posterior class probabilities estimated using Bayes’ theorem is the method that provides the discrimination rule with minimal expected misclassification error (Mardia et al., 1997; Hastie et al., 2001). The Bayesian threshold calculation assumes that the predicted $\hat{y}$ values, obtained from a PLS model built for two or more logical classes, follow a distribution similar to what will be observed for future samples. A threshold is selected at the point where these two estimated distributions cross; this is the $y$-value at which the number of false positives and false negatives should be minimized for future predictions, at which the classes are split with the least probability of false classifications. A sample giving this $y$-value has a 50% chance of being in ‘class 1’ or ‘class 0’. The calculation assumes that the distribution of predicted $\hat{y}$ values for a class is approximately normal (Wise et al., 2006). Discriminant analysis is more sensitive to violations of normality than regression, and this is particularly true when the size of the two groups is very different (such as when a group is below 10%). Small numbers of samples in any class may bias the threshold. PLS-DA works reliably when each class is "tight" and occupies a small and separate volume in the $X$-space; furthermore, the number of modeled classes must not be too high (2÷4) (Ericsson et al., 2006a).

PLS-DA makes it possible to accomplish a rotation of the projection to give latent variables that focus on class separation (discrimination) (Ericsson et al., 2006a); a maximum separation among the classes is obtained. The maximum number of PLS-DA components that can be observed is higher than $G - 1$ (with $K > G$). On the contrary, the maximum number of LDA discriminant axes is less than or equal to $\min(K, G - 1)$. This
leads to a not intuitive geometrical interpretation of higher than $G - 1$ order components (Sabatier et al., 2003).

After relevant latent vectors are extracted (directions for discriminating among classes of observations), an appropriate classifier can be applied. A classifier is a model of data with a categorical response (i.e. each of its levels is a category or group); it is built from calibration data, for which classifications are known. The classifier assigns new test data to one of the categorical levels of the response. As well as for regression models, classifiers require cross-validation procedures to analyze the predictive classification capabilities on unknown objects. The prediction ability estimation of classification models is performed on different parameters with respect to regression methods, since the PLS-DA modeled response is qualitative and not quantitative. Several parameters can be used, such as the percentage of correctly classified objects with respect to the total number of available objects or the percentage of correctly classified objects of a category of interest. Even if these parameters can be calculated with the same procedures involved in the validation of regression models (leave-one-out, venetian blinds, contiguous blocks, random subsets, etc.), the percentage of objects retained in each cross validation group has to be considered, when classification models are validated (Ballabio and Todeschini, 2008). If the data are split into, f. e., 10 contiguous blocks, a correct validation procedure should at least retain objects of all the considered classes in each training group.

There are indices related to the classification quality of a single class. The threshold value selected by the PLS-DA algorithm may be used to calculate the sensitivity and specificity of the implemented model. The sensitivity describes the model ability to correctly recognize objects belonging to the $g$-th class (percentage of true positive); it is given by the number of samples predicted as in the class divided by number actually in the class (the fraction of ‘in-class’ samples which are above the given threshold). The specificity characterizes the ability of the $g$-th class to reject the objects of all the other classes (percentage of true negative); it is given by the number of samples predicted as not in the class divided by actual number not in the class (the fraction of ‘not-in-class’ samples which are below the given threshold). Sensitivity and specificity may be plotted as the threshold value is varied. In these plots, the abscissa axis is the threshold value used to classify objects into one group or the other and is the same axis as the predicted $\hat{y}$ value. Ideally, these lines cross while still at a value of 1. Crossing below a value of 1 indicates that, as the threshold is increased, the specificity increases (i.e. the false
positive rate decreases), likewise the sensitivity begins to suffer (drops below 1, i.e. the false negative rate increases) before the model is completely specific (Wise et al., 2006). The point where the two curves meet is the balance between false positives and false negatives. The higher up the ordinate axis this point is, the better the model. The exact value where the curves meet is: 1 - misclassification rate. As the classes start to overlap (e.g., due to a model which cannot sufficiently separate them), the sensitivity/specificity plots will drop in the middle (increased misclassification rate).

These plots help to determine how well PLS-DA separates samples which are in a given class from those not ‘in class’. On the basis of the classification aim, it is possible to decide which is the optimal balance of sensitivity and specificity and consequently set the best threshold value. The mean of sensitivity and specificity defines the efficiency of a classification function (Derde and Massart, 1986).

Once the PLS2 model has been checked and validated, it is possible to classify new samples. The evaluation of the model is based on whether or not the model places objects into groups more accurately than would occur by chance (or better than a random classifier).

6.2 Prediction probability and classification threshold for PLS-DA

Classification methods are expected to assign a sample to the class of origin. Classification rules must provide reliable classifications of forthcoming samples and must minimize the number of misclassifications (i.e. the expected error rate) (Botella et al., 2009).

PLS-DA calculates a prediction probability and a classification threshold for each modeled class.

PLS-DA modeling algorithm is based on PLS regression. A PLS-DA model can be calculated by regressing \( y \) (a column of \( Y \)) on \( X \) using an adequate number of factors (LVs), \( a \). In the dependent variable vector \( y \), the class of each sample is coded with the integer 1 if the sample belongs to the class of interest (class 1), or 0 if the sample belongs to a different class (class 0). When the training samples belong to \( G > 2 \) classes, one of the classes is defined as the class of interest (class 1) and the remaining \( G - 1 \) classes are included together in ‘class 0’. A PLS-R model is calculated with this training set to relate the independent variables to the integer \( y \) that designates the class of the sample. PLS-DA classification of an unknown sample is derived from the value \( \hat{y} \) predicted by the PLS-R model. This value is a real number, not an integer, which
should be ideally close to zero if the sample belongs to ‘class 0’ and close to one if the sample belongs to ‘class 1’. A cut-off value between 0 and 1 is established so that a sample is assigned to ‘class 1’ if the prediction is larger than the cut-off value, or assigned to ‘class 0’ otherwise. The simplest approach is to use an arbitrary cut-off value, such as 0.5. An advanced approach is to assume that the predictions for each logical class in the training set follow a Gaussian distribution. Then, the mean and the standard deviation of these predictions are used to estimate a probability density function (PDF) for each class (Pérez et al., 2009). The two constructed PDFs are close to representing the true profiles of all samples in the populations of ‘class 0’ and ‘class 1’. This allows to calculate $P(\hat{y}|1)$, the probability of observing a value of $\hat{y}$ given a sample from ‘class 1’, and $P(\hat{y}|0)$, the probability of observing a value of $\hat{y}$ given a sample from ‘class 0’. These two probabilities are estimated from the $y_i$ values observed in the training data.

Density methods (Coomans and Broeckaert, 1986; Coomans and Massart, 1981) or kernel methods are classification methods that are based on calculating a potential function around a point. A potential function can be calculated for each calibration sample $i$ with the shape of a Gaussian curve, which is a commonly used kernel function (Coomans and Broeckaert, 1986), centered at $\hat{y}_i$ and with standard deviation (smoothing parameter) equal to $SEP_i$ (Standard Error of Prediction for sample $i$):

$$SEP_i = \text{RMSEC}_a \sqrt{1 + h_i}$$

where RMSEC$_a$ is the root mean square error of calibration and $h_i$ is the leverage of the sample (Botella et al., 2009). The leverage is a measure of the influence of an observation on the PLS model (Ericsson et al., 2006b). It is computed as

$$h_i = t_i^T(T^T T)^{-1} t_i$$

where $t_i$ is the scores vector for sample $i$ on $a$ factors (LVs) and $T$ is the scores matrix of the data ($X$) for the $a$ factors (Pérez, et al., 2009). The width of the Gaussian kernel for sample $i$ depends on $SEP_i$, which is particular to that sample, and depends on the relative position of the sample in the multivariate space (Botella et al., 2009). The PDF of ‘class 0’, $P(\hat{y}|0)$, is calculated as the average of the individual kernel functions of the $N_0$ training samples of ‘class 0’; likewise, $P(\hat{y}|1)$ is calculated referring to the $N_1$ training samples of ‘class 1’. Figure 6.1 shows the Gaussian functions calculated from the predictions of three training samples of ‘class 0’ and four samples of ‘class 1’, where $\hat{y}$ is the variable in the abscissa axis.
The predicted class for sample \( i \) is obtained using the Bayes formula. The probability that a sample with prediction \( \hat{y}_i \) belongs to the ‘class 0’, or ‘class 1’, is given by (Duda et al., 2000):

\[
P(0|\hat{y}_i) = \frac{P(\hat{y}_i|0) \cdot P(0)}{P(\hat{y}_i)} = \frac{P(\hat{y}_i|0) \cdot P(0)}{P(\hat{y}_i|0) \cdot P(0) + P(\hat{y}_i|1) \cdot P(1)}
\]

or

\[
P(1|\hat{y}_i) = \frac{P(\hat{y}_i|1) \cdot P(1)}{P(\hat{y}_i)} = \frac{P(\hat{y}_i|1) \cdot P(1)}{P(\hat{y}_i|0) \cdot P(0) + P(\hat{y}_i|1) \cdot P(1)}
\]

where \( P(\hat{y}_i|0) \) and \( P(\hat{y}_i|1) \) are the conditional probabilities evaluated from the PDFs of classes 0 and 1 (Botella et al., 2009), \( P(0) \) and \( P(1) \) are the a priori probabilities that can be calculated from the training data set after assuming that the number of samples of each class in the training set is representative of the entire population, i.e. \( P(0) = N_0/(N_0 + N_1) \) and \( P(1) = N_1/(N_0 + N_1) \) (Pérez, et al., 2009); the probability of observing 1 or 0 (in the future) is similar to how many samples of 1 and 0 were in the original data set. The law of total probability \( P(\hat{y}) = P(\hat{y}|0)P(0) + P(\hat{y}|1)P(1) \) has been used; it expresses the total probability of an outcome which can be realized via several distinct events.

The Bayes rule assigns the sample to the class (0 or 1) in which it has the highest a posteriori probability of belonging (Duda et al., 2000):

‘class 0’ if \( P(0|\hat{y}_i) > P(1|\hat{y}_i) \)

‘class 1’ if \( P(1|\hat{y}_i) > P(0|\hat{y}_i) \)

This rule is optimal in the sense that no other rule can yield a lower error probability (Botella et al., 2009). The Bayesian decision rule tries to minimize the probability of error that may occur when assigning a sample to a class.

Figure 6.2 shows the a posteriori probabilities \( P(0|\hat{y}) \) and \( P(1|\hat{y}) \), where \( \hat{y} \) is the variable in the abscissa axis.
Since each sample for which the prediction is made belongs to one of the two classes (0 is the complementary event of 1), then \( P(1|\hat{y}) + P(0|\hat{y}) = 1 \), that is, the probabilities are normalized to 1. There is a single point where \( P(0|\hat{y}) = P(1|\hat{y}) = 0.5 \). This point can be selected as a threshold for the PLS-DA.

When the probability that the sample belongs to ‘class 0’ is similar to the probability that it belongs to ‘class 1’, there is a high risk of misclassification and the reliability of the classification is low; in these situations the reject option might be an advantageous addition to the decision rule.

Since \( P(0|\hat{y}_i) \) and \( P(1|\hat{y}_i) \) formulae have the same denominator, the classification rule can be rewritten as:

\[
\begin{align*}
\text{‘class 0’} & \quad \text{if } P(\hat{y}_i|0) \cdot P(0) > P(\hat{y}_i|1) \cdot P(1) \\
\text{‘class 1’} & \quad \text{if } P(\hat{y}_i|0) \cdot P(0) < P(\hat{y}_i|1) \cdot P(1)
\end{align*}
\]

By combining the PDFs and the Bayes theorem (Bishop, 2006) the cut-off is defined as the value of \( \hat{y} \) at which the \textit{a posteriori} probability of both classes is equal (Wise et al., 2005). This cut-off value depends on the distribution of the predictions. At the point \( \hat{y}_{\text{cut-off}} \), the threshold value is

\[
P(\hat{y}_{\text{cut-off}}|0) \cdot P(0) = P(\hat{y}_{\text{cut-off}}|1) \cdot P(1)
\]

\text{i.e.}

\[
P(\hat{y}_{\text{cut-off}}|0) \cdot N_0/(N_0 + N_1) = P(\hat{y}_{\text{cut-off}}|1) \cdot N_1/(N_0 + N_1)
\]

Figure 6.3 shows the probability density functions \( P(\hat{y}_i|0) \) and \( P(\hat{y}_i|1) \) for the classes 0 and 1 (on the left) and each of them multiplied by the \textit{a priori} class probability (on the right) \( P(\hat{y}_i|0) \cdot P(0) \) and \( P(\hat{y}_i|1) \cdot P(1) \), where \( \hat{y} \) is the variable in the abscissa axis.
The two distributions typically cross in only one place, unless one is really broad in comparison to the other, in which case they will cross twice.

The classification decision is based on the actual predicted value. However, ideally it should be based on the true value $y_i$.

When the calibration contains more than two classes, thresholds to distinguish all classes will be determined. In this case, the primary misclassification threat is from the adjacent classes.

Some authors (Wang et al., 1999; Eker et al., 2001) used PLS scores and the Bayesian decision theory to calculate a posteriori probabilities and assigned a sample to the class where the a posteriori probability is the largest.

6.3 Variable Importance in the Projection (VIP)

The aim of PLS is to predict one or more response variables ($Y$) from the predictor data ($X$). In addition to regression coefficients used in describing the relationship between $X$ and $Y$ blocks, Variable Importance in the Projection (VIP) can be used; it compactly summarizes the importance of the $X$-variables for both the $X$ and $Y$-model parts (Ericsson et al., 2006a). VIP is a weighted sum of squares of the PLS weights, taking into account the amount of explained $Y$-variance of each PLS component, $a$ (Ericsson et al., 2006b). In this way, the VIP parameter identifies the relative importance of the $X$ when predicting $Y$ (Peolsson, 2008).

In case of $M$ response variables, $K$ predictors and $A$ latent variables, the VIP is defined for each $k$-th predictor variable, $x_k$ ($k = 1,.., K$), as (Tenenhaus, 1998; Eriksson et al., 1999)

$$VIP_k = \frac{K \sum_{a=1}^{A} \sum_{m=1}^{M} R^2(y_m t_a) w^2_{ak}}{\sum_{a=1}^{A} \sum_{m=1}^{M} R^2(y_m t_a)}$$

where $R^2(y_m t_a)$ (coefficient of determination) stands for the squared correlation between items in vector $y_m$ and $t_a$, whereas $w^2_{ak}$ measures the contribution of each variable $x_k$ to the $a$-th PLS component (in the PLS regression algorithm $w_a$ is a vector of norm 1). Thus, $VIP_k$ quantifies the influence on the response of each variable summed over all components and categorical responses (for more than two categories in $Y$), relative to the total sum of squares of the model (Perez-Enciso and Tenenhaus, 2003).

In case of a single response $y$, the VIP score for the $k$-th variable can be calculated as (Zhang et al., 2006)
\[ \text{VIP}_k = \frac{K \sum_{\alpha=1}^{A} R^2(y; t_\alpha)w_{ak}^2}{\sum_{\alpha=1}^{A} R^2(y; t_\alpha)w_{ak}^2} \]

where \( R^2(y; t_1, \ldots, t_a) = \sum_{\alpha=1}^{A} R^2(y; t_\alpha) \) and \( R(y; t_\alpha) = \text{corr}(y, t_\alpha) \), or by eq. (Chong and Jun, 2005; Afanador et al., 2013; Peerbhay et al., 2013)

\[ \text{VIP}_k = \frac{K \sum_{\alpha=1}^{A} \left( c_{\alpha}^2 t_\alpha^T t_\alpha \cdot \left( \frac{w_{ak}}{\|w_{ak}\|} \right)^2 \right)}{\sum_{\alpha=1}^{A} c_{\alpha}^2 t_\alpha^T t_\alpha} \]

where the expression \( c_{\alpha}^2 t_\alpha^T t_\alpha \) computes the variance explained by each component, and \( \left( \frac{w_{ak}}{\|w_{ak}\|} \right)^2 \) represents the importance of the \( k \)-th variable (Mehmood et al., 2012).

Since \( w_{ak}^2 \) (or \( \left( \frac{w_{ak}}{\|w_{ak}\|} \right)^2 \)) takes a large value if \( x_k \) is important in building up \( t_\alpha \), and \( R^2(y; t_\alpha) \) is large if \( t_\alpha \) is strongly correlated with \( y \), then VIP\(_k\) is large in that situation (Zhang et al., 2006). In addition, for given \( y \) and \( X \), VIP has also the property that the sum of squares of all the scores is equal to the number of terms in the model, i.e.:

\[ \sum_{k=1}^{K} \text{VIP}_k^2 = K; \] it is a constant. Hence, the average VIP value would be equal to 1. Predictors with a VIP score larger than 1 may be most influential for the model. (Ericsson et al., 2006a); it does not mean that a variable with a low VIP is not relevant for that model. The VIP parameters are not used to point out an absolute cut-off number, but used with the purpose of demonstrating the most important variables explaining the prediction variables (Peolsson, 2008). So, for the response variable \( y \), the explanatory variables with larger VIP\(_k\) values tend to be more important than others. That reflects the relative importance of the input variables.
REFERENCES


Hand D J. Construction and Assessment of Classification Rules. 1997, Chichester: Wiley


Stähle L, Wold S. Partial least squares analysis with cross validation for the two-class problem: a Monte Carlo study, J. Chemom., 1987, 1, 185-96.


POSTERS AND MANUSCRIPTS

Posters in conferences

21-st European Biomass Conference and Exhibition, Copenhagen, Denmark, June 3-7 2013.


Manuscripts


STUDY OF WOOD CHIP SIZES: SEARCHING FOR A METHOD BASED ON IMAGE ANALYSIS

L. Pari, P. Febbi, A. Assirelli, V. Civitarese, E. Santangelo
Consiglio per la ricerca e la sperimentazione in agricoltura - Via della pascolare 16, 00016 Monterotondo (Roma), Italy

Introduction
The size distribution of wood chips is recognized as one of the most important parameters for efficient combustion since it affects the storage, the efficiency of energy conversion and environmental emissions. The standard method for the determination of the size distribution (UNI EN 15149-1:2011) may have a resolution and repeatability that can be subject to a certain degree of variability and uncertainty. Image analysis can give highly accurate measure of the axial dimensions defined for each individual particle. Since the chip form influences the sieve results, the image analysis could provide a new method which is sensitive to the geometrical form for determining a more accurate measure of size integrated with form. CRA-ING (Consiglio per la ricerca e la sperimentazione in agricoltura) is conducting a research activity for developing a new method, based on the image analysis, for the classification of wood particles, depending on both size and shape.

Materials and Methods
A sample of poplar chips (8 liters) was sieved through horizontally oscillating sieves, using five different screen hole diameters (3.15, 8, 16, 45, 63 mm). The wood chips were sorted in decreasing size classes and the cumulative particle size distribution of wood chips was determined.

The true size distributions of a wood chip subsample, according to length and width, have been determined by hand using a digital caliper and by weighing each of the chips. A comparison is then made between the traditional sieve result and the hand measured sizes (length and width).

Digital images of all chips in the subsample have been acquired using a digital scanner (Fig. 1). 2-D numerical data have been processed by software and a sequence of pixels has been extracted from each shape. The number of pixels has been converted to area and lengths (Feret diameters). The form of chips has been transformed in distance coefficients in order to be compared.

Cumulative curves of the size distribution based on image analysis use the actual lengths of Feret diameters (Max and min) to construct the curves. Hand measured length and width, Feret diameters (Max and min) and particle weight are used to construct the size-distribution curves with respect to the cumulative percent mass (Fig. 2) and area (Fig. 3).

The median value of the particles is established by the intersection of the cumulative distribution curve with the 50%-line.

Results
A comparison among sieve, image analysis and hand-made measurement results shows a lateral displacement among the curves. The curves relate the different sizes of the chips. The further apart the curves are the greater the difference between the three axial dimensions of the chip is (Fig. 2 and 3).

The sizes of Feret min and width are very similar. If the wood chips are to be classified by width, there is some compliance of the horizontal screening results with the results from image analysis method (Fig. 4).

The first results, deriving from a comparison among the standard method, the image analysis and the hand-made measurement of particle size, show that the standard technique (using oscillating sieves) is not devoid of inaccuracies that may lead to significant deviations from the real distribution of the particle size, whereas the use of image analysis assures a better approximation to the curve obtained from manual measurement.

Conclusions
The image analysis size is considered to be more accurate that the hand measured size, due to human error and subjectivity associated with the hand measurements. With image analysis it is possible to quantify size and shape distribution of all the particles more accurately. In order for the industry to be able to use image analysis results, it is necessary to correlate sieve size with image-analysis results. CRA-ING is looking for a new method in order to determine sieve size from image analysis.

This is a preliminary work whose main objective will be, in the coming years, to shift from a 2-dimensional static method (which is the one studied so far) to a 3-dimensional dynamic acquisition process (that is, with the transposition of the moving image), being the final goal the setting up of a fully automated system for product classification.

Acknowledgements
The author would like to thank Dr. Paolo Menesatti and Dr. Corrado Costa for their support and software analysis.
Determining Wood Chip Size: Image Analysis and Clustering Methods

Paolo Febbi, Corrado Costa, Paolo Menesatti, Luigi Pari
Consiglio per la Ricerca e la sperimentazione in Agricoltura
via della pascolare 16, 00016 Monterotondo (Roma), Italy

The oscillating screen method (EN 15149-1:2010) is one of the standard methods for the determination of the size distribution of wood chips. Recent literature demonstrated how image analysis could return highly accurate measure of the dimensions defined for each individual particle, and could promote a new method depending on the geometrical shape to determine the chip size in a more accurate way.

Five samples of wood chips (8 litres each) were sieved through horizontally oscillating sieves; the wood chips were sorted in decreasing size classes and the mass of all fractions was used to determine the size distribution of the particles.

Since the chip shape and size influence the sieving results, Wang’s theory was considered: geometric forms (rectangles, triangles and diamonds) fall into different areas of a plot of the shape factor, $S_h$ (the ratio between the particles perimeter and area, $P/A$), against the ratio between maximum and minimum Feret diameters, $D_{\text{Max}}/D_{\text{min}}$.

A cluster analysis on the shape and size descriptors was applied to observe the chips distribution according to their descriptors derived from image analysis. The UPGMA algorithm was applied on Euclidean distance. The obtained dendrogram (on the right) shows a group separation according with the original three sieving fractions (in yellow, green and red). The table gives a comparison between the traditional sieve and clustering results.

<table>
<thead>
<tr>
<th>Sieving</th>
<th>Fraction 8+16</th>
<th>Fraction 16+45</th>
<th>Fraction 45+63</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction 8+16</td>
<td>483</td>
<td>-</td>
<td>-</td>
<td>483</td>
</tr>
<tr>
<td>Fraction 16+45</td>
<td>-</td>
<td>220</td>
<td>-</td>
<td>220</td>
</tr>
<tr>
<td>Fraction 45+63</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clustering</th>
<th>Fraction 8+16</th>
<th>Fraction 16+45</th>
<th>Fraction 45+63</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction 8+16</td>
<td>478</td>
<td>35</td>
<td>-</td>
<td>514</td>
</tr>
<tr>
<td>Fraction 16+45</td>
<td>5</td>
<td>184</td>
<td>-</td>
<td>189</td>
</tr>
<tr>
<td>Fraction 45+63</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

This preliminary result shows how the image analysis-based method has a high potential for the characterization of wood chip size distribution and could be further investigated. An improvement of the results is expected by using supervised multivariate methods that utilize known class memberships.

The main objective of the future activities will be to shift the analysis from a 2-dimensional method to a 3-dimensional acquisition process.
Determining Wood Chip Size: Image Analysis and Clustering Methods
Paolo Febbi*, Corrado Costa, Paolo Menesatti, Luigi Pari
Consiglio per la Ricerca e la sperimentazione in Agricoltura, Unità di ricerca per l’ingegneria agraria, Via della pascolare 16, 00016 Monterotondo (Roma), Italy.
*Corresponding author: Paolo Febbi, UniTus-DAFNE, e-mail: paolo.febbi@entecra.it

Abstract
One of the standard methods for the determination of the size distribution of wood chips is the oscillating screen method (EN 15149-1:2010). Recent literature demonstrated how image analysis could return highly accurate measure of the dimensions defined for each individual particle, and could promote a new method depending on the geometrical shape to determine the chip size in a more accurate way. Five samples of wood chips (8 litres each) were sieved through horizontally oscillating sieves, using five different screen hole diameters (3.15, 8, 16, 45, 63 mm); the wood chips were sorted in decreasing size classes and the mass of all fractions was used to determine the size distribution of the particles. Since the chip shape and size influence the sieving results, Wang’s theory, which concerns the geometric forms, was considered. A cluster analysis on the shape descriptors (Fourier descriptors) and size descriptors (area, perimeter, Feret diameters, eccentricity) was applied to observe the chips distribution. The UPGMA algorithm was applied on Euclidean distance. The obtained dendrogram shows a group separation according with the original three sieving fractions. A comparison has been made between the traditional sieve and clustering results. This preliminary result shows how the image analysis-based method has a high potential for the characterization of wood chip size distribution and could be further investigated. Moreover, this method could be implemented in an online detection machine for chips size characterization. An improvement of the results is expected by using supervised multivariate methods that utilize known class memberships. The main objective of the future activities will be to shift the analysis from a 2-dimensional method to a 3-dimensional acquisition process.

Keywords: image analysis, wood chip, size distribution, clustering, sieve result.
Introduction

The size distribution of wood chips is recognized as one of the most important parameters for efficient combustion since it affects the storage, the efficiency of energy conversion and environmental emissions (Nati et al., 2010). The dimensions of wood chips are specified by the international standard EN 14961-1:2010 (CEN, 2010) (Tab. 1).

Table 1: Specification of the dimensions for wood chips

<table>
<thead>
<tr>
<th>Particle size distribution</th>
<th>Main fraction (min. 75 w-%)</th>
<th>Cross sectional area</th>
<th>Coarse fraction, max length of particle</th>
</tr>
</thead>
<tbody>
<tr>
<td>P16</td>
<td>3.15÷16 mm</td>
<td>&lt; 1 cm²</td>
<td>&lt; 31.5/12 mm</td>
</tr>
<tr>
<td>P45</td>
<td>8÷45 mm</td>
<td>&lt; 5 cm²</td>
<td>&lt; 120/350 mm</td>
</tr>
<tr>
<td>P63</td>
<td>8÷63 mm</td>
<td>&lt; 10 cm²</td>
<td>&lt; 350 mm</td>
</tr>
<tr>
<td>P100</td>
<td>16÷100 mm</td>
<td>&lt; 18 cm²</td>
<td>&lt; 350 mm</td>
</tr>
</tbody>
</table>

For fuel chips, mechanically or manually operated screening devices are commonly applied and there is a large variety of applicable systems. The reference methods for size classification of samples are EN 15149-1:2010 (CEN, 2010; oscillating screen method) and CEN/TS 15149-3:2006 (CEN, 2006; rotary screen method). The average relative repeatability limits for horizontal (< 2 w-%) and rotary screenings are exceedingly low. Generally, the relative reproducibility limits based on the median values for horizontal (< 10 w-%) and rotary screening results seem acceptable. On average, reproducibility is better for horizontal screenings than for rotary screening. However, if different measuring principles are applied, particle size analysis of biofuels is associated with high measuring uncertainties (Hartmann et al., 2006).

Since the chip form (shape and size) influences the sieve results, the image analysis could provide a method which is sensitive to the geometrical shape for determining a more accurate measure of size integrated with shape (Fernlund, 1998).

Consiglio per la Ricerca e la sperimentazione in Agricoltura, Unità di ricerca per l’ingegneria agraria (CRA-ING) is conducting a research activity for developing a new method, based on the image analysis, for the classification of wood particles, depending on both size and shape.
Materials and Methods

Five samples of wood chips (8 litres each) were sieved through horizontally oscillating sieves, using five different screen hole diameters (3.15, 8, 16, 45, 63 mm). The wood chips were sorted in decreasing size classes and the size distribution of wood chips was determined. The result is expressed as a percentage of the total mass of all fractions.

The test samples were divided and a wood chip subsample of 706 chips was processed in sequential sieving operations for the determination of its size distribution. Three fractions were considered: 8÷16 mm, 16÷45 mm, 45÷63 mm. Digital images of the 706 chips in the subsample were acquired using a high resolution digital scanner A3 (600 dpi; Fig. 1). 2-D numerical data were processed in Matlab (rel. 7.1) environment, specifically developed by CRA-ING.

![Figure 1: Scanned image of chips (on the left) and segmented image (on the right)](image)

After image segmentation, the following parameters from each object were extracted: size descriptors (area, perimeter, Feret diameters, eccentricity, etc.) and shape descriptors (99 Fourier descriptors). The Feret diameter is defined as the distance between two parallel tangential lines restricting the object perpendicular to that direction; it measures a particle size along a specified direction. To determine the length and width of a particle area, the algorithm measures Feret diameters in 32 azimuth directions and takes the longest distance as D_{\text{Max}} (length) and the shortest distance as D_{\text{Min}} (width). The Fourier coefficients can summarize the shape of an object in the frequency domain. Complex shapes can be represented with a small number of invariant coefficients, which can be viewed as features extracted from the original shapes. Generally, a subset of the components is often enough to capture the overall features of the shape and discriminate different shapes (Zhang and Lu, 2002).
Wang (Wang, 1994) presented a theory suggesting that geometric forms, rectangles, triangles and diamonds, should fall into different areas of a plot of the shape factor, $Sh$ (the ratio between the particles perimeter and area, $P/A$), against the ratio between $D_{\text{Max}}/D_{\text{min}}$.

In the present work, some ideal geometric shapes have been considered: rectangles, diamonds, squares and triangles. In order to obtain scale invariant contours from forms, each of them was normalized by the length of $D_{\text{min}}=1$; so, the shape classification is size independent. This is represented in figure 2, where diamonds, rectangles and two different kinds of triangles are reported at ratio $D_{\text{Max}}/D_{\text{min}}$ varying in the range $\sqrt{2}$÷10.

![Figure 2: Different areas occupied by the geometric shapes](image)

Fernlund (Fernlund, 1998) reported a plot of the shape factor ($Sh$) against the ratio between $length$ and $width$ made for rectangular, triangular and diamond shaped particles. Mathematically the different shapes fall into different areas in this diagram. However, the geometric forms are not easily differentiated using the 32 azimuth directions to compute the Feret diameters. The computed values do not agree with the theoretical ones and so there is a great deal of overlap. The reason for this is assumed to be that the particles are more irregular than they are similar to the typical geometric forms and their perimeters and areas do not agree with the theoretical perimeters and areas (Jansson and Muhr, 1995). It has been demonstrated that in a two dimensional plot it’s quite difficult to distinguish enough the different shapes. Wang (Wang, 1994) considered four descriptors (perimeter, area, $D_{\text{Max}}$, $D_{\text{min}}$) to represent forms in a 2-dimensional space; the relative positions of objects in that space cannot always be distinguished. This method reduces the reference space, consequently reducing the
distances between objects. Distances between shapes in a 2-dimensional space are smaller than in a $n$-dimensional space.

To discriminate the real forms of chips, it could be helpful to represent the objects in a multidimensional way, with many axes. This is equivalent to dilating the reference space, but augmenting the complexity of the analyses. The shape of chips can be transformed in similarity coefficients in order to quantify the resemblance between objects. Similarities are higher when the two objects are identical and lower when the two objects are completely different. A similarity index ($S$) can be transformed into a distance ($D$) by computing its one-complement. For a similarity measure varying between 0 and 1, the corresponding distance may be computed as $D = 1 − S$ (Legendre, 1998).

In an Euclidean space, the objects (chips) may be represented along axes using shape and size descriptors (Costa et al., 2011). The relative position of an object with respect to another one may be used to obtain clusters. Clustering objects is a multivariate operation which partitions the objects into two or more subsets (clusters), using pre-established rules of agglomeration or division, such that each object belongs to one and only one subset of that partition. The subsets form a series of mutually exclusive classes, among which the objects are included.

The adopted clustering method is based on the Euclidean distances among objects, computed using shape and size descriptors (109 variables). The hierarchical clustering routine produces a dendrogram showing how objects can be clustered. A dendrogram is made of branches that meet at nodes which are drawn at the similarity value where fusion of branches takes place.

The dendrogram based on Euclidean distance was performed with the Unweighted Pair-Group Average (UPGMA) algorithm. The highest similarity, or smallest distance, identifies the next cluster to be formed. Following this event, the method computes the arithmetic average of the similarities or distances between a candidate object and each of the cluster members or, in the case of previously formed cluster, between all members of the two clusters. All objects receive equal weights in the computation. Clusters are joined based on the average distance between all members in the two groups. Clustering proceeds by agglomeration as the similarity criterion is relaxed. Each cluster collects objects (chips) that are sufficiently similar, given the variables considered.
Results
The result of cluster analysis (UPGMA based on Euclidean distance) on both shape descriptors (Fourier descriptors) and size descriptors (area, perimeter, Feret diameters, eccentricity, etc.) is reported in figure 3 to observe the chips distribution according to their descriptors derived from image analysis. For graphical convenience, vertical lines are used to connect branches at the similarity levels of the nodes. Figure 3 represents a summarized version of the whole dendrogram, being it composed by 706 observations. The separation of the main groups is in accordance with the original three sieving fractions. The percentage of chips classified according to the sieving procedure grouping is reported.

<table>
<thead>
<tr>
<th>Euclidean distance</th>
<th>1600</th>
<th>1200</th>
<th>800</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 (100 % Gr. 3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>189 (97.35 % Gr. 2; 2.65 % Gr. 1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>514 (93 % Gr. 1; 7 % Gr. 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: Summarized dendrogram for the cluster analysis on the chips descriptors (size and shape). The groups numbers refers to the sieving procedure grouping.

Comparing the original assignment of chips made by the sieve method to the assignment made by the clustering method (Tab. 2), it is possible to determine the number of chips classified in accordance. The method quality is expressed through the percentage of correctly classified units.

Table 2: Chips counts in the three fractions using sieving and clustering methods

<table>
<thead>
<tr>
<th>Sieving</th>
<th>Fraction 8÷16</th>
<th>Fraction 16÷45</th>
<th>Fraction 45÷63</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction 8÷16</td>
<td>483</td>
<td></td>
<td></td>
<td>483</td>
</tr>
<tr>
<td>Fraction 16÷45</td>
<td></td>
<td>220</td>
<td></td>
<td>220</td>
</tr>
<tr>
<td>Fraction 45÷63</td>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clustering</th>
<th>Fraction 8÷16</th>
<th>Fraction 16÷45</th>
<th>Fraction 45÷63</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction 8÷16</td>
<td>478</td>
<td>36</td>
<td></td>
<td>514</td>
</tr>
<tr>
<td>Fraction 16÷45</td>
<td>5</td>
<td>184</td>
<td></td>
<td>189</td>
</tr>
<tr>
<td>Fraction 45÷63</td>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Conclusions

The examined data sets consist of chips that belong to three different fractions or groups. A comparison was made between the traditional sieving and clustering results. The preliminary result shows that the image analysis-based method has a high potentiality for the characterization of wood chip size distribution and could be further investigated. The ability to observe particle groups in accordance with the fractions determined by the screen methods allows the development of a new method for the determination of chip size distribution compatible with the standards EN 15149-1/3. Moreover, this new method could be implemented in an online detection machine for chips size characterization.

Cluster analysis is an unsupervised method that attempts to identify classes (or groups) without using pre-established class memberships. Other works within different scientific fields demonstrated how the integration of complex shape descriptors (such as Fourier descriptors; Menesatti et al., 2009) with basic morphometric descriptors (Feret diameters, area, perimeter, etc.) could return better and more accurate descriptions of a phenomena.

In light of this, an improvement of the results is expected by using supervised multivariate methods that utilize known class memberships. Cumulative curves of the size distribution based on image analysis could be constructed. The main objective of the future activities will be to shift the analysis from a 2-dimensional method to a 3-dimensional acquisition process.

References


Automated determination of wood chip size distribution based on combined image and multivariate analyses

Paolo Febbi\textsuperscript{1,2}, Paolo Menesatti\textsuperscript{1}, Corrado Costa\textsuperscript{*1}, Luigi Pan\textsuperscript{1}, Massimo Cecchini\textsuperscript{2}

\textsuperscript{1}Consiglio per la Ricerca e la sperimentazione in Agricoltura - Unità di ricerca per l’ingegneria agraria - Via della Pascolare 16, 00015, Monterotondo scalo (Rome), Italy.
\textsuperscript{2}Università della Tuscia - Dipartimento di Scienze e Tecnologie per l’Agricoltura le Foreste la Natura e l’Energia - Via S. C. de Lellis s.n.c., 01100, Viterbo, Italy.

* Corresponding author: Corrado Costa, e-mail corrado.costa@entecra.it

Abstract

The European technical standard EN 14961 on solid biofuels determines the fuel quality classes and specifications for wood chips. Sieving methods are currently used for the determination of particle size distribution. Some authors suggested that image analysis tools could provide methods for a more accurate measure of size integrated with shape. This work for the first time analyzes how image analysis combined with multivariate modeling methods could be used to construct cumulative size distribution curves based on chip mass (or weight). This has been done through a Partial Least Squares Regression model for the weight prediction of poplar chips and Partial Least Squares Discriminant Analysis models for estimation of chips size classification. Images of 7583 chips were analyzed to extract size and shape descriptors (area, major and minor axis lengths, perimeter, eccentricity, equivalent diameter, fractal dimension index, Feret diameters and Fourier descriptors). The weight prediction model showed an high accuracy ($r = 0.94$). The chip classification based on three size fractions (8÷16 mm, 16÷45 mm and 45÷63 mm), with or without Fourier descriptors, showed accuracies equal to 92.9% of correct classification for both models in the independent test. The combination of image analysis with multivariate modeling approaches allow a better conversion of image analysis results to sieve results using the esteemed weight. The proposed method could allow to standardize processes applicable by biofuels laboratories and machinery certifiers.

Keywords: Cumulative size distribution curve; Sieving; Size classification; Biofuel quality determination; Modeling; Partial Least Squares.
Abbreviations

$C_{\text{min}}$ = chip width measured by digital caliper  
$C_{\text{max}}$ = chip length measured by digital caliper  
$D_{\text{min}}$ = minimum Feret diameter  
$D_{\text{max}}$ = maximum Feret diameter  
FD = Fourier descriptor  
LV = Latent Variable  
P = designation for particle size distribution  
PLS-DA = Partial Least Squares Discriminant Analysis  
PLS-R = Partial Least Squares Regression  
RMSEC = Root-Mean-Square Error of Calibration  
RMSECV = Root-Mean-Square Error of Cross-Validation  
RPD = Ratio of Percentage Deviation  
SRF = Short Rotation Forestry  
VIP = Variable Importance in the Projection

1 Introduction

The standard EN 14961-1:2010 determines the fuel quality classes and specifications for solid biofuels. The classification is based on the biofuel origin and source. Woody biomass is biomass from trees, bushes and shrubs that may only have been subjected to size reduction, debarking, drying or wetting. Solid biofuels are traded in many different sizes and shapes, which influence the handling of the fuel as well as its combustion properties. Energy conversion and emissions are also influenced by the particle sizes. In the case of wood chips, the properties of dimensions, moisture and ash content are normative in the specification, whereas other properties (net calorific value, bulk density, ash melting behavior) are informative [1]. Particle size is important also during storage, as it affects drying, calorific value and durability [2]. Together with moisture content, particle size distribution defines the product quality.

The European technical standard EN 14961-4: 2011 [3] on solid biofuels determines the fuel quality classes and specifications for non-industrial wood chips (used in smaller scale appliances, such as in households and small commercial and public sector buildings). The sensitivity to the fuel quality imposes a tighter specifications for small-size plants, whose small conveying ducts can be blocked by oversize particles [4].
An oscillating screen method (using sieve apertures of 1 mm and above) is currently used for the determination of particle size distribution, according to the standard EN 15149-1:2010 [5]. The results of sieving are presented in cumulative size distribution curves: the cumulative percentage mass of each fraction with respect to the total mass of all fractions versus the particle size in mm. Size-sorting by mechanical screening is used in many different industries, including conventional solid fuels, such as coal. However, the screening of wood chips is not a generalized practice yet [4].

Particle form (i.e., the combination of shape and size) influences if a particle passes a given sieve, and the least cross sectional area is an important factor that has effect on the wood chips sieving results. However, there are some disadvantages associated with sieve analysis, which is considered a crude method of determining size [6] and not giving an exact measure of any dimension of the particles. If a particle passes through a sieve is not only dependent upon its length and width but also on thickness and shape; as the size difference between width and thickness increases, the particle tends to pass through the sieve [7].

Image analysis could provide a method which is sensitive to the geometrical shape for determining a more accurate measure of size integrated with shape [6]. Image analysis methods are generally based on two-dimensional images of particle projection area and provide an accurate measure of particle size. Typically, the result of this analysis is presented in size distribution curves, which are not based on the cumulative percentage mass, as it is for sieve analysis, but refer to the percentage of the total particle projection area (or particle number). Dynamic online image analysis systems are considered particularly interesting because they can sort the particle sizes according to more than just one size parameter [8]. Hartmann et al. [8] suggested it would be useful to launch a standardization process in order to include the image analysis method to the scope of applicable standard laboratory principles for biofuels, too, so as to overcome the drawbacks of the screening methods. Following this suggestion, the aim of this study is to analyze how image analysis combined with multivariate modeling methods can be used to construct cumulative size distribution curves based on chip mass (or weight), which can be compared with the sieving results required by EN standards. This has been done through a Partial Least Squares Regression (PLS-R) model for the weight prediction of poplar chips and PLS Discriminant Analysis (PLS-DA) models for estimation of chips size classification. A comparison between two PLS-DA models was conducted to investigate the influence of the shape on the size fraction determination.
2 Materials and Methods

2.1 Sample preparation

Wood fuel chips consist of chipped woody biomass in the form of pieces with a defined particle size produced by mechanical treatment with sharp tools, such as knives. Chips have a sub-rectangular shape with a typical length of 5÷100 mm and a low thickness compared to other dimensions; wood chips in non-industrial situations have typical length of 5÷50 mm [3].

After air-drying, five 8-liter samples of wood chips deriving from the harvesting of short-rotation forestry (SRF) poplar (root four- and stem two- years old: R4S2) were considered. Whole trees without roots were comminuted with a drum chipper, prototyped by CRA-ING [9]. The samples were sieved (oscillating screen method) for the determination of particle size distribution, according to the standard EN 15149-1:2010 [5]. Sieves were used to separate the wood chips in six dimensional classes: <3.15 mm, 3.15÷8 mm, 8÷16 mm, 16÷45 mm, 45÷63 mm, 63÷100 mm. The total weight of the particles between the sieve intervals were determined with a precision scale and the result was expressed as a percentage of the total mass of all fractions.

Wood chips may be delivered in specific trade classes, mostly on the basis of the main fraction (minimum 75 weight-percentage): P16 (3.15÷16 mm), P45 (8÷45 mm), P63 (8÷63 mm), P100 (16÷100 mm). The designation symbol for particle size distribution (P) is used in combination with a number (P-class) for dimension referring to the particle sizes passing through the mentioned round hole sieve size. The average numerical value from the whole lot (or defined portion from the lot) determines the class to be used. The cross sectional area of the oversized particles shall be less than a given value, in cm² (Tab. 1). Gross fraction identifies the particles quantity with dimensions exceeding those reference values; fines fraction refers to the dimensions below the lower limit (< 3.15 mm).

<table>
<thead>
<tr>
<th>Particle size distribution</th>
<th>Main fraction (mm) (min. 75 w-%)</th>
<th>Cross sectional area (cm²)</th>
<th>Coarse fraction, max length of particle (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P16</td>
<td>3.15÷16</td>
<td>&lt; 1</td>
<td>&lt; 31.5/120</td>
</tr>
<tr>
<td>P45</td>
<td>8÷45</td>
<td>&lt; 5</td>
<td>&lt; 120/350</td>
</tr>
<tr>
<td>P63</td>
<td>8÷63</td>
<td>&lt; 10</td>
<td>&lt; 350</td>
</tr>
<tr>
<td>P100</td>
<td>16÷100</td>
<td>&lt; 18</td>
<td>&lt; 350</td>
</tr>
</tbody>
</table>
The commercial classes P16, P45 and P63 describe high grade chips, suitable for the feeding of small-size domestic and residential boilers [10]. Generally, more uniform products have been required.

While chips produced from logs always contain a smaller proportion of oversize particles and a higher proportion of accepts (chips in the selected P-class) [2], chips produced from tops, branches and small stems tend to present a higher incidence of oversize particles [11]. For this reason the analyzed chips were quite irregular in shape and size, and the sieved fractions were not very uniform.

2.2 Chip size determination

The 7583 particles of the three wood chip size fractions (i.e., dimensional classes 8÷16 mm, 16÷45 mm, 45÷63 mm) were divided in 2 sub-samples, by mass: a sub-sample corresponding to 1/8 of the mass of each fraction (706.4 g, 706 chips) and a sub-sample corresponding to remaining mass (4946.9 g, 6877 chips). Each single chip in the first sub-sample (706 chips) was weighed, at constant temperature and moisture content below 20 weight-percentage, with a resolution of 0.01 g; then individual length ($C_{\text{max}}$), width ($C_{\text{min}}$) and thickness were measured with a resolution of 0.1 mm, using a digital caliper (Borletti CDJ 15).

The chip length was measured as the maximum expansion of the particle when oriented in a stable position (longitudinal direction); the chip width was recorded as the second longest expansion (perpendicular to the longitudinal direction), while the thickness was the third longest expansion perpendicular to both length and width [8]. The particles were saved and numbered for future analysis. Chips in the size fraction that passed the 8 mm sieve were not individually measured. No wood chips in the 63÷100 mm dimensional class were obtained.

Digital images (Fig. 1-left) of all chips in both sub-samples were acquired using a high resolution (600 dpi) digital scanner A3 Epson GT-10000+. The 2-D numerical data were

![Figure 1: Wood chips; original scanned image (on the left) and image after segmentation (i.e., binarization) (on the right).](image)
processed in Matlab (rel. 7.1) environment. Images were segmented using the following procedure: i. a median filter (7 x 7) was applied to each RGB channel, ii. for each pixel an Euclidean distance was calculated basing on the RGB values and iii. a minimum error thresholding algorithm [12] was applied to binarize the image. After image segmentation (Fig. 1-right), the following 10 size descriptors were extracted from each object: area, major and minor axis lengths, perimeter, eccentricity, equivalent diameter, fractal dimension index [13], maximum, minimum and mean Feret diameters. Feret diameter is defined as the distance between two parallel tangential lines restricting the object perpendicular to that direction; it measures a particle size along a specified direction. The maximum and minimum Feret diameters, $D_{\text{max}}$ and $D_{\text{min}}$, are often used as the dimensions of the particles; they represent the longest ($\text{length}$) and intermediate ($\text{width}$) dimensions of the particle projected area. Typically, cumulative distribution curves of the size distribution based on image analysis use the actual lengths of Feret diameters. Pixels were converted into metric scales through a scale factor (25.4/600). Moreover, 99 Fourier coefficients were extracted; they summarize the shape of an object in the frequency domain. Complex shapes can be represented with a small number of invariant coefficients, which can be viewed as features extracted from the original shape boundaries [14]. Generally, a subset of the components of the Fourier descriptors (lower frequencies) is enough to capture the overall features of the shape and to discriminate the different shapes [15].

2.3 Multivariate modeling

PLS-R is a particular type of multivariate analysis which uses a two-blocks predictive PLS model. It relates the two data matrices ($X$ and $Y$) by a linear multivariate model and models also the structure of $X$ ($K$ column vectors: $x_1, \ldots, x_K$) and $Y$ ($M$ column vectors: $y_1, \ldots, y_M$); both these blocks are assumed to be, at least partly, modeled by the same latent variables (not directly observed or measured), LVs [16, 17]. The regression analysis objective is achieved by using the equation that minimizes the residual mean square error, or maximizes the coefficient of multiple determination $r^2$, which is the most commonly used statistic to measure the forecasting potential of a multiple regression equation [18]. The predictive ability of the model depends also on the number of latent vectors used. Because fit and prediction are different aspects of a model’s performance, Root-Mean-Square Error of Calibration (RMSEC) and Root-Mean-Square Error of Cross-Validation (RMSECV) values for PLS were calculated as a
function of the number of LVs in the model. RMSEC is a measure of how well the model fits the data; RMSECV is a measure of a model’s ability to predict new samples that were not used to build the model. Generally, a good predictive model should have high values of Pearson correlation coefficient \(r\), low values for RMSE and maximum Ratio of Percentage Deviation (RPD). RPD is the ratio of the standard deviation of the laboratory measured (reference) data to the RMSE of the cross-validation [19]. RPD values between 2.0 and 2.5 indicate very good, quantitative model and/or predictions; RPD values major than 2.5 indicate excellent model and/or predictions. [20].

PLS regression modeling was applied in order to estimate the weight of chips from size and shape descriptors obtained by image analysis. The model adopted for weight prediction was selected from 540 PLS linear regression models considering the combination among \(X\) pre-processings (Abs, Autoscale, Baseline, Detrend, Mean center, Median center, None, Normalize, Snv), \(y\) pre-processings (Autoscale, Median center, None) and number of LVs, from 1 to 20 (the pre-processing techniques are summarized in [21]). The PLS-R models were developed from a calibration set (training/evaluation set [22]) of 530 chips (75% of the 706-chip sub-sample) with 109 \(X\)-variables (10 size and 99 shape descriptors) and 1 \(y\)-variable (weight or mass). The PLS-R models (cross-validated) were then validated on an internal test set of 176 chips (25% of the 706-chip sub-sample). The partitioning were conducted optimally choosing the Euclidean distances based on the algorithm of Kennard and Stone [23] that selects objects without the \textit{a priori} knowledge of a regression model. The PLS-R model selection was mainly based on the efficiencies and robustness parameters described above. Once selected, the model was applied to the entire sample (7583 chips).

PLS-DA [24, 25, 26] is a PLS regression where the response variable is categorical, expressing the class membership of the statistical units. The objective of PLS-DA is to find a model, developed from a training set of observations of known class membership, that separates classes of objects on the basis of their \(X\)-variables. The multistate and qualitative response variable, \(y\), may be split into a set of dummy variables (\(Y\) block) whose number is equal to the number of categories or classes. Its modeling efficiency is the mean of sensitivity and specificity [27]. The sensitivity of the model is given by the number of samples predicted as in the class divided by number actually in the class (percentage of true positive); the specificity is given, per each category, by the number of samples predicted as not in the class divided by actual number not in the class (percentage of true negative).
Some authors did not observe a direct correlation between the intermediate axis of the particle, often defined as the ‘image analysis size’ of the particle, and sieve size; so, alternative methods were investigated [7]. The determination of particle size distribution, as reported in the oscillating screen method [5], provides a series of mutually exclusive size fractions (or classes) among which the chips are included. This response variable (the particle class) was the classification criterion used in discriminant analysis. PLS-DA was utilized as a supervised modeling method using SIMPLS algorithm [28] to calculate a model for correlating image-analysis with sieving results.

Two different PLS-DA analyses were conducted. The first one considering 109 X-variables: 10 size descriptors and 99 FDs; the second one considering only 10 explanatory variables, the same 10 size descriptors, without the 99 FDs. The models adopted for the size fraction prediction of chips were selected considering the combination between X pre-processing (Abs, Autoscale, Baseline, Detrend, Diff1, gls weighting, Groupscale, Log1suR, Mean center, Median center, Msc (mean), None, Normalize, Snv) and the number of LVs (the pre-processing techniques are summarized in [21]). There was no Y pre-processing. The models with FDs, using between 1 and 20 LVs, were 280; the models without FDs, using between 1 and 9 LVs, were 126. The chip membership to the three fractions was known before the analysis. The PLS-DA models were calibrated and validated on 100 chips of fraction 8÷16 mm, 100 chips of fraction 16÷45 mm and 25 chips of fraction 45÷63 mm; the 225 particles were randomly extracted from the 706-chip sub-sample. This dataset was divided into a calibration set of 169 chips (75% of each group) and an internal validation set of 56 chips (25% of each group). This was done optimally choosing the Euclidean distances based on the algorithm of Kennard and Stone [23] that selects objects without the a priori knowledge of a regression model. The percentages of correct classification were calculated for calibration and validation phases, and then used for model selection, in both analyses (with and without FDs). The PLS-DA model selection was mainly based on the efficiencies and robustness parameters described above.

In PLS-R and PLS-DA methods, a summary of the relative importance of the X-variables for both Y and X model parts is given by Variable Importance in the Projection (VIP) [29]. It is a weighted sum of squares of the PLS weights, taking into account the amount of explained Y-variance in each PLS component [30]. VIP has also the property of \( \sum_{k=1}^{K} VIP_k^2 = K \), where K is the number of predictor variables [31]; the average of squared VIP scores equals 1. The explanatory variables with larger VIP...
values tend to be more important than others [32], even if it does not mean that a variable with a low VIP is not relevant for the classification. In the case of PLS-DA, for each response variable, $y_m$ ($m$-th column vector of $Y$), a regression model on the $X$-components was considered. For each $k$-th predictor variable, the $VIP_k$ value quantifies the influence on the response of each variable summed over all components and categorical responses (for more than two categories in $Y$) [31]. The $VIP_k$ scores are a set of values equal in length to the number of $X$-variables included in the model ($K$) and were calculated according to Chong et al. [33]. To summarize the contribution of all the Fourier coefficients, a single summary variable was considered, whose value was calculated as the root square of the sum of square of all 99 FDs scores.

3 Results and discussion

The determination of particle size distribution (standard EN 15149-1:2010 [5]) requires samples (> 8 liters) taken from stock or from deliveries (e.g. shipload, truckload) in accordance with the sampling methods for solid biofuels (EN 14778:2011); methods for reducing combined samples (or increments) to laboratory samples and laboratory samples to sub-samples and general analysis samples, are described in ‘Solid biofuels - Sample preparation’ (EN 14780:2011). In addition, it is requested to identify the particles over 100 mm, to specify their number and size fraction, and to record the length of the longest particle overall. Image analysis accurately measures several parameters allowing to satisfy the standard requirements. Generally, image analysis is based on two-dimensional images of particles. Although the shape of the minimum projected area of a particle theoretically influences how particles pass the sieves [7], its automated measurement is not easy to carry out and also the particle could be so long that it cannot rotate and pass the screen holes vertically, even if it were thin enough to pass through.

3.1 Typical cumulative distribution curves

The manual measured length ($C_{max}$) and width ($C_{min}$), Feret diameters, $D_{max}$ and $D_{min}$, and particle weight (i.e., observed) were used to construct the size-distribution curves with respect to the cumulative percentage mass or to the cumulative percentage area (Fig. 2). The $x$-axis presents the chip / hole size (mm). For data evaluation and statistical processing, the median value of the particles is considered being less susceptible to outliers with respect to mean value [5, 8]. The median value is the observed particle size
of a sample that separates the cumulative size distribution into two equal parts (half of the particle mass is below and half is above); graphically, it is given by the intersection point of the cumulative size distribution curve with the 50%-horizontal line. The comparison among sieving, image analysis and manual measurements showed a lateral relative displacement among the curves. The curves relate the different sizes of the chips in the first sub-sample (706 chips). The further apart the curves are, the more the difference among the axial dimensions of the chips. Observing the median values, there is some compliance of the horizontal screening (sieve) result with the results from image analysis method ($D_{\text{min}}$).

Hartmann et al. [8] used a digital caliper as a reliable way of chip size determination and such manual measurements were applied to build a reference distribution curve. Figure 2 shows that the cumulative distribution curves based on $C_{\text{min}}$ and $D_{\text{min}}$ are very close each others. Even though manual measurements of chip width ($C_{\text{min}}$) are affected by errors and a slight tilting of the chip position changes its projection area, and then its $D_{\text{min}}$ size, the linear correlation coefficient (Pearson $r$) between $D_{\text{min}}$ and $C_{\text{min}}$ of the 706 chips was very high, 0.989. It is confirmed that image analysis sizing could be considered a more accurate way of measuring chips with respect to the manual one, due to human error and subjectivity associated with the manual measurements [34].
3.2 Weight (or mass) prediction

For the first time a cumulative distribution curve completely based on 2-D image analysis has been constructed. The PLS-R model adopted for the weight prediction presents the characteristics and principal results reported in Table 2. The selected model was not pre-processed. The estimated number of LVs was 10; this number of significant components minimizes the residual errors for the validation phase (RMSECV). The first two LVs, which capture most of the variance in X-block (99.97%) and Y-block (92.47%), have the maximum ability for the predictive model. The cumulated variance of X-block was 100%, and the cumulated variance of Y-block was 95.1%. The model correlation coefficient was 0.96, while the test correlation coefficient was 0.89; these values are relatively high. The model RPD_{RMSE} value was 3.65, while test RPD_{RMSE} value was 2.16; these values indicate an excellent / very good model [20]. The bias value was negligible (< 10^{-3}). The loading values and VIP scores obtained by the PLS regression showed that area and perimeter variables are truly significant for the model (data not shown).

Table 2: Characteristics and principal results of the Partial Least Squares Regression model to estimate the chips’ weight (or mass). LVs = Latent Vectors; RMSEC = Root-Mean-Square Error of Calibration; RMSECV = Root-Mean-Square Error of Cross-Validation; RPD = Ratio of Percentage Deviation.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of particles</td>
<td>530</td>
</tr>
<tr>
<td>n° LVs</td>
<td>10</td>
</tr>
<tr>
<td>Cumulated variance (%) X-block</td>
<td>100</td>
</tr>
<tr>
<td>Cumulated variance (%) Y-block</td>
<td>95.14</td>
</tr>
<tr>
<td>RMSEC</td>
<td>0.1937</td>
</tr>
<tr>
<td>RMSECV</td>
<td>0.2020</td>
</tr>
<tr>
<td>Bias</td>
<td>-0.00067</td>
</tr>
<tr>
<td>r model</td>
<td>0.96</td>
</tr>
<tr>
<td>r test</td>
<td>0.89</td>
</tr>
<tr>
<td>RPD_{RMSE} model</td>
<td>3.65</td>
</tr>
<tr>
<td>RPD_{RMSE} test</td>
<td>2.16</td>
</tr>
</tbody>
</table>

Figure 3-left shows the predicted weights vs. the observed weights of both 530 model and 176 test chips; the Pearson r value between the observed and predicted weights of the 706 chips was 0.94. The predicted weight of the 706-chip sample was 682.7 g, 96.6% of the observed weight. The deviation between observed and predicted weight of
chips was higher at higher weight values. Many existing methods based their conversion of image-analysis size to sieve size on the intermediate axis [7]. Typically, minimum Feret diameter ($D_{\text{min}}$) and manually weighted mass of each particle were used to construct the cumulative distribution curves with respect to the cumulative percentage mass [7, 8]. Figure 3-right shows the size-distribution curves with respect to the cumulative percentage mass, where the ‘$D_{\text{min}}$ predicted mass’ distribution curve refers to the predicted mass. The cumulative distribution curve completely based on 2-D image analysis is almost completely overlapping the cumulative distribution curve constructed from the manually weighted mass of each chip.

![Graph showing comparison between observed and predicted weight vs. size (mm)](image)

**Figure 3**: Comparison between the observed weight (or mass) and the predicted one using the Partial Least Squares Regression model (on the left). Comparison between cumulative distribution curves applied on the 706-chip sub-sample, referring to observed and predicted mass (on the right). Sieve refers to sieving; $D_{\text{min}}$ is the minimum Feret diameter.

### 3.3 Chip size fraction prediction

The two selected PLS-DA models (with and without FDs) adopted for the size fraction prediction of chips present the characteristics and principal results reported in Table 3. In both models, the $X$ and $Y$ blocks were not pre-processed. The models were selected not only considering their performances but hence considering their robustness (*sensu* [35]). The selected numbers of LVs, 12 and 9 respectively, minimized the respective RMSECV curves. The mean sensitivity and specificity were always high: 94.7% and 93.1% for the model with FDs, 95.1% and 92% for the model without FDs. Their efficiencies were 93.9% and 93.6%, respectively. The mean classification error was low and equal to 6.1% for the model with FDs and 6.4% for the model without FDs. The
mean percentage of correct classification, calculated on 169 chips (training/evaluation set), was 92.3\% for the model with FDs and 92.9\% for the model without FDs; the mean percentage of correct classification for the internal test (56 particles) was 92.9\% in both cases.

The discrimination ability is expressed through the percentage of well classified units. The random probability to assign a generic chip to one of three fractions is 33\%. Both PLS-DA models were applied to all 7583 wood chips (external test) to evaluate their ability to assign the chip size fraction more accurately than what would occur by chance. The classification results (confusion matrices) are reported in Table 4. The mean percentage of correct classification for the external test was 89.6\% for the model with FDs and 89.2\% for the model without FDs. The efficiencies of both models were slightly higher than 88\%.

Table 3: Characteristics and principal results of the selected Partial Least Squares Discriminant Analysis models, with Fourier descriptors (FDs) and without Fourier descriptors (no FD), to predict the three size fractions (8÷16, 16÷45, 45÷63). LVs = Latent Vectors; RMSEC = Root-Mean-Square Error of Calibration.

<table>
<thead>
<tr>
<th>PLSDA model</th>
<th>FDs</th>
<th>no FD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of particles</td>
<td>169</td>
<td></td>
</tr>
<tr>
<td>n° size fractions (Y-block)</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>n° LVs</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>X pre-processing</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Cumulated variance (%) X-block</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Cumulated variance (%) Y-block</td>
<td>82.2</td>
<td>80.3</td>
</tr>
<tr>
<td>Mean sensitivity (%)</td>
<td>94.7</td>
<td>95.1</td>
</tr>
<tr>
<td>Mean specificity (%)</td>
<td>93.1</td>
<td>92</td>
</tr>
<tr>
<td>Efficiency</td>
<td>93.9</td>
<td>93.6</td>
</tr>
<tr>
<td>Random probability (%)</td>
<td>33.3</td>
<td></td>
</tr>
<tr>
<td>Mean classification error (%)</td>
<td>6.1</td>
<td>6.4</td>
</tr>
<tr>
<td>Mean RMSEC</td>
<td>0.2382</td>
<td>0.2501</td>
</tr>
<tr>
<td>Mean % correct classification model</td>
<td>92.3</td>
<td>92.8</td>
</tr>
<tr>
<td>Mean % correct classification test</td>
<td>92.9</td>
<td>92.9</td>
</tr>
<tr>
<td>Mean % correct classification external test</td>
<td>89.6</td>
<td>89.2</td>
</tr>
</tbody>
</table>
Table 4: Confusion matrices of the 7583-chip external test (3 classes/fractions: 8÷16, 16÷45, 45÷63) obtained from the two Partial Least Squares Discriminant Analysis models with and without Fourier descriptors (FDs).

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted (with FDs)</th>
<th>Predicted (no FD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fraction 8÷16</td>
<td>Fraction 16÷45</td>
</tr>
<tr>
<td>Fraction 8÷16</td>
<td>4838</td>
<td>106</td>
</tr>
<tr>
<td>Fraction 16÷45</td>
<td>641</td>
<td>1936</td>
</tr>
<tr>
<td>Fraction 45÷63</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

The VIP scores of both models (12 LVs with FDs and 9 LVs without FDs) were reported in Figure 4. Figure 4-up shows the VIP_k parameters per each fraction of the model with Fourier descriptors, where the FDs variable summarizes the overall contribution of all the Fourier coefficients. All the VIP_k scores were higher than 1, demonstrating that all the explanatory variables were important for explaining the prediction variables. Area variable had the higher VIP scores, therefore it was the most significant variable in comparing the difference among the three size fractions.

![Figure 4](image_url): Partial Least Squares Discriminant Analysis models for size fraction prediction. VIP (Variable Importance in the Projection) scores of X variables for the three size fractions (8÷16, 16÷45, 45÷63), with Fourier descriptors in contracted form (i.e., root square of the sum of square of all 99 FDs scores) for the 12 LVs model (in the upper) and without Fourier descriptors for the 9 LVs model (in the lower). The y axes have different magnitude because refer to two different models, and the VIP scores express the relative importance of the X-variables within each model.
Figure 4-low shows the $VIP_k$ parameters per each fraction of the model without Fourier descriptors. The $VIP_k$ scores expressed a relative weight of the explanatory variables: area, equivalent diameter and mean Feret diameter were more important predictors in comparing the difference among the three considered fractions; eccentricity and fractal dimension index did not seem to be significant.

Even if the model with FDs is slightly more performing, there are not significant differences between the two models, when a round hole sieve is used for determination of particle size distribution. Some papers [36, 37] state that area, perimeter and Feret diameters are able to mathematically distinguish and differentiate the particle shapes.

The obtained results demonstrated that multivariate analyses are able to technically realize, in a multidimensional space, this shape distinction and to discriminate among classes, even if FDs are more informative.

3.4 *Sieving software simulation*

While traditional image analysis constructs size distribution curves with respect to cumulative percentage area or percentage number of particles [6, 8], sieve analysis is typically presented in percentage cumulative weight [5]. In order to use image analysis and construct size distribution curves based on cumulative percentage mass, some researchers [38, 39, 40, 41, 42] used the determination of the volume of a particle to calculate its mass. The results of these methods were considered not perfectly accurate [7].

Image and multivariate analyses recognize shapes and relate these results to sieve size. Considering the size fraction assignment obtained by the two PLS-DA models and the weight predicted by the PLS-R model to each particle of the entire sample (7583 wood chips), a conversion from image-analysis morphometries to sieving result was realized. This conversion depended upon an appropriate estimate of the chip weights, even if an exact determination of mass is not necessary for constructing the size distribution curve [7]. The predicted weight of the entire three-fraction sample (7583 chips) was 5539.3 g, 98% of the observed weight. The cumulative size distribution curves completely based on image analysis were reported in Figure 5, as result of 706-chip and 7583-chip tests. The cumulative curves of the two models are almost completely overlapping and very close to the traditional curve based on sieve analysis. They give a good quantification of particle size distribution. The advantage of image and multivariate analyses is to construct cumulative distribution curves showing a good agreement with the sieve
results and without the need of manual measurements, generally time consuming. A direct comparison between the result of the proposed method based on image analysis and traditional sieving result makes real.

Figure 5: Comparison among cumulative distribution curves determined by horizontal sieving (sieve - observed) and image analysis coupled with the Partial Least Squares Discriminant Analysis models, with (12LVs with FDs) and without Fourier descriptors (9LVs without FDs), applied on the 706-chip (on the left) and 7583-chip (on the right) samples.

4 Conclusions
The European Standard EN 14961-1:2010 concerns the general requirements of solid biofuel specifications and classes. Specification of the dimensions property for wood chips (particle size distribution) refers to the standard EN 15149-1:2010 concerning the oscillating screen method, which is subject to a relative inaccuracy. However, the screening of wood chips is carried out occasionally, often at a semi-experimental level, while in many other applications (food processing, fruit marketing, seed selection, construction, coal) mechanical screening is widely used and profitable, due to the higher value of the final product.

The proposed method, combining (for the first time within the biofuel sector) image analysis with multivariate modeling, could be implemented on an online detection machine for particle size characterization, revealing particularly helpful when the P-class to be specified (the number of property levels) shall be determined from large quantities (the whole lot) or when frequent sampling is required (e.g., for an internal quality system). More samples or large amounts would reduce the uncertainty arising from sampling and potentially could increase the quality and value of wood chips.

The detailed information of chip form and other dimensional aspects provided by the method proposed in this work could help: i. quality managers of large biofuel suppliers
or purchasers to check the fulfillment of particle size demands of the quality classes given in the standards, ii. chipper machine constructors to verify the prototype performances depending from different settings (knives position and number, cutting and feeding speeds, cutting and sharpness angles, anvil height, cutting direction, etc.) in a given experimental situation, not easily related to sieve size, and to optimize parameters of comminution devices in engineering situation, and iii. engineering machine certification in order to fix standard methodologies highly replicable. An improvement of the results is expected introducing a thickness measure, which could contribute to better express particle form and mass prediction, and to achieve an improved modeling. The main objective of the future activities will be to shift the analysis from a 2-dimensional method to a 3-dimensional acquisition process. A 3-D analysis could prove to be significant.

Acknowledgments
The authors would like to thank Angelo Del Giudice for the provision of sieved wood chips, and Andrea Acampora for his help in manual measurements.

References


Further analyses developed apart the manuscripts

Image analysis coupled with multivariate modeling has been developed to determine the quality classes (EN 14961-1:2010) and to specify the dimension (EN 15149-1:2010) of wood chips.

Image analysis is typically based on two-dimensional images of objects. Digital images of 7583 wood chips belonging to three size fractions (8÷16 mm, 16÷45 mm and 45÷63 mm) were acquired using a high resolution digital scanner, A3 Epson GT-10000+ (Fig. 1). After image segmentation (i.e. binarization), size and shape descriptors were extracted from each particle.

100 wood chips were randomly selected for the determination of moisture content (oven dry method, EN 14774-2); the mean value was about 9%, the standard deviation was about 0.84%.

PLS-R models were calculated with SIMPLS algorithm to predict chips’ weight (or mass). The predictive ability of a PLS-R model depends also on the number of latent vectors used by the model. Because fit and prediction are different aspects of a model’s performance, Root-Mean-Square Error of Calibration (RMSEC) and Root-Mean-Square Error of Cross-Validation (RMSECV) were calculated and plotted (Fig. 2) as a function
of the number of LVs in the considered PLS-R model (1+20). The estimated number of significant components (10 LVs) was based on the minimization of the residual errors for the validation phase (RMSECV).

![Figure 2: Relationship between Root-Mean-Square Error of Calibration and Root-Mean-Square Error of Cross-Validation curves in Partial Least Squares Regression model for mass prediction.](image)

The first two LVs of the selected model capture most of the variance in X-block (99.97%) and Y-block (92.47%). These two factors have the highest ability for the predictive model; their loadings are reported in Figure 3.

![Figure 3: Loadings of the first two Latent Variables in the PLS-R model for mass prediction.](image)

In Figure 4, the 10+99 VIP scores show and confirm that area and perimeter give a higher contribution to the PLS-R model construction.

![Figure 4: VIP (Variable Importance in the Projection) scores of X-variables for mass prediction in Partial Least Squares Regression model.](image)
In figure 5, the FDs variable summarizes the contribution of all the Fourier coefficients; its value was calculated as the root square of the sum of square of the 99 FDs scores, so that \( \sum_{k=1}^{10} VIP_k^2 = 109 \).

![Figure 5: VIP (Variable Importance in the Projection) scores of X-variables for mass prediction, with Fourier descriptors in contracted form, for the 12 LVs PLS-R model.](image)

PLS-DA was utilized as a supervised modeling method to elaborate a model able to correlate image-analysis with sieving results and then to predict the chips’ size fraction. The predictive ability depends also on the number of the LVs used in the model (1÷20). RSMEC and RSMECV were calculated and plotted (Fig. 6) as a function of the number of LVs in the considered PLS-DA model with FDs. The optimal value (12 LVs) maximizes the percentage of correct classification in the independent test set.

![Figure 6: Relationship between Root-Mean-Square Error of Calibration and Root-Mean-Square Error of Cross-Validation curves for size fraction prediction in Partial Least Squares Discriminant Analysis model.](image)

As a comparison, models without Fourier descriptors were also calculated. The loadings of the most significant latent variables (LV1, LV2, LV7) of the two selected models (with and without FDs) are reported in figures 7 and 8 respectively. These three LVs capture most of the variance in the X-block and Y-block.
Figure 7: Loadings of the three most informative Latent Variables in the Partial Least Squares Discriminant Analysis model, with 12 LVs and Fourier descriptors, for size fraction prediction.

Figure 8: Loadings of the three most informative Latent Variables in the Partial Least Squares Discriminant Analysis model, with 9 LVs and without Fourier descriptors, for size fraction prediction.

The model and test scores of the three latent variables (LV1, LV2, LV7) are reported in figure 9, respectively on the left and on the right.

Figure 9: Model scores (75% of the particles, on the left) and test scores (25% of the particles, on the right) of the 12 LVs Partial Least Squares Discriminant Analysis model.
Figure 10 shows the scores of the PLS-DA 12 LVs 7583-chip test.

![Figure 10: 7583-chip test scores of the 12 LVs Partial Least Squares Discriminant Analysis model.](image)

The multivariate analysis also expresses the VIP values calculated for both the selected models according to Chong and Jun, 2005. The VIP scores for each column of the calibration Y-block are reported in Figures 11 and 12. The VIP parameters show the most important predictor variables explaining the prediction variables. The lower and higher frequency Fourier descriptors (FD2, FD98, FD99) are characterized by relatively higher VIP scores.

![Figure 11: VIP (Variable Importance in the Projection) scores of X-variables of the three size fractions (8÷16, 16÷45, 45÷63), for the 12 LVs PLS-DA model with Fourier descriptors.](image)
Figure 12: VIP (Variable Importance in the Projection) scores of $X$ variables of the three size fractions (8÷16, 16÷45, 45÷63), for the 9 LVs PLS-DA model without Fourier descriptors.

The proposed PLS-R and PLS-DA models allow the quality requirements of wood chips to be assessed with regard to their dimensions.
GENERAL CONCLUSIONS
In this PhD thesis, a rigorous overview of Partial Least Squares Regression and Partial Least Squares Discriminant Analysis has been developed so as to provide the reader with the basic tools to use these techniques in various sectors of research and innovation (e.g., the classification of food products). An understanding in depth of their working principles allows the user to intervene on an informed basis, though in a limited manner, in the implementation and improvement of the final solution. The extensive bibliography provides the opportunity for additional insight.

The detailed information on the methods used for the evaluation of qualitative aspects of solid biofuels allow the obtained results to be replicated.

The proposed method, combining image analysis with multivariate modeling, could be extended to an on-line processing system based on a laser sheet-of-light (triangulation) technique and high speed cameras for three-dimensional form measurement, where the thickness measurement can help in achieving improved modeling performance.