SAPIENZA
UNIVERSITÀ DI ROMA

Dottorato di Ricerca in Statistica Metodologica

Tesi di Dottorato XXXI Ciclo – anno 2018/2019

Dipartimento di Scienze statistiche

Dimensionality reduction and simultaneous classification approaches for complex data: methods and applications

P.hD candidate
Dr. Mario Fordellone

Tutor
Prof. Maurizio Vichi

Contents

List of Figures

List of Tables

Abstract

Statistical learning (SL) is the study of the generalizable extraction of knowledge from data (Friedman et al. 2001). The concept of learning is used when human expertise does not exist, humans are unable to explain their expertise, solution changes in time, solution needs to be adapted to particular cases. The principal algorithms used in SL are classified in: *(i)* supervised learning (e.g. regression and classification), it is trained on labelled examples, i.e., input where the desired output is known. In other words, supervised learning algorithm attempts to generalize a function or mapping from inputs to outputs which can then be used speculatively to generate an output for previously unseen inputs; *(ii)* unsupervised learning (e.g. association and clustering), it operates on unlabeled examples, i.e., input where the desired output is unknown, in this case the objective is to discover structure in the data (e.g. through a cluster analysis), not to generalize a mapping from inputs to outputs; *(iii)* semi-supervised, it combines both labeled and unlabeled examples to generate an appropriate function or classifier.

In a multidimensional context, when the number of variables is very large, or when it is believed that some of these do not contribute much to identify the groups structure in the data set, researchers apply a continuous model for dimensionality reduction as principal component analysis, factorial analysis, correspondence analysis, etc., and sequentially a discrete clustering model on the object scores computed as K-means, mixture models, etc. This approach is called *tandem analysis* (TA) by Arabie & Hubert (1994).

However, De Sarbo et al. (1990) and De Soete & Carrol (1994) warn against this approach, because the methods for dimension reduction may identify dimensions that do not necessarily contribute much to perceive the groups structure in the data and that, on the contrary, may obscure or mask the groups structure that could exist in the data. A solution to this problem is given by a methodology that includes the simultaneous detection of factors and clusters on the computed scores. In the case of continuous data, many alternative methods combining cluster analysis and the search for a reduced set of factors have been proposed, focusing on factorial methods, multidimensional scaling or unfolding analysis and clustering (e.g., Heiser 1993,

De Soete & Heiser 1993). De Soete & Carroll (1994) proposed an alternative to the K-means procedure, named reduced K-means (RKM), which appeared to equal the earlier proposed projection pursuit clustering (PPC) (Bolton & Krzanowski 2012). RKM simultaneously searches for a clustering of objects, based on the K-means criterion (MacQueen 1967), and a dimensionality reduction of the variables, based on the principal component analysis (PCA). However, this approach may fail to recover the clustering of objects when the data contain much variance in directions orthogonal to the subspace of the data in which the clusters reside (Timmerman et al. 2010). To solve this problem, Vichi & Kiers (2001), proposed the factorial K-means (FKM) model. FKM combines K-means cluster analysis with PCA, then finding the best subspace that best represents the clustering structure in the data. In other terms FKM works in the reduced space, and simultaneously searches the best partition of objects based on the use of K-means criterion, represented by the best reduced orthogonal space, based on the use of PCA.

When categorical variables are observed, TA corresponds to apply first multiple correspondence analysis (MCA) and subsequently the K-means clustering on the achieved factors. Hwang et al (2007) proposed an extension of MCA that takes into account cluster-level heterogeneity in respondents' preferences/choices. The method involves combining MCA and k-means in a unified framework. The former is used for uncovering a low-dimensional space of multivariate categorical variables while the latter is used for identifying relatively homogeneous clusters of respondents. In the last years, the dimensionality reduction problem is very known also in other statistical contexts such as structural equation modeling (SEM). In fact, in a wide range of SEMs applications, the assumption that data are collected from a single homogeneous population, is often unrealistic, and the identification of different groups (clusters) of observations constitutes a critical issue in many fields.

Following this research idea, in this doctoral thesis we propose a good review on the more recent statistical models used to solve the dimensionality problem discussed above. In particular, in the first chapter we show an application on hyperspectral data classification using the most used discriminant functions to solve the high dimensionality problem, e.g., the partial least squares discriminant analysis (PLS-DA); in the second chapter we present the multiple correspondence K-means (MCKM) model proposed by Fordellone & Vichi (2017), which identifies simultaneously the best partition of the N objects described by the best orthogonal linear combination of categorical variables according to a single objective function; finally, in the third chapter we present the partial least squares structural equation modeling K-means (PLS-SEM-KM) proposed by Fordellone & Vichi (2018), which identifies simultaneously the best partition of the N objects described by the best causal relationship among the latent constructs.

Chapter 1

Partial Least Squares Discriminant Analysis: a dimensionality reduction method to classify hyperspectral data

1.1 Introduction

The recent development of more sophisticated spectroscopic approaches allows for the acquisition of high dimensional datasets from which valuable information may be extracted via different multivariate statistical techniques. The high data dimensionality greatly enhances the informational content of the dataset and provides an additional opportunity for the current techniques for analyzing such data (Jimenez & Landgrebe 1998). For example, automatic classification (clustering and/or classification) of data with similar features is an important problem in a variety of research areas such as biology, chemistry, and medicine (Hardy et al. 2006, Galvan et al. 2006). When the labels of the clusters are available, a supervised classification method is applied. Several classification techniques are available and described in the literature. However, data derived by spectroscopic detection represent a hard challenge for the researcher, who faces two crucial problems: data dimensionality larger than the observations, and high correlation levels among the variables (multicollinearity).

Usually, in order to solve these problems *(i)* a first data compression or reduction method, such as principal component analysis (PCA) is applied to shrink the number of variables; then, a range of discriminant analysis techniques is used to

solve the classification problem, while *(ii)* in other cases, non-parametric classification approaches are used to classify directly the original data without using any dimensionality reduction methods (Jimenez & Landgrebe 1998, Agrawal et al. 1998, Bühlmann & Van De Geer 2011, Kriegel et al. 2009, Ding & Gentleman 2005).

In this work, the dataset consists of three different varieties of olives (*Moraiolo*, *Dolce di Andria*, and *Nocellara Etnea*) monitored during ripening up to harvest (Bellincontro et al. 2012). Samples containing olives from 162 trees (54 for each variety), and 601 spectral detections (i.e., dimensions/variables) were performed using a portable near infrared acousto-optically tunable filter (NIR-AOTF) device in diffuse reflectance mode from 1100 nm to 2300 nm with an interval of 2. The use of NIRS on olive fruits and related products is already known; applications for the determination of oil and moisture content are now considered routine analyses in comparison with relatively new methodologies, such as nuclear magnetic resonance (NMR), or more traditional analytical determinations (Garcia et al. 1996, Gallardo et al. 2005, León et al. 2004, Cayuela & Camino 2010).

Bellincontro et al. (2012) affirm that the determination of the optimal fruit ripening stage in virgin olive oil production is a critical choice based on the best combination of oil quantity and oil quality. Some of the most important aspects related to virgin olive oil quality are deeply affected by the olive ripening stage. The modification of the phenolic fraction, in particular, has been extensively investigated: the concentration of oleuropein reaches relatively high levels in immature fruit during the growth phase and declines with the physiological development of the fruit. Then, because of the well-known importance of the phenolic fraction for oil stability and the sensory and health properties, it is essential to identify the harvest period that ensures the ripening stage corresponding to the optimal phenolic content. Many approaches have been proposed in recent years for the evaluation of the optimal harvesting period, and Near-infrared spectroscopy (NIRS) can be considered an interesting, alternative technique for the nondestructive measurement of quality parameters in food crops, including fresh fruit and vegetables.

This work is based on the use of partial least squares discriminant Analysis (PLS-DA). The idea is to test some different chemometric applications of NIR spectra, with the aim of predicting qualitative attributes and discriminating cultivar origins using PLS-DA. PLS-DA is a dimensionality reduction technique, a variant of partial least squares regression (PLS-R) that is used when the response variable is categorical. It is a compromise between the usual discriminant analysis and a discriminant analysis on the principal components of the predictor variables. In particular, PLS-DA instead of finding hyperplanes of maximum covariance between the response and independent variables finds a linear regression model by projecting the predicted variables and the observed variables into a new space (Kemsley 1996).

PLS-DA can provide good insight into the causes of discrimination via weights and loadings, which gives it a unique role in exploratory data analysis, for example in metabolomics via visualization of significant variables such as metabolites or spectroscopic peaks (Kemsley 1996, Brereton & Lloyd 2014, Wehrens & Mevik 2007).

However, for comparison purposes, we also analyze the results obtained by other commonly used non-parametric classification models such as K-nearest neighbor (KNN), support vector machine (SVM) (Balabin et al. 2010, Misaki et al. 2010, Tran et al. 2006, Joachims 2005), and some variants of discriminant functions for sparse data as such as diagonal linear discriminant analysis (DLDA), maximum uncertainty linear discriminant analysis (MLDA), and shrunken linear discriminant analysis (SLDA). All the three regularization techniques compute linear discriminant functions (Hastie et al. 1995, Clemmensen et al. 2011, Thomaz et al. 2006, Fisher & Sun 2011, Dudoit et al. 2002, Guo et al. 2006).

The chapter is structured as follows: in Section 1.2 we provide a background on the most commonly used non-parametric statistical methodologies to solve the classification problem of sparse data (i.e., KNN and SVM) and an overview of different classifiers derived from linear discriminant analysis (LDA), in Section 1.3 we focus on the PLS-DA model with a deeper examination of the PLS algorithm, in Section 1.4 we show a comparison of the results obtained by the application of PLS-DA and those obtained by the other common classification methods, and finally in Section 1.5 we provide some suggestions and ideas for future research.

1.2 Background

In this section, we present a brief overview of different classifiers that have been highly successful in handling high dimensional data classification problems, starting with popular methods such as K-nearest neighbor (KNN) and support vector machines (SVM) (Dudoit et al. 2002, Zhang et al. 2006) and variants of discriminant functions for sparse data (Clemmensen et al. 2011). We also examine dimensionality reduction techniques and their integration with some existing algorithms (i.e., partial least squares discriminant analysis (PLS-DA)) (Kemsley 1996, Brereton & Lloyd 2014).

1.2.1 K-Nearest Neighbor (KNN)

The KNN method was first introduced by Fix and Hodges (Fix & Hodges 1989) based on the need to perform discriminant analysis when reliable parametric esti-

mates of probability densities are unknown or difficult to determine. In this method, a distance measure (e.g., Euclidean) is assigned between all points in the data. The data points, K-closest neighbors (where K is the number of neighbors), are then found by analyzing a distance matrix. The K-closest data points are then found and analyzed in order to determine which class label is the most common among the set. Finally, the most common class label is then assigned to the data point being analyzed (Balabin et al. 2010).

The KNN classifier is commonly based on the Euclidean distance between a test sample and the specified training samples. Formally, let $\mathbf{x}_i$ be an input sample with J features $(\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,J})$, and n be the total number of input samples $(i = 1, \ldots, n)$. The Euclidean distance between sample $\mathbf{x}_i$ and $\mathbf{x}_l$ $(l = 1, \ldots, n)$ is defined as

$$d(\mathbf{x}_i, \mathbf{x}_l) = \sqrt{(\mathbf{x}_{i,1} - \mathbf{x}_{l,1})^2 + \cdots + (\mathbf{x}_{i,J} - \mathbf{x}_{l,J})^2}. \tag{1.1}$$

Using the latter characteristic, the KNN classification rule is to assign to a test sample the majority category label of its K nearest training samples. In other words, K is usually chosen to be odd, so as to avoid ties. The $K = 1$ rule is generally called the 1-nearest-neighbor classification rule.

Then, let $\mathbf{x}_i$ be a training sample and $\mathbf{x}_i^*$ be a test sample, and let ω be the true class of a training sample and $\hat{\omega}$ be the predicted class for a test sample $(\omega, \hat{\omega} = \ldots, \Omega)$, where Ω is the total number of classes. During the training process, only the true class ω of each training sample to train the classifier is used, while during testing the class $\hat{\omega}$ of each test sample is predicted. With 1-nearest neighbor rule, the predicted class of test sample $\mathbf{x}_i^*$ is set equal to the true class ω of its nearest neighbor, where $\mathbf{z}_i$ is a nearest neighbor to $\mathbf{x}_i^*$ if the distance

$$d(\mathbf{z}_i, \mathbf{x}_i^*) = \min_j \{d(\mathbf{z}_j, \mathbf{x}_i^*)\}. \tag{1.2}$$

For the K-nearest neighbors rule, the predicted class of test sample $\mathbf{x}_i^*$ is set equal to the most frequent true class among the K nearest training samples.

1.2.2 Support Vector Machine (SVM)

The SVM approach was developed by Vapnik (Suykens & Vandewalle 1999, Cortes & Vapnik 1995). Synthetically, SVM is a linear method in a very high dimensional feature space that is nonlinearly related to the input space. The method maps input vectors to a higher dimensional space where a maximal separating hyperplane is constructed (Joachims 2005). Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data and maximizes the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or

distance between these parallel hyperplanes, the better the generalization error of the classifier will be.

SVM was initially designed for binary classification. To extend SVM to the multi-class scenario, a number of classification models were proposed (Wang & Xue 2014). Formally, given training vectors $\mathbf{x}_i \in \Re^J$, $i = 1, \ldots, n^*$, in two classes, and the label vector $\mathbf{Y} \in \{-1, 1\}^{n^*}$ (where n^* in the size of the training samples), the support vector technique requires the solution of the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w}\in H, b\in\Re, \xi_i\in\Re} & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n^*}\xi_i, \\ \textit{subject to} \quad & y_i(\mathbf{w}^T\varphi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \quad i = 1, \ldots, n^*, \end{aligned} \tag{1.3}$$

where $\mathbf{w} \in \Re^J$ is the weights vector, $C \in \Re_+$ is the regularization constant (i.e., the "cost" parameter), ξ are the data points to classify, and the mapping function φ projects the training data into a suitable feature space H.

For a K-class problem, many methods use a single objective function for training all K-binary SVMs simultaneously and maximize the margins from each class to the remaining ones (Wang & Xue 2014, Weston & Watkins 1998). An example is the formulation proposed by Weston and Watkins (Weston & Watkins 1998). Given a labeled training set represented by $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n^*}, y_{n^*})\}$, where $\mathbf{x}_i \in \Re^J$ and $y_i \in \{1, \ldots, K\}$, this formulation is given as follows:

$$\begin{aligned} \min_{\mathbf{w}_k\in H, b\in\Re^K, \xi\in\Re^{n^*\times K}} & \frac{1}{2}\sum_{k=1}^{K}\mathbf{w}_k^T\mathbf{w}_k + C\sum_{i=1}^{n^*}\sum_{t\neq y_i}\xi_{i,t}, \\ \textit{subject to} \quad & \mathbf{w}_{y_i}^T\varphi(\mathbf{x}_i) + b_{y_i}) \geq \mathbf{w}_t^T\varphi(\mathbf{x}_i) + b_t + 2 - \xi_{i,t}, \\ & \xi_{i,t} \geq 0, \quad i = 1, \ldots, n^*, \quad t \in \{1, \ldots, K\}. \end{aligned} \tag{1.4}$$

The resulting decision function is given in Equation 1.5 (Wang & Xue 2014).

$$\underset{k}{\operatorname{argmax}}\, f_m(\mathbf{x}) = \underset{k}{\operatorname{argmax}}(\mathbf{w}_k^T\varphi(\mathbf{x}_i) + b_k). \tag{1.5}$$

1.2.3 Discriminant Analysis functions

In this section we present a comprehensive overview of different classifiers derived by Linear Discriminant Analysis (LDA), and that have been highly successful in handling high dimensional data classification problems: Diagonal Linear Discriminant Analysis (DLDA), Maximum uncertainty Linear Discriminant Analysis (MLDA), and Shrunken Linear Discriminant Analysis (SLDA). All the three regularization techniques compute Linear Discriminant Functions, by default after a preliminary

variable selection step, based on alternative estimators of a within-groups covariance matrix that leads to reliable allocation rules in problems where the number of selected variables is close to, or larger than, the number of available observations.

The main purpose of discriminant analysis is to assign an unknown subject to one of K classes on the basis of a multivariate observation $x = (x_1, \ldots, x_J)'$, where J is the number of variables. The standard LDA procedure does not assume that the populations of the distinct groups are normally distributed, but it assumes implicitly that the true covariance matrices of each class are equal because the same within-class covariance matrix is used for all the classes considered (Thomaz et al. 2006, Wichern & Johnson 1992). Formally, let $\mathbf{S}_b$ be the between-class covariance matrix defined as

$$\mathbf{S}_b = \sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T, \tag{1.6}$$

and let $\mathbf{S}_w$ be the within-class covariance matrix defined as

$$\mathbf{S}_w = \sum_{k=1}^{K} (n_k - 1)\mathbf{S}_k = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\bar{x}_{k,i} - \bar{x}_k)(\bar{x}_{k,i} - \bar{x}_k)^T, \tag{1.7}$$

where $x_{k,i}$ is the J-dimensional pattern i from the k-th class, n_k is the number of training patterns from the k-th class, and K is the total number of classes (or groups) considered. The vector $\bar{x}_k$ and matrix $\mathbf{S}_k$ are respectively the unbiased sample mean and sample covariance matrix of the k-th class, while the vector $\bar{x}$ is the overall unbiased sample mean given by

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{K} n_k \bar{x}_k = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} x_{k,i}, \tag{1.8}$$

where n is the total number of samples $n = n_1 + \cdots + n_K$.

Then, the main objective of LDA is to find a projection matrix (here defined as $\mathbf{P}_{LDA}$) that maximizes the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix (Fisher's criterion). Formally,

$$\mathbf{P}_{LDA} = \underset{\mathbf{P}}{\operatorname{argmax}} \frac{det\left(\mathbf{P}^T \mathbf{S}_b \mathbf{P}\right)}{det\left(\mathbf{P}^T \mathbf{S}_w \mathbf{P}\right)}. \tag{1.9}$$

It has been shown (Devijver & Kittler 1982) that Equation (1.9) is in fact the solution of the following eigenvector system problem:

$$\mathbf{S}_b \mathbf{P} - \mathbf{S}_w \mathbf{P} \Lambda = 0. \tag{1.10}$$

Note that by multiplying both sides by $\mathbf{S}_w^{-1}$, Equation (1.10) can be rewritten as

$$\begin{aligned} &\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{P} - \mathbf{S}_w^{-1}\mathbf{S}_w\mathbf{P}\Lambda = 0 \\ &\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{P} - \mathbf{P}\Lambda = 0 \\ &(\mathbf{S}_w^{-1}\mathbf{S}_b)\mathbf{P} = \mathbf{P}\Lambda, \end{aligned} \tag{1.11}$$

where $\mathbf{P}$ and Λ are respectively the eigenvector and eigenvalue matrices of the $\mathbf{S}_w^{-1}\mathbf{S}_b$ matrix. These eigenvectors are primarily used for dimensionality reduction, as in principal component analysis (PCA) (Rao 1948).

However, the performance of the standard LDA can be seriously degraded if there are only a limited number of total training observations n compared to the number of dimensions of the feature space J. In this context, in fact the $\mathbf{S}_w$ matrix becomes singular. To solve this problem, Thomaz et al. (2006), Yu & Yang (2001) have developed a direct LDA algorithm (called DLDA) for high dimensional data with application to face recognition that diagonalizes simultaneously the two symmetric matrices $\mathbf{S}_w$ and $\mathbf{S}_b$. The idea of DLDA is to discard the null space of $\mathbf{S}_b$ by diagonalizing $\mathbf{S}_b$ first and then diagonalizing $\mathbf{S}_w$.

The following steps describe the DLDA algorithm for calculating the projection matrix $\mathbf{P}_{DLDA}$:

1. diagonalize $\mathbf{S}_b$, that is, calculate the eigenvector matrix $\mathbf{V}$ such that $\mathbf{V}^T\mathbf{S}_b\mathbf{V} = \Lambda$;
2. let $\mathbf{Y}$ be a sub-matrix with the first m columns of $\mathbf{V}$ corresponding to the $\mathbf{S}_b$ largest eigenvalues, where $m \leq rank(\mathbf{S}_b)$. Calculate the diagonal $m \times m$ sub-matrix of the eigenvalues of Λ as $\mathbf{D}_b = \mathbf{Y}^T\mathbf{S}_b\mathbf{Y}$;
3. let $\mathbf{Z} = \mathbf{Y}\mathbf{D}_b^{-1/2}$ be a whitening transformation of $\mathbf{S}_b$ that reduces its dimensionality from J to m (where $\mathbf{Z}^T\mathbf{S}_b\mathbf{Z} = \mathbf{I}$). Diagonalize $\mathbf{Z}^T\mathbf{S}_w\mathbf{Z}$, that is, compute $\mathbf{U}$ and $\mathbf{D}_w$ such that $\mathbf{U}^T(\mathbf{Z}^T\mathbf{S}_w\mathbf{Z})\mathbf{U} = \mathbf{D}_w$;
4. calculate the projection matrix as $\mathbf{P}_{DLDA} = \mathbf{D}_w^{-1/2}\mathbf{U}^T\mathbf{Z}^T$.

Note that by replacing the between-class covariance matrix $\mathbf{S}_b$ with total covariance matrix $\mathbf{S}_T$ ($\mathbf{S}_T = \mathbf{S}_b + \mathbf{S}_w$), the first two steps of the algorithm become exactly the PCA dimensionality reduction technique (Yu & Yang 2001).

Two other approaches commonly used to avoid both the critical singularity and instability issues of the within-class covariance matrix $\mathbf{S}_w$ are SLDA and the MLDA (Thomaz et al. 2006). Firstly, it is important to note that the within-class covariance matrix $\mathbf{S}_w$ is essentially the standard pooled covariance matrix $\mathbf{S}_p$ multiplied by the scalar $(n - K)$. Then,

$$\mathbf{S}_w = \sum_{k=1}^{K}(n_k - 1)\mathbf{S}_k = (n - K)\mathbf{S}_p. \tag{1.12}$$

From this property, the key idea of some regularization proposals of LDA (Guo et al. 2006, Campbell 1980, Peck & Van Ness 1982) is to replace the pooled covariance matrix $\mathbf{S}_p$ of the within-class covariance matrix $\mathbf{S}_w$ with the following convex combination:

$$\hat{\mathbf{S}}_p(\gamma) = (1-\gamma)\mathbf{S}_p + \gamma\bar{\lambda}\mathbf{I}, \tag{1.13}$$

where $\gamma \in [0,1]$ is the shrinkage parameter, which can be selected to maximize the leave-one-out classification accuracy (Cawley & Talbot 2003), $\mathbf{I}$ is the identity matrix, and $\bar{\lambda} = J^{-1}\sum_{j=1}^{J}\lambda_j$ is the average eigenvalue, which can be written as $J^{-1}trace(\mathbf{S}_p)$. This regularization approach, called SLDA, would have the effect of decreasing the larger eigenvalues and increasing the smaller ones, thereby counteracting the biasing inherent in eigenvalue sample-based estimation (Thomaz et al. 2006, Hastie et al. 1995).

In contrast, in the MLDA method a multiple of the identity matrix determined by selecting the largest dispersions regarding the $\mathbf{S}_p$ average eigenvalue is used. In particular, if we replace the pooled covariance matrix $\mathbf{S}_p$ of the covariance matrix $\mathbf{S}_w$ (shown in Equation (1.12)) with a covariance estimate of the form $\hat{\mathbf{S}}_p(\delta) = \mathbf{S}_p + \delta\mathbf{I}$ (where $\delta \geq 0$ is an identity matrix multiplier), then the eigen-decomposition of a combination of the covariance matrix $\mathbf{S}_p$ and the $J \times J$ identity matrix $\mathbf{I}$ can be written as

$$\begin{aligned}\hat{\mathbf{S}}_p(\delta) &= \mathbf{S}_p + \delta\mathbf{I} \\ &= \sum_{j=1}^{r}\lambda_j\phi_j(\phi_j)^T + \delta\sum_{j=1}^{J}\phi_j(\phi_j)^T \\ &= \sum_{j=1}^{r}(\lambda_j+\delta)\phi_j(\phi_j)^T + \sum_{j=1}^{J}\delta\phi_j(\phi_j)^T,\end{aligned} \tag{1.14}$$

where r is the rank of $\mathbf{S}_p$ (note that $r \leq J$), λ_j is the j-th eigenvalue of $\mathbf{S}_p$, ϕ_j is the j-th corresponding eigenvector, and δ is the identity matrix multiplier previously defined. In fact, in Equation (1.14) the identity matrix is defined as $\mathbf{I} = \sum_{j=1}^{J}\phi_j(\phi_j)^T$. Now, given the convex combination shown in Equation (1.13), the eigen-decomposition can be written as

$$\begin{aligned}\hat{\mathbf{S}}_p(\gamma) &= (1-\gamma)\mathbf{S}_p + \gamma\bar{\lambda}\mathbf{I} \\ &= (1-\gamma)\sum_{j=1}^{r}\lambda_j\phi_j(\phi_j)^T + \gamma\sum_{j=1}^{J}\bar{\lambda}\phi_j(\phi_j)^T.\end{aligned} \tag{1.15}$$

The steps of the MLDA algorithm are shown follows:

1. Find the Φ eigenvectors matrix and Λ eigenvalues matrix of $\mathbf{S}_p$, where $\mathbf{S}_p =$

$(n-K)\mathbf{S}_w$ (from Equation (1.12));
2. Calculate $\mathbf{S}_p$ average eigenvalues as $J^{-1}trace(\mathbf{S}_p)$;
3. Construct a new matrix of eigenvalues based on the following largest dispersion values :

$$\Lambda^* = diag\left[max(\lambda_1, \bar{\lambda}), \ldots, max(\lambda_J, \bar{\lambda})\right];$$

4. Define the revised within-class covariance matrix:

$$\mathbf{S}_w^* = (n-K)\mathbf{S}_p^* = (n-K)(\Phi\Lambda^*\Phi^T).$$

Then, the MLDA approach is based on replacing $\mathbf{S}_w$ with $\mathbf{S}_w^*$ in the Fisher's criterion formula described in Equation (1.9).

1.3 Partial Least Squares Discriminant Analysis (PLS-DA)

Multivariate regression methods like principal component regression (PCR) and partial least squares regression (PLS-R) enjoy large popularity in a wide range of fields and are mostly used in situations where there are many, possibly correlated, predictor variables and relatively few samples, a situation that is common, especially in chemistry, where developments in spectroscopy since the seventies have revolutionized chemical analysis (Wehrens & Mevik 2007, Pérez-Enciso & Tenenhaus 2003). In fact, the origin of PLSR lies in chemistry (Wehrens & Mevik 2007, Martens 2001, Wold 2001).

PCR performs a principal components analysis on the predictors and then fits a linear regression on the chosen reduced dimension. PLS-R, on the other hand, performs the dimensionality reduction by repeatedly regress the response variable on each single predictor: in fact, the response variable participates to the dimensional reduction (Friedman et al. 2001).

Partial least squares discriminant Analysis (PLS-DA) is a variant of PLS-R that can be used when the response variable $\mathbf{Y}$ is categorical. Under certain circumstances, PLS-DA provides the same results as the classical approach of Euclidean distance to centroids (EDC) (Davies & Bouldin 1979) and under other circumstances, the same as that of linear discriminant analysis (LDA) (Izenman 2013). However, in different contexts this technique is specially suited to deal with models with many more predictors than observations and with multicollinearity, two of the main problems encountered when analyzing hyperspectral detection data (Pérez-Enciso & Tenenhaus 2003).

1.3.1 Model and algorithm

PLS-DA is derived from PLS-R, where the response vector $\mathbf{Y}$ assumes discrete values. In the usual multiple linear regression model (MLR) approach we have

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F}, \tag{1.16}$$

where $\mathbf{X}$ is the $n \times J$ data matrix, $\mathbf{B}$ is the $J \times 1$ regression coefficients matrix, $\mathbf{F}$ is the $n \times 1$ error vector, and $\mathbf{Y}$ is the $n \times 1$ response variable vector. In this approach, the least squares solution is given by $\mathbf{B} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y}$.

In many cases, the problem is the singularity of the $\mathbf{X^T X}$ matrix (e.g., when there are multicollinearity problems in the data or the number of predictors is larger than the number of observations). Both PLS-R and PLS-DA solve this problem by decomposing the data matrix $\mathbf{X}$ into P orthogonal scores $\mathbf{T}$ $(n \times P)$ and loadings matrix $\mathbf{\Lambda}$ $(J \times P)$, and the response vector $\mathbf{Y}$ into P orthogonal scores $\mathbf{T}$ $(n \times P)$ and loadings matrix $\mathbf{Q}$ $(1 \times P)$. Then, let $\mathbf{E}$ and $\mathbf{F}$ be the $n \times J$ and $n \times 1$ error matrices associated with the data matrix $\mathbf{X}$ and response vector Y, respectively. There are two fundamental equations in the PLS-DA model:

$$\begin{aligned} \mathbf{X} &= \mathbf{T\Lambda^T} + \mathbf{E} \\ \mathbf{Y} &= \mathbf{TQ^T} + \mathbf{F}. \end{aligned} \tag{1.17}$$

Now, if we define a $J \times P$ weights matrix $\mathbf{W}$, we can write the scores matrix as

$$\mathbf{T} = \mathbf{XW}(\mathbf{\Lambda^T W})^{-1}, \tag{1.18}$$

and by substituting it into the PLS-DA model, we obtain

$$\mathbf{Y} = \mathbf{XW}(\mathbf{\Lambda^T W})^{-1}\mathbf{Q^T} + \mathbf{F}, \tag{1.19}$$

where the regression coefficient vector $\mathbf{B}$ is given by

$$\hat{\mathbf{B}} = \mathbf{W}(\mathbf{\Lambda^T W})^{-1}\mathbf{Q^T}. \tag{1.20}$$

In this way, an unknown sample value of $\mathbf{Y}$ can be predicted by $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$, i.e. $\hat{\mathbf{Y}} = \mathbf{XW}(\mathbf{\Lambda^T W})^{-1}\mathbf{Q^T}$. The PLS-DA algorithm estimates the matrices $\mathbf{W}$, $\mathbf{T}$, $\mathbf{\Lambda}$, and $\mathbf{Q}$ through the following steps (Brereton & Lloyd 2014).

Algorithm 1 Partial Least Squares

1: Fixed P, initialize the residuals matrices $\mathbf{E_0} = \mathbf{X}$ and $\mathbf{F_0} = \mathbf{Y}$;
2: **for** $p = 1$ to P **do**
3: Calculate PLS weights vector
$\mathbf{W_p} = \mathbf{E_0^T F_0}$;
4: Calculate and normalize scores vector
$\mathbf{T_p} = \mathbf{E_0 W_p} (\mathbf{W_p^T E_0^T E_0 W_p})^{-1/2}$;
5: Calculate the $\mathbf{X}$ loadings vector
$\mathbf{\Lambda_p} = \mathbf{E_0^T T_p}$;
6: Calculate $\mathbf{Y}$ loading
$\mathbf{Q_p} = \mathbf{F_0^T T_p}$;
7: Update the $\mathbf{X}$ residuals vector
$\mathbf{E_0} = \mathbf{E_0} - \mathbf{T_p \Lambda_p^T}$;
8: Update the $\mathbf{Y}$ residuals vector
$\mathbf{F_0} = \mathbf{F_0} - \mathbf{T_p Q_p^T}$;
9: **end for**
10: Obtain output matrices $\mathbf{W}$, $\mathbf{T}$, $\mathbf{\Lambda}$, $\mathbf{Q}$.

1.4 Application on real data

In this section we show an application of the method to real data. In particular, we compare the results obtained by partial least squares discriminant analysis (PLS-DA) and the other classification techniques discussed in Section 1.2. Brereton & Lloyd (2014) report good motivations which brings the researchers to compare PLS-DA with other discriminant functions. Moreover, we have also added other two non-parametric approaches in our simulation study (i.e., SVM and KNN) which are reference classification approaches to solve the high dimensionality problems.

1.4.1 Dataset

The dataset consists of 162 drupes of olives harvested in 2010 belonging to three different cultivars (response variable): 54 *Dolce di Andria* (low phenolic concentration), 54 *Moraiolo* (high phenolic concentration), and 54 *Nocellara Etnea* (medium phenolic concentration). Spectral detection is performed using a portable NIR device (diffuse reflectance mode) in the 1100–2300 nm wavelength range, with 2 nm wavelength increments (601 observed variables) (Bellincontro et al. 2012). In Figure 1.1 the NIR spectra in function of wavelength range is presented.

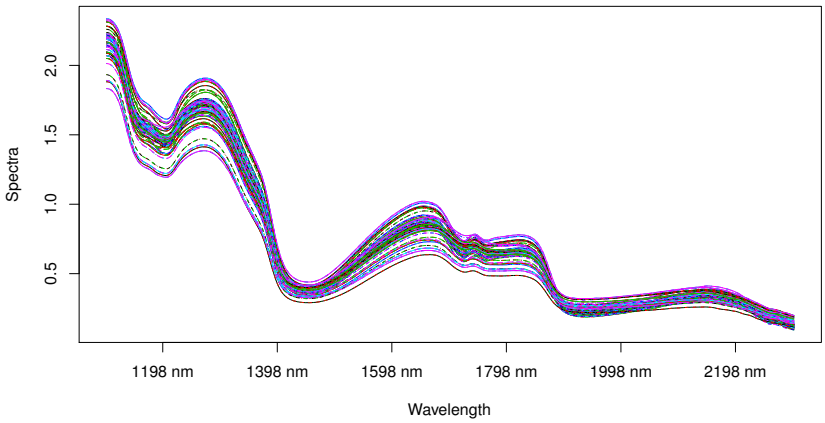


Figure 1.1: Representation of spectral detections performed on the 1100–2300 nm wavelength range

1.4.2 Principal results

In order to evaluate the prediction capability of the model, the entire data set has been randomly divided into a *training set* composed of 111 balanced observations (i.e., about 70% of the entire sample, with each class composed of 37 elements), and a *test set* (drawn from the sample) composed of 51 observations balanced across the three cultivars (i.e., about 30% of the entire sample and each class composed by 17 elements) (Guyon et al. 1998).

The first step of the analysis consists in selecting the optimal number of components P, i.e., the number of latent scores to consider for representing the original variable space. For this purpose, the latent subspace must explain the largest possible proportion of the total variance to guarantee the best model estimation. Table 1.1 shows the proportion of the total variance explained by the first five components identified by PLS-DA. The table shows that the first two components explain about 97% of the total variance, and only the first two latent scores have a very high contribution.

Table 1.1: Cumulative proportion of the total variance explained by the first five components (percent values)

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Exp.Variance	61.152	35.589	0.892	0.982	1.167
Cum. Sum	61.152	96.741	97.633	98.615	99.782

Thus, it seems that the best latent subspace is represented by the plane composed of the first two identified components. However, in order to guarantee the best

model estimate, it is also useful to understand its prediction quality with regard to the different subspace dimensions. In other words, the selection of the optimal number of components must be related to some criterion that ensures the maximum prediction quality of the estimated model. In this work, we propose the maximum reduction of the *misclassification error rate* criterion - applied on the comparison between the real training partition and the predicted training partition - in order to choose the number of components of PLS-R. Figure 1.2 represents the error rate values for different numbers of components (i.e., from 2 to 10 selected components).

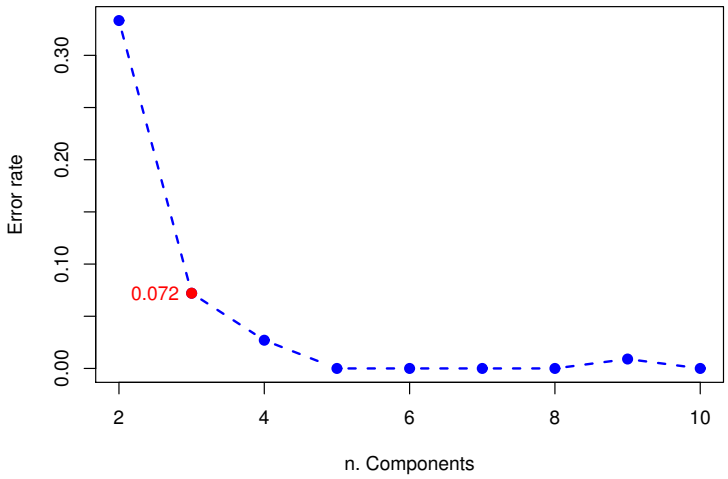


Figure 1.2: Error rate values with respect to different choices of components number

The scree-plot shown in Figure 1.2 suggests $P = 3$ as the optimal number of components, where the minimum value of the misclassification error rate is equal to 0.07. Then, we can select three components to estimate the model.
Figure 1.3 shows the loadings distributions and the squared of the loadings distributions of the three **T**s' latent scores, measured on all the observed variables (i.e., on the 1100–2300 nm wavelength range). By observing the behavior of the loadings, we can say that the wavelengths from about 1100 nm to about 1500 nm have a high negative contribution to the first two components, while they have a positive contribution to the third component; the wavelengths from about 1500 nm to about 1900 nm have a negative contribution to all three components, with the largest contribution to the first component; finally, the wavelengths from about 1900 nm to about 2300 nm have a positive contribution to both the first and the third component, while they have a negative contribution to the second component.

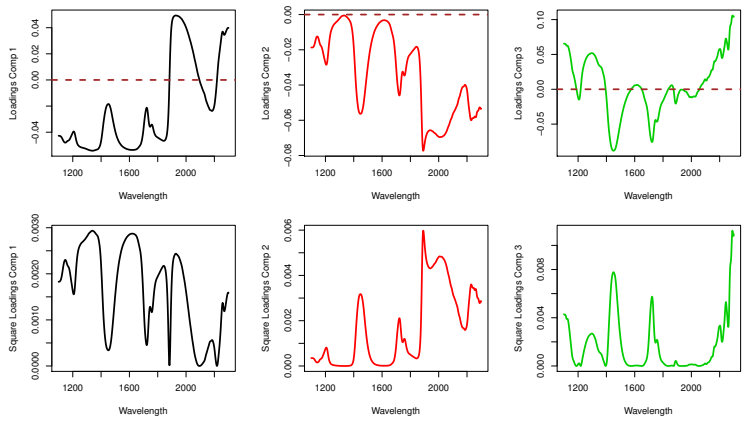


Figure 1.3: The loadings distributions (top) and squared loadings distributions (bottom) of the three latent scores measured on all the observed variables

The partition obtained by PLS-DA on the three latent scores is represented in Figure 1.4.

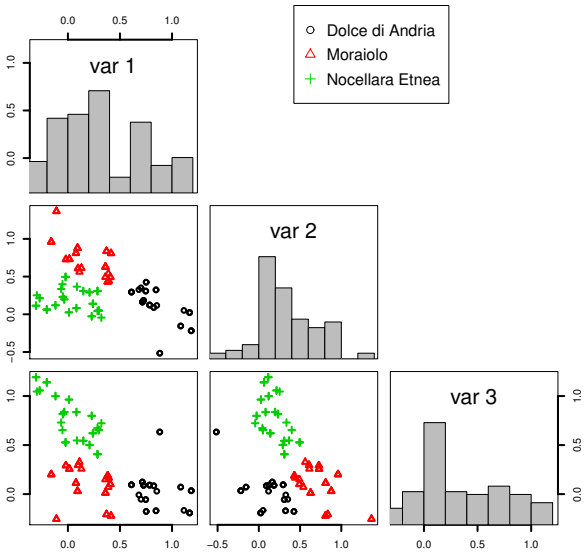


Figure 1.4: Partition obtained by PLS-DA represented on the three estimated latent scores

From the figure, we can see that the partition identified by PLS-DA shows very separated and homogeneous groups maintaining the same features of the original partition of data. Note that PLS-DA partition has a very low misclassification rate, equal to 0.002.

Now, we compare the classification results obtained by the PLS-DA procedure with results obtained by other classifiers, including K-nearest neighbor (KNN), support vector machine (SVM), diagonal linear discriminant analysis (DLDA), maxi-

mum uncertainty linear discriminant analysis (MLDA), and shrunken linear discriminant analysis (SLDA). For the measurement of the model prediction quality, we have used *misclassification rate* (MIS) and the *chi-squared* test (χ^2). The measures have been computed on the comparison between the real data partition and the predicted partition.

Formally, let Table 1.2 be the $K \times K$ confusion matrix where the real data partition (called R) and the predicted partition (called P) have been compared,

$$MIS = 1 - n^{-1} \left[\sum_{r=1}^{R} \sum_{=1}^{C} n_{rc} \right].$$

Table 1.2: An example of a confusion matrix between the real data partition and the predicted partition

		Predicted partition			
		P_1	$\cdots$	P_C	
Real partition	R_1	n_{11}	$\cdots$	n_{1C}	$n_{1\cdot}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	R_R	n_{R1}	$\cdots$	n_{RC}	$n_{R\cdot}$
		$n_{\cdot 1}$	$\cdots$	$n_{\cdot C}$	n

Table 1.3 shows the results for the quality of the model predictions obtained on the training set and the test set.

Table 1.3: Model prediction quality computed on the training set and the test set

	Training set		*Test set*	
	MIS	χ^2	MIS	χ^2
PLS-DA	0.002	153.283	0.008	77.182
KNN	0.027	151.744	0.157	65.294
SVM	0.072	152.688	0.137	69.750
DLDA	0.241	101.599	0.255	46.714
MLDA	0.078	149.577	0.010	72.311
SLDA	0.005	150.456	0.011	75.899

From the results, we can see that PLS-DA has the best performance on both the training set and the test set. This result is confirmed by the representation of the predicted partition on the first three **T**s' latent scores (i.e., on about 97% of

the total data variance) as shown in Figures 1.5 and 1.6 (training set and the test set, respectively). In fact, we can see that, with respect to the other discriminant function, PLS-DA identifies more homogeneous and better-separated classes.

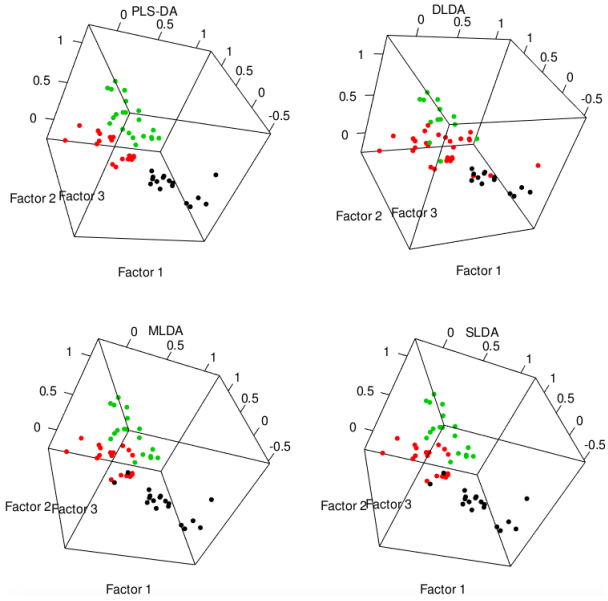


Figure 1.5: Representation of the predicted partition on the three latent scores (training set). The colors black, red, and green represent *Dolce di Andria*, *Moraiolo*, and *Nocellara Etnea*, respectively

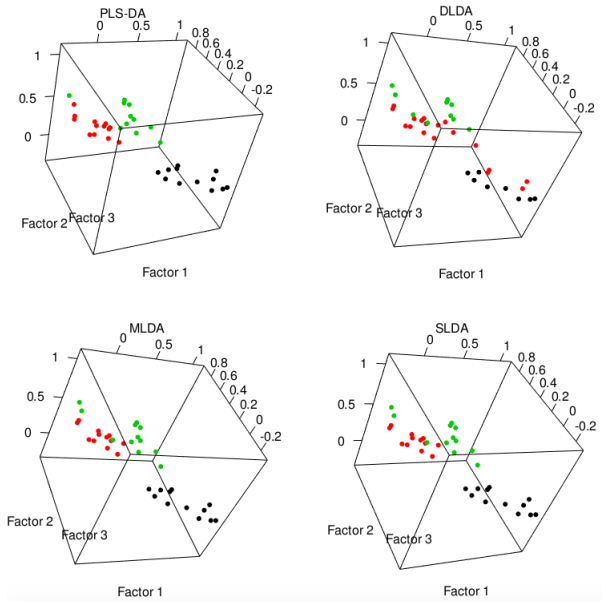


Figure 1.6: Representation of the predicted partition on the three latent scores (test set). The colors black, red, and green represent *Dolce di Andria*, *Moraiolo*, and *Nocellara Etnea*, respectively

From the scatter plot 3D we can see that in terms of classification, an appreciable separation among all observations referring to the three cultivars used has been obtained by a good discrimination among samples of the cultivar *Dolce di Andria* (points in black color) and the two other cultivars, while the separation between samples of the cultivars *Moraiolo*, and *Nocellara Etnea* (points in color red and green, respectively) appears was a bit difficult.

1.5 Concluding remarks

Data acquired via spectroscopic detection represent a hard challenge for researchers, who face two crucial problems: data dimensionality larger than the number of observations, and high correlation levels among the variables. In this work, partial least squares discriminant analysis (PLS-DA) modeling was proposed as a method to classify hyperspectral data. The results obtained on real data show that PLS-DA identifies classes that are more homogeneous and better-separated than other commonly used methods, such as other discriminant functions and some other non-parametric classifiers.

Moreover, we think that PLS-DA is a very important tool in terms of dimensionality reduction, as it can maximize the total variance of data using just a few components (i.e., the $\mathbf{T}$s' latent scores). In fact, the PLS-DA components enable a good graphical representation of the partition, which is not possible with other approaches.

Chapter 2

Multiple Correspondence K-Means: simultaneous vs sequential approach for dimensionality reduction and clustering

2.1 Introduction

In the era of "*big data*", complex phenomena - representing reality in economic, social and many other fields - are frequently described by a large number of statistical units and variables. Researchers who have to deal with this abundance of information are often interested to explore and extract the relevant relationships by detecting a reduced set of prototype units and a reduced set of prototype latent variables, both representing the "*golden knowledge*" mined from the observed data. This dimensionality reduction of units and variables is frequently achieved through the application of two types of methodologies: a discrete classification method, producing hierarchical or non-hierarchical clustering and a dimensionality reduction model that defines the latent factors. The two methodologies, generally are independently applied. In fact, firstly, the factorial method is used to determine a reduced set of latent variables and then the clustering algorithm is computed on the achieved factors. This sequential strategy of analysis has been called tandem analysis (TA) by Arabie & Hubert (1996). With applying first the factorial method it is believed that all the relevant information regarding the relationships of variables is selected

by the factorial method, while, the residual information represents noise that can be discarded. Then, the clustering of units complete the dimensionality reduction of data by producing prototype units generally described by centroids, that is, mean profiles of units belonging to clusters.

However, some authors have noted that TA in some situations cannot be reliable because the factorial models applied first may identify factors that do not necessarily include all the information on the clustering structure of units (Desarbo et al. 1991). In other words the factorial method may filter out some of the relevant information for the subsequent clustering. A solution to this problem is given by a methodology that includes the simultaneous detection of factors and clusters on the observed data. Many alternative methods combining cluster analysis and the search for a reduced set of factors have been proposed, focusing on factorial methods, multidimensional scaling or unfolding analysis and clustering (Heiser 1993, De Soete & Heiser 1993). De Soete & Carroll (1994) proposed an alternative model to the K-means procedure, named reduced K-means (RKM), which appeared to equal projection pursuit clustering (PPC) earlier proposed by Bolton & Krzanowski (2003). RKM simultaneously searches for a clustering of objects, based on the K-means criterion (MacQueen et al. 1967), and a dimensionality reduction of the variables, based on component analysis. However, this approach may fail to recover the clustering of objects when the data have much variance in directions orthogonal to the subspace of the data in which the clusters are allocated (Timmerman et al. 2010). To solve this problem, Vichi & Kiers (2001) proposed the factorial K-means (FKM) model. FKM combines K-means cluster analysis with PCA, then finding the best subspace that best represents the clustering structure in the data. In other words, FKM selects the most relevant variables by producing factors that best identify the clustering structure in the data. Both RKM and FKM proposals are good alternative to TA in the case of numeric variables have been considered.

When categorical (nominal) variables are observed, TA corresponds to apply first multiple correspondence analysis (MCA) and subsequently the K-means clustering on the achieved factors (i.e., latent scores). As far as we know there are no studies that verify if TA has the same problems observed for quantitative variables. Thus, the first aim of this work is to discuss if there are limits of TA in the case of categorical data, the second and most relevant aim of the work is to present a methodology, named multiple correspondence K-means (MCKM), for simultaneous dimensionality reduction and clustering in the case of categorical data. The chapter is structured as follows: in section 2.2 a background on the sequential and simultaneous approaches is provided, showing an example where TA for categorical data fails to identify the correct clusters structure in the data. This is a good motivating example that justifies the use of a simultaneous methodology. In section 2.3 details

on the MCKM model are shown and the alternative least-square (ALS) algorithm is proposed for MCKM parameters estimation. In section 2.4 the main theoretical and applied proprieties of the MCKM are discussed and finally, in section 2.5 an application on a real benchmark data set is given to show the MCKM performance.

2.2 Statistics background and motivating example

Let $\mathbf{X} = [x_{ij}]$ be a $N \times J$ data matrix corresponding to N units (objects) on which J categorical (nominal) variables have been observed, tandem analysis (TA) (Arabie & Hubert 1996, Desarbo et al. 1991) is the statistical multivariate procedure that uses two methodologies: *(i)* a dimensionality reduction (factorial) method for finding a set of P factors (generally, $P < J$) that better reconstructing the J observed variables (e.g., principal component analysis (PCA) or factor analysis (FA)); and *(ii)* a clustering method that partitions the N multivariate objects into K homogeneous and isolated clusters (e.g., K-means (KM), or gaussian mixture models (GMM)).

In TA first the factorial method is applied to define a matrix of component scores; then, the clustering method is applied, sequentially, on the component score matrix to identify the clusters structure. The first methodology detects the maximal part of the total variance by using a reduced set of P components; while the second method maximizes the between variance of the total variance explained in the first analysis. Thus, the variance explained by the factorial method could not be all the between variance of the original variables necessary for the successive clustering methodology. Actually, it may happen that some noise masking the successive clustering could have been included in the P components. Vichi & Kiers (2001) show an instructive example where a data set formed by variables with a clusters structure and other variables without clusters structure (noise), but having high variance, has been considered. When TA is applied on this typology of data, PCA generally explains also part of the noise data (i.e., where the maximum variance there is). These last tend to mask the observed clusters structure, and as a consequence, several units are misclassified.

If the J variables considered in the matrix $\mathbf{X}$ are categorical, then TA corresponds, usually, to apply multiple correspondence analysis (MCA) and K-means (KM), where this last is sequentially applied on the factors identified by MCA. The researcher may ask if TA for the categorical variables has the same limits discussed for the quantitative case. Before considering this, let us first formalize TA in the categorical data case.

The MCA model can be written as

$$J^{1/2}\mathbf{JBL}^{1/2} = \mathbf{YA}^T + \mathbf{E}_{MCA} \ , \tag{2.1}$$

where $\mathbf{Y} = J^{1/2}\mathbf{JBL}^{1/2}\mathbf{A}$ is the $N \times P$ score matrix of MCA; $\mathbf{A}$ is the $J \times P$ column-wise orthonormal loadings matrix (i.e., $\mathbf{A}^T\mathbf{A} = \mathbf{I}_P$); $J^{1/2}\mathbf{JBL}^{1/2} = \mathbf{X}$ is the centered data matrix corresponding to the J qualitative variables, with the binary block matrix $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_j]$ composed by J indicator binary matrices $\mathbf{B}_j$ with elements $b_{ijm} = 1$ if the i^{th} has assumed category m for variable j, $b_{ijm} = 0$ otherwise; $\mathbf{L} = diag(\mathbf{B}^T\mathbf{1}_N)$; $\mathbf{J} = \mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N^T$ is the idempotent centering matrix with $\mathbf{1}_N$ the N-dimensional vector of unitary elements.
The KM model applied on the MCA scores matrix $\hat{\mathbf{Y}} = J^{1/2}\mathbf{JBL}^{1/2}\hat{\mathbf{A}}$ can be written as

$$\hat{\mathbf{Y}} = \mathbf{U}\bar{\mathbf{Y}} + \mathbf{E}_{KM} \ , \tag{2.2}$$

where $\mathbf{U}$ is the $N \times K$ binary and row stochastic memberships matrix, i.e., $u_{ik} \in \{0, 1\}$ with $i = 1, \ldots, N$ and $k = 1, \ldots, K$ and $\mathbf{U1}_K = \mathbf{1}_N$, identifying a partition of objects and $\bar{\mathbf{Y}}$ is the $K \times P$ corresponding centroid matrix in the P-dimensional space. Note that $\mathbf{Y} = \mathbf{XA}$, while $\bar{\mathbf{Y}} = \bar{\mathbf{X}}\mathbf{A}$. Finally, $\mathbf{E}_{MCA}$ and $\mathbf{E}_{KM}$ are the $N \times J$ error matrices of MCA and KM, respectively.
The least-squares (LS) estimation procedure of the model shown in Equation (2.1) corresponds to minimize the loss function

$$\begin{cases} ||J^{1/2}\mathbf{JBL}^{1/2} - \mathbf{YA}^T||^2 \underset{\mathbf{A}}{\rightarrow} min \\ \mathbf{A}^T\mathbf{A} = \mathbf{I}_P \\ \mathbf{Y} = J^{1/2}\mathbf{JBL}^{1/2} \end{cases} , \tag{2.3}$$

while LS estimation of model shown in Equation (2.2) relates to minimize the loss function

$$\begin{cases} ||\hat{\mathbf{Y}} - \mathbf{U}\bar{\mathbf{Y}}||^2 \underset{\mathbf{U},\bar{\mathbf{Y}}}{\rightarrow} min \\ \mathbf{U} \in \{0, 1\} \\ \mathbf{U1}_K = \mathbf{1}_N \end{cases} , \tag{2.4}$$

Thus, given the LS estimates $\hat{\mathbf{A}}$, $\hat{\mathbf{U}}$, $\hat{\bar{\mathbf{Y}}}$ of MCA and KM, and considering $\mathbf{Y} = J^{1/2}\mathbf{JBL}^{1/2}\hat{\mathbf{A}}$, the TA procedure has an overall objective function equal to the sum (or mean) of the two objective functions of MCA and KM; formally,

$$f(\hat{\mathbf{Y}}, \hat{\mathbf{A}}, \hat{\mathbf{U}}, \hat{\bar{\mathbf{Y}}}) = \frac{1}{2}\left(||J^{1/2}\mathbf{JBL}^{1/2} - \hat{\mathbf{Y}}\hat{\mathbf{A}}^T||^2 + ||\hat{\mathbf{Y}} - \hat{\mathbf{U}}\hat{\bar{\mathbf{Y}}}||^2\right) \ . \tag{2.5}$$

Therefore, TA is the procedure that optimizes sequentially the two objective functions of MCA and KM, which loss can be summarized by the quantity shown in

Equation (2.5). However, we now show with an example that this sequential estimation has some limits similar to those highlighted in the quantitative variables case. In Figure 2.1, the heat-map of the data matrix of 90 units according to 6 qualitative categorical variables, each one with 9 categories, is shown.

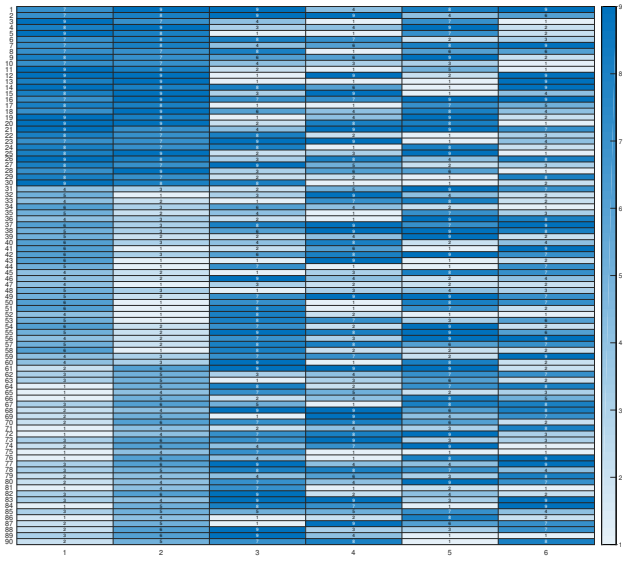


Figure 2.1: Heat-map of the 90×6 categorical variables with 9 categories for each variable

This is a synthetic data set composed by multinomial distributions. The first two variables are a mixture of three multinomial distributions with values from 1 to 3, from 4 to 6 and from 7 to 9, respectively, thus defining three clusters of units, each one with equal size (30 units). The other four variables are multinomial distributions with values from 1 to 9 with equal probabilities, thus these do not define any clusters structure of data. We suppose that this is an example of a simulated data set of 90 customers who have expressed their preferences on 6 products on the basis of a Likert scale from 1 (like extremely) to 9 (dislike extremely), passing through 5 (neither like nor dislike).

The heat-map in Figure 2.1 is a graphical representation of data where the individual values contained in the matrix are represented as different levels of blur from white (value 1) to blue (value 9) (1 like extremely, 2 like very much, 3 like moderately, 4 like slightly, 5 neither like nor dislike, 6 dislike slightly, 7 dislike moderately, 8 dislike very much, 9 dislike extremely). By examining the columns of the heat-map (corresponding to products) it can be confirmed that the first two (products A, B) have a well-defined clusters structure. In fact, the first 30 customers dislike (moderately, very much and extremely) the two products having chosen attributes from 7 to 9, for both products. Customers from 31 to 60 having values from 4 to 6 and from 1 to 3, for the first and second column, respectively, are almost neutral on

the first product (like slightly, nether like nor dislike, dislike slightly), but they like the second product (extremely, very much or moderately). Finally, customers from 61 to 90 have values from 1 to 3 and from 4 to 6 in the first and second column, respectively, thus, they like the first product and are substantially neutral for the second. For the other four products (C, D, E, F) the 90 customers do not show a systematic clusters pattern with values that range randomly with equal probability from 1 to 9. Therefore, the 90 customers have two patterns of preferences: "clustered" for products A, B and "random" for products C, D, E and F.

On the 90×6 data matrix defined, TA was applied by computing first the MCA model and subsequently, by running the K-means algorithm on the first two components identified by MCA. Figure 2.2, shows the biplot of categories of the 6 variables named A, B, C, D, E, F and followed by a number between 1 and 9 to distinguish categories. The total loss of the function shown in Equation (2.5) is 7.39.

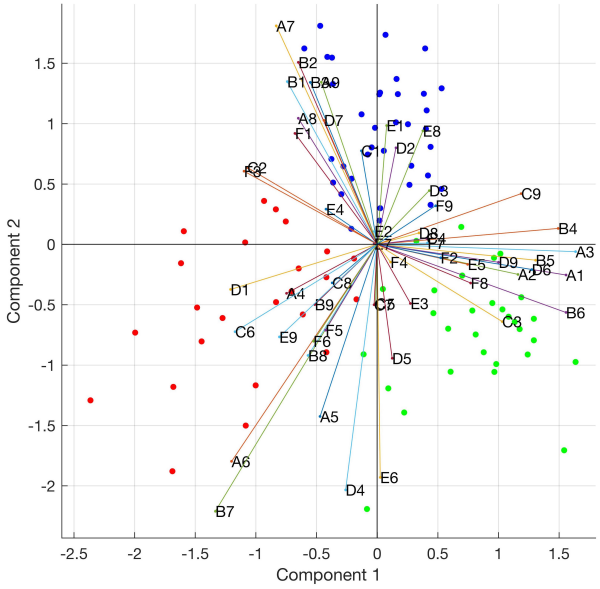


Figure 2.2: Biplot of the 90×6 qualitative variables (A, B, C, D, E, F) with categories from 1 to 9. The three generated clusters are represented by three different colors

It can be clearly seen from the biplot that the most relevant categories are those of the two variables A and B together with other categories e.g., F7, C7, E9, D1 from variables F, C, E and D. Thus, the clustered and the random patterns of the customers are assorted and not clearly distinguishable in the biplot. Furthermore, TA tends to mask the three clusters of costumers, each one originally formed by 30 customers, as shown in Table 2.1. In fact, the points classified in the three groups are 40, 28 and 22, respectively. Thus, 11 customers (12%) are misclassified (3 from the second cluster and 8 from the last cluster). The adjusted Rand index (ARI) discussed in Rand (1971) and Hubert & Arabie (1985) computed on the comparison

between the generated three clusters and the three clusters obtained by K-means is equal to 0.6579.

Table 2.1: Contingency table between K-Means groups and simulated groups

		K-means groups			
		Group 1	Group 2	Group 3	Total
Simulated groups	Group 1	30	0	0	30
	Group 2	3	27	0	30
	Group 3	7	1	22	30
	Total	40	28	22	90

Then, TA describes imprecisely the three clusters and defines components which do not clearly distinguish the two different preference patterns: the clustered for products A, B and the random for the products C, D, E, F.

2.3 Multiple Correspondence *K*-Means: model and algorithm

2.3.1 Model

Hwang et al. (2006) propose a convex combination of the homogeneity both for the criterion MCA and for the criterion K-means; in this work let us use a different approach by specifying a model for the data, replacing Equation (2.2) into Equation (2.1). Thus, it follows that

$$J^{1/2}\mathbf{JBL}^{1/2} = (\mathbf{U}\bar{\mathbf{Y}} + \mathbf{E}_{KM})\mathbf{A}^T + \mathbf{E}_{MCA}\ , \tag{2.6}$$

and rewriting the error term $\mathbf{E}_{MCKM} = \mathbf{E}_{KM}\mathbf{A}^T + \mathbf{E}_{MCA}$, the resulting equation is here named multiple correspondence K-means (MCKM) model:

$$J^{1/2}\mathbf{JBL}^{1/2} = (\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^T + \mathbf{E}_{MCKM})\ . \tag{2.7}$$

MCKM model identifies, simultaneously, the best partition of the N objects described by the best orthogonal linear combination of variables according to a single objective function. The coordinates of the projections onto the basis are given by the components y_{ip} collected in the matrix $\mathbf{Y} = \mathbf{XA}$. Within this subspace, hence, with these components, a partition of objects is sought such that the objects are "closest" to the centroids of the clusters (Vichi & Kiers 2001). When $\mathbf{X} = J^{1/2}\mathbf{JBL}^{1/2}$ is actually a quantitative data matrix, the least-squares (LS) estimation of the model

shown in Equation (2.7) is equal to the reduced K-means (RKM) model proposed by De Soete & Carroll (1994). Additionally, when Equation (2.7) is post-multiplied both sides by $\mathbf{A}$, the RKM model is transformed into the factorial K-means (FKM) model, proposed by Vichi & Kiers (2001). Both models have been formalized for numeric data.
The LS estimation of MCKM corresponds to minimize the objective function

$$\begin{cases} ||J^{1/2}\mathbf{JBL}^{1/2} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^T||^2 \underset{\mathbf{A},\mathbf{U},\bar{\mathbf{Y}}}{\rightarrow} min \\ \mathbf{A}^T\mathbf{A} = \mathbf{I}_P \\ \mathbf{U} \in \{0,1\} \\ \mathbf{U}\mathbf{1}_K = \mathbf{1}_N \end{cases} . \tag{2.8}$$

2.3.2 Alternating least-squares algorithm

The quadratic constrained problem of minimizing Equation (2.8) can been solved by an alternative least-squares (ALS) algorithm, which is structured on three fundamental steps, as follows:

Step 0: Firstly, initial values are chosen for $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$; in particular, initial values for $\mathbf{A}$ and $\mathbf{U}$ can be chosen randomly satisfying the constraints shown in Equation (2.8), while initial values for $\bar{\mathbf{Y}}$ are then given at once by $(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{Y}$.

Step 1: Minimize $F([u_{ik}]) = ||J^{1/2}\mathbf{JBL}^{1/2} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^T||^2$ with respect to $\mathbf{U}$, given the current values of $\mathbf{A}$ and $\bar{\mathbf{Y}}$. The problem is solved for the rows of $\mathbf{U}$ independently by taking $u_{ik} = 1$ if $F([u_{ik}]) = min\{F([u_{iv}]) : v = 1, \ldots, P; (v \neq k)\}$; $u_{ik} = 0$, otherwise.

Step 2: Given $\mathbf{U}$, update $\mathbf{A}$ and implicitly $\bar{\mathbf{Y}}$ by minimizing the loss function in Equation (2.8). The problem is solved by taking the first p eigenvectors of $\mathbf{X}^T(\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U})^T\mathbf{X}$ (Vichi & Kiers 2001).

Step 3: Compute the objective function in Equation (2.8) for the current values of $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$. When the updates of $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$ have decreased the function value, repeat the step 1 and 2; otherwise, the process has converged.

ALS algorithm monotonically decreases the loss function and, because the constraints on $\mathbf{U}$, the method can be expected to be rather sensitive to *local minima*. For this reasons, it is recommended the use of many randomly started runs to find the best solution. In some test, it has been valued that, for a good solution (a good

local optimum value), the use of 500 random starts usually suffices. Note that the algorithm is very fast.

2.4 Theoretical and applied properties

2.4.1 Theoretical Property

Proof 1: The least-squares solution of MCKM obtained by solving the quadratic problem shown in Equation (2.8) subject to constraints $\mathbf{A}^T\mathbf{A} = \mathbf{I}_P$, $\mathbf{U} \in \{0,1\}$, and $\mathbf{U}\mathbf{1}_K = \mathbf{1}_N$ is equivalent to the minimization of the objective function shown in Equation (2.5) used to give an overall estimation of the loss produced by tandem analysis results. In other words, the following equality can be proved

$$2f(\hat{\mathbf{Y}}, \hat{\mathbf{A}}, \hat{\mathbf{U}}, \hat{\bar{\mathbf{Y}}}) = ||J^{1/2}\mathbf{JBL}^{1/2} - \hat{\mathbf{Y}}\hat{\mathbf{A}}^T||^2 + ||\hat{\mathbf{Y}} - \hat{\mathbf{U}}\hat{\bar{\mathbf{Y}}}||^2 = ||\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^T||^2 \ , \quad (2.9)$$

where $\mathbf{X} = J^{1/2}\mathbf{JBL}^{1/2}$.
In fact, after some algebra the objective function of multiple correspondence K-means can be written as

$$||\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^T||^2 = ||\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^T||^2 = tr(\mathbf{X}^T\mathbf{X}) - tr(\mathbf{X}^T\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^T) \ . \quad (2.10)$$

Thus, it is necessary to prove that the objective function of TA is equal to Equation (2.10).

$$\begin{aligned}
&||\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}^T||^2 + ||\mathbf{X}\mathbf{A} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A}||^2 = \\
&tr\{(\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}^T)^T(\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}^T)\} + tr\{(\mathbf{X}\mathbf{A} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A})^T(\mathbf{X}\mathbf{A} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A})\} = \\
&tr(\mathbf{X}^T\mathbf{X}) - tr(\mathbf{X}^T\mathbf{X}\mathbf{A}\mathbf{A}^T) - tr(\mathbf{A}\mathbf{A}^T\mathbf{X}^T\mathbf{X}) + tr(\mathbf{A}\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A}\mathbf{A}^T) + \\
&+ tr(\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A}) - tr(\mathbf{A}^T\mathbf{X}^T\mathbf{U}\bar{\mathbf{X}}\mathbf{A}) - tr(\mathbf{A}^T\bar{\mathbf{X}}^T\mathbf{U}^T\mathbf{X}\mathbf{A}) + tr(\mathbf{A}^T\bar{\mathbf{X}}^T\mathbf{U}^T\mathbf{U}\bar{\mathbf{X}}\mathbf{A}) \ .
\end{aligned} \quad (2.11)$$

Now, knowing that $\mathbf{U}\bar{\mathbf{X}} = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{X} = \mathbf{P}_\mathbf{U}\mathbf{X}$, where $\mathbf{P}_\mathbf{U}$ is the idempotent projector of matrix $\mathbf{U}$, Equation (2.11) can be written as

$$\begin{aligned}
&tr(\mathbf{X}^T\mathbf{X}) - tr(\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A}) - tr(\mathbf{A}\mathbf{A}^T\mathbf{X}^T\mathbf{X}) + tr(\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A}) + \\
&+ tr(\mathbf{A}\mathbf{A}^T\mathbf{X}^T\mathbf{X}) - tr(\mathbf{A}^T\mathbf{X}^T\mathbf{P}_\mathbf{U}\mathbf{X}\mathbf{A}) - tr(\mathbf{A}^T\mathbf{X}^T\mathbf{P}_\mathbf{U}\mathbf{X}\mathbf{A}) + tr(\mathbf{A}^T\mathbf{X}^T\mathbf{P}_\mathbf{U}\mathbf{P}_\mathbf{U}\mathbf{X}\mathbf{A}) = \\
&= tr(\mathbf{X}^T\mathbf{X}) - tr(\mathbf{A}^T\mathbf{X}^T\mathbf{P}_\mathbf{U}\mathbf{X}\mathbf{A}) - tr(\mathbf{A}^T\mathbf{X}^T\mathbf{P}_\mathbf{U}\mathbf{X}\mathbf{A}) + tr(\mathbf{A}^T\mathbf{X}^T\mathbf{P}_\mathbf{U}\mathbf{X}\mathbf{A}) = \\
&= tr(\mathbf{X}^T\mathbf{X}) - tr(\mathbf{X}^T\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}^T) \ ,
\end{aligned} \quad (2.12)$$

which complete the prof.

2.4.2 Applied Property

Let us apply MCKM on the 90×6 data set used in section 2.2 to show the limits of TA in the case categorical data are considered. In this case, the loss value obtained minimizing the function shown in Equation (2.8) is equal to 7.23, better than the loss obtained by TA, with an improvement of 2%. Even if the improvement seems small this time the biplot of MCKM in Figure 2.3 shows a very clear synthesis of the data.

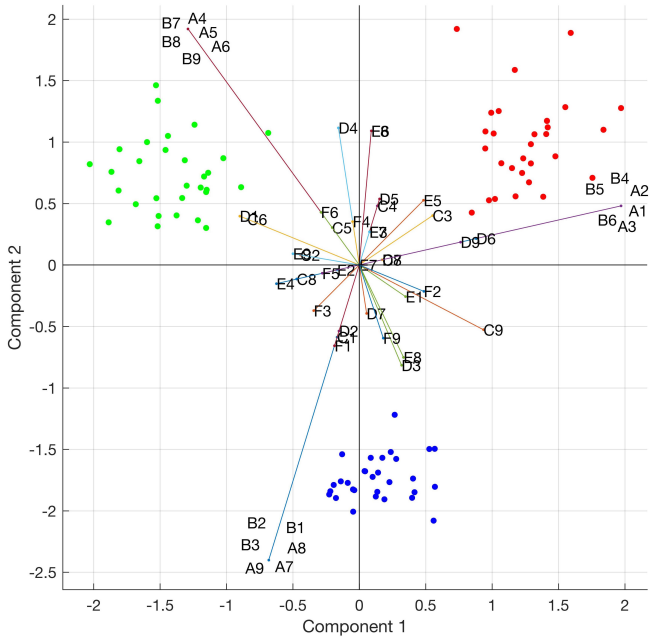


Figure 2.3: Biplot of the multiple correspondence K-means . It can be clearly observed that the three cluster are homogeneous and well-separated

Categories of products A and B are well-distinguished from categories of products C, D, E, F and therefore the clustered and random patterns of preferences of customers are clearly differentiated. Furthermore, the clusters structure of the customers is well represented in the biplot. In fact, the three clusters are composed each one by 30 customers, as expected, and they are more homogeneous and well-separated with respect to the clusters in the biplot of TA (Figure 2.2).
In particular, the red cluster is composed by customers who like products A and are neutral on the product B (the first 30 rows, in the data set). The blue cluster is composed by customers who like the second product B and dislike the first product A (the second 30 rows of the data set). Finally, the green cluster of customers is composed by people that dislike the product B and are neutral of on product A (the third and last 30 rows of the data set). Then, this time no misclassifications are observed for the clusters (see Table 2.2) and the two different patterns of products

are differently represented in the plot as expected.

Table 2.2: Contingency table between MCKM groups and simulated groups

		K-Means			
		Group 1	Group 2	Group 3	Total
Simulated groups	Group 1	30	0	0	30
	Group 2	0	30	0	30
	Group 3	0	0	30	30
	Total	30	30	30	90

2.5 Application on South Korean underwear manufacturer data set

The empirical data presented in this section, is part of a large survey conducted by a South Korean underwear manufacturer in 1997 (Hwang et al. 2006), where 664 South Korean consumers were asked to provide responses for three multiple-choice items. In Table 2.3 the frequency distributions of the three categorical variables are shown.

Table 2.3: Frequency distributions of the South Korean underwear manufacturer data

BRAND (A)		ATTRIBUTES (B)		AGE (C)	
A01. BYC	201	B01. Comfortable	398	C01. 10 - 29	239
A02. TRY	131	B02. Smooth	65	C02. 30 - 49	242
A03. VICMAN	30	B03. Superior fabrics	29	C03. 50 and over	183
A04. James Dean	72	B04. Reasonable price	33		
A05. Michiko-London	11	B05. Fashionable design	67		
A06. Benetton	13	B06. Favorable advertisements	7		
A07. Bodyguard	166	B07. Trendy color	15		
A08. Calvin Klein	40	B08. Good design	4		
		B09. Various colors	4		
		B10. Elastic	11		
		B11. Store is near	3		
		B12. Excellent fit	20		
		B13. Design quality	6		
		B14. Youth appeal	1		
		B15. Various sizes	1		

In particular, the first item asked which of eight brands of underwear the consumer

most prefers (A): (A01) BYC, (A02) TRY, (A03) VICMAN, (A04) James Dean, (A05) Michiko-London, (A06) Benetton, (A07) Bodyguard, and (A08) Calvin Klein; then, both domestic (A01, A02, A03, A04, and A07) and international (A05, A06, and A08) brands were included. The second item asked the attribute of underwear most sought by the consumers (B): (B01) comfortable, (B02) smooth, (B03) superior fabrics, (B04) reasonable price, (B05) fashionable design, (B06) favourable advertisements, (B07) trendy colour, (B08) good design, (B09) various colors, (B10) elastic, (B11) store is near, (B12) excellent fit, (B13) design quality, (B14) youth appeal, and (B15) various sizes. The last item asked the age class of each consumer (C): (C01) 10-29, (C02) 30-49, and (C03) 50 and over.

The analysis starts with the application of multiple correspondence analysis and, subsequently, the application of K-means on the computed scores (i.e., we apply tandem analysis). Hwang et al. (2006), suggested to apply MCA by fixing the number of components equal to 2 since sizes of the adjusted inertias appeared to decrease slowly after the first two. The results obtained by the MCA are shown in the Table 2.4.

Table 2.4: Results of the MCA model applied on the South Korean underwear manufacturer data

Sing. Value	Inertia	Chi-square	Inertia (%)	Cum.Inertia (%)
0.726	0.527	1048.930	6.870	6.870
0.644	0.414	824.878	5.400	12.270
Total	0.941	1873.808	12.270	

p-value = 0, Degrees of freedom = 196

From Table 2.4, it is worthy to note that the explained variance of the two computed factors is equal to 12.27% of the total inertia. Note that Greenacre (1984) recommends to adjust the inertias greater than $1/J$ using the formula proposed by Benzécri (1979). In the Table 2.5 the computed loadings among the two components and each category of the data are shown.

From the table, it easy to note that the categories with bigger contributions on the first component are: the first two brands of underwear (A01 and A02) and the seventh brand (A07); the fifth attribute (B05); the first and third class of the age (C01 and C03). Whereas, the categories with bigger contribution on the second component are: the third, fourth and fifth brand (A03, A04 and A05); the third, fourth, tenth and thirteenth attribute (B03, B04, B10 and B13); second and third class of the age (C01 and C03). Then, the two component scores represent a very

high number of the categories. However, the variables brands (A) and age (C) are more represented than attributes (B).

Table 2.5: Loading matrix of the MCA model applied on the South Korean underwear manufacturer data

Component 1			Component 2		
Brand	Attributes	Age	Brand	Attributes	Age
-0.250	-0.133	0.467	0.177	-0.152	0.102
-0.302	-0.065	-0.163	0.090	0.184	-0.374
-0.134	-0.008	-0.346	-0.363	0.285	0.312
0.135	-0.047	-	-0.291	0.234	-
0.161	0.373	-	0.311	0.064	-
0.181	-0.046	-	-0.031	-0.036	-
0.334	0.108	-	0.038	0.030	-
0.175	0.123	-	-0.077	0.017	-
-	-0.097	-	-	0.027	-
-	-0.082	-	-	-0.278	-
-	-0.020	-	-	0.162	-
-	-0.002	-	-	-0.164	-
-	0.152	-	-	-0.231	-
-	0.099	-	-	0.049	-
-	-0.067	-	-	-0.073	-

Subsequently, according to TA approach, the K-means model on the two component scores has been applied (Figure 2.4).

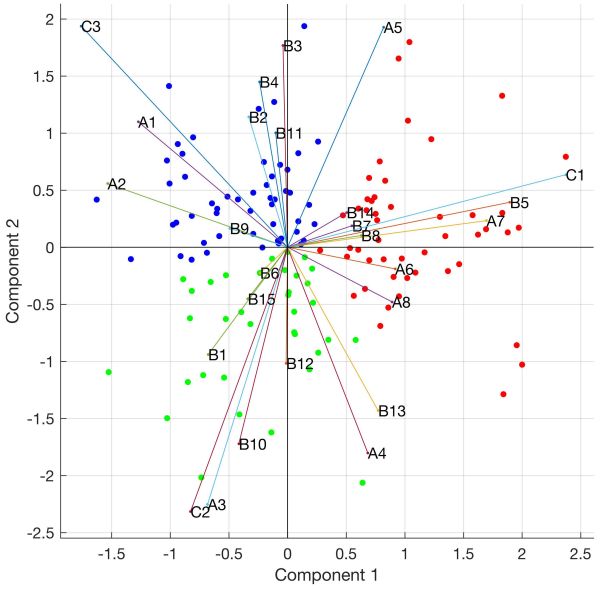


Figure 2.4: Biplot of the sequential approach applied on South Korean underwear manufacturer data

The fixed number of groups is $K = 3$ as suggested by Hwang et al. (2006). The plot in Figure 2.4 shows the projection of the single category on the bi-dimensional factorial plane and the distributions of the computed scores. We can note that the three defined groups are underlined with different colors. The biplot shows that the groups are not well separated and they are characterized by a high inside heterogeneity. In fact, it is very hard to understand the preferences of the consumers that belong to the three groups.

Different results have been obtained with the multiple correspondence K-means approach. Fixing the same number of components and groups, the explained variance of the two components are around to 20%. The component loadings of the MCKM are represented in the Table 2.6.

Table 2.6: Loading matrix of the MCKM model applied on the South Korean underwear manufacturer data

Component 1			Component 2		
Brand	Attributes	Age	Brand	Attributes	Age
0.429	0.029	-0.252	0.159	0.040	-0.057
0.346	0.028	0.062	0.128	0.068	-0.018
0.158	0.034	0.216	0.046	-0.045	0.086
-0.123	0.025	-	-0.609	0.007	-
-0.048	-0.161	-	-0.238	-0.074	-
-0.052	0.031	-	-0.259	0.007	-
-0.694	-0.016	-	0.449	-0.018	-
-0.092	-0.046	-	-0.454	0.005	-
-	0.061	-	-	0.022	-
-	0.011	-	-	0.034	-
-	0.052	-	-	0.019	-
-	0.036	-	-	-0.093	-
-	-0.052	-	-	-0.132	-
-	-0.054	-	-	0.035	-
-	0.030	-	-	0.011	-

In the MCKM results the categories with bigger contributions on the first component are: the first two brands of underwear (A01 and A02) and the seventh brand (A07); the first and the third class of the age (C01 and C03). The categories with bigger contribution on the second component are the fourth, fifth, sixth, seventh and eighth brand (A04, A05, A06, A07 and A8) only. Then, unlike TA, in the MCKM model the variable attributes (B) do not give a relevant contribution. In Figure 2.5 is shown the biplot where are represented the component scores and the three defined groups. From the plot we can note that the groups are well separated and homogeneous. In fact, it easy to note that the green group (166 observations) are the consumers that prefer the seventh brand (A07); the blue group (361 obser-

vations) are the consumers that prefer the first three brands (A01, A02 and A03) and they have mainly an age of 50 years and over (C03); finally the red groups (137 observations) are the consumers that prefer the fourth, fifth, sixth and eight brand (A04, A05, A06, and A08).

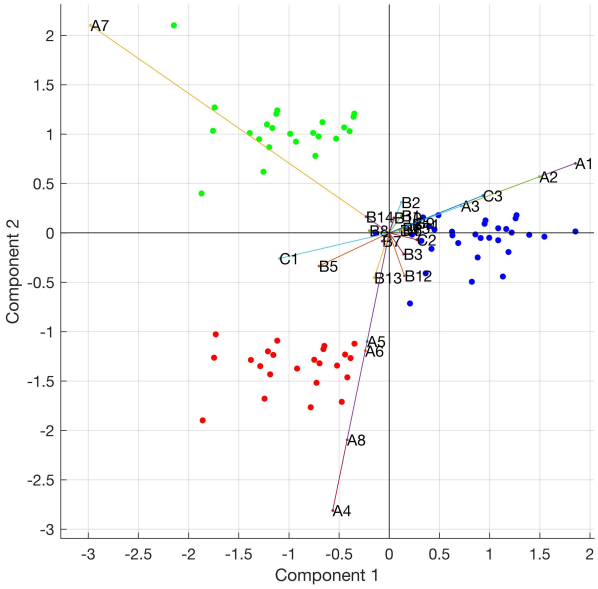


Figure 2.5: Biplot of the simultaneous approach applied on South Korean underwear manufacturer data

It is possible to verify these results observing the frequency distributions of the three categorical variables shown in Table 2.3.

2.6 Concluding remarks

Tandem Analysis (TA) is a well-known sequential procedure for clustering and dimensionality reduction. It is frequently used in applications for quantitative data, although is has several limitations. In particular, it can fail to find the correct clusters structure with a reduced set of factors (Vichi & Kiers 2001). TA is also frequently used when categorical variables are considered. It corresponds to apply multiple correspondence analysis (MCA) on the original data and subsequently to apply K-means model on the component scores matrix obtained by MCA to cluster the statistical units.

In this work it was proved that also this TA has serious problems to correctly classify units and synthesize the relationships of the observed categorical variables. Thus, a model called multiple correspondence K-means (MCKM) was proposed and estimated via least-squares (LS) by using an alternating least squares (ALS)

algorithm. It has been proved that the LS estimation of MCKM corresponds to optimize the loss function TA which is only imprecisely estimated by the sequential application of MCA and K-means.

Chapter 3

Structural Equation Modeling and simultaneous clustering through the Partial Least Squares algorithm

3.1 Introduction

In the last years, structural equation modeling (SEM) has become one of the reference statistical methodologies in the analysis of the statistical relationships between observable (manifest) and non-observable (latent) variables. SEM is often used for both to assess non-observable *hidden* constructs (i.e., latent variables) by means of observed variables, and to evaluate the relations among latent constructs and among manifest variables. In SEM, variables (manifest or latent) are considered (*i*) endogenous if they are dependent, i.e., related to a set of variables that explain them; (*ii*) exogenous if they are independent, i.e., explain a set of variables. Note that endogenous variables may also cause other endogenous variables. SEM has the property to estimate the multiple and interrelated dependencies in a single analysis by combining factor analysis and multivariate regression analysis. SEM has been used in many different fields, as in economics and social sciences, in marketing for example to assess customer satisfaction (Steenkamp & Baumgartner 2000, Richter et al. 2016, Rigdon 2016). Then, SEM allows to build latent variables (LVs), such as customer satisfaction, through a network of manifest variables (MVs).

Covariance structure approach (CSA) (Jöreskog 1978) and partial least squares (PLS) (Lohmöller 1989) are the two alternative statistical techniques for estimating

such models. The reader may refer to Sarstedt et al. (2016) and Rigdon et al. (2017) in order to understand PLS-SEM as a different statistical method from CSA.

An important research objective in the PLS-SEM context is the assessment of the potential validity threats if the researchers do not account for unobserved heterogeneity. In this direction Jedidi et al. (1997) propose a simultaneous procedure based on finite mixture estimated via the expectation-maximization (EM) algorithm (Dempster et al. 1977, McLachlan et al. 2004). Hahn et al. (2002) affirm that this technique extends CSA, but it is inappropriate for PLS-SEM. They propose the finite mixture partial least squares (FIMIX-PLS) approach that joins a finite mixture procedure with an EM algorithm specifically regarding the ordinary least predictions of PLS. Sarstedt (2008) and Sarstedt & Ringle (2010) review this technique and concludes that FIMIX-PLS can currently be viewed as the most comprehensive and commonly used approach for capturing heterogeneity in PLS-SEM. Following the guidelines of Jedidi et al. (1997) and Hahn et al. (2002), Ringle et al. (2010) present FIMIX-PLS, implemented for the first time in a statistical software application, called Smart-PLS (Ringle et al. 2005). Vinzi et al. (2008) propose a new method for unobserved heterogeneity detection in PLS-SEM: response-based procedure for detecting unit segments in PLS (REBUS-PLS). REBUS-PLS does not require distributional hypotheses but may lead to local models that are different in terms of both structural and measurement models. In fact, separate PLS-SEM are estimated for each cluster, and the results are compared in order to identify, if possible, differences among component scores, structural coefficients and different loadings. This is certainly an interesting feature, which has the unique problem of complicating the interpretation of results, since the number of the SEM parameters to be mentioned increases at the increasing of the number of clusters. Following this idea, Squillacciotti (2010) proposes a technique, called PLS typological path modeling (PLS-TPM), that allows to take into account the predictive purpose of PLS techniques when the classes are defined.

Other methods of PLS-SEM segmentation approach include prediction oriented segmentation in PLS path models (PLS-POS) proposed by Becker et al. (2013), genetic algorithm segmentation in partial least squares path modeling (PLS-GAS) proposed by Ringle et al. (2014), and particularly segmentation of PLS path models through iterative reweighted regressions (PLS-IRRS) proposed by Schlittgen et al. (2016). For more details see also Sarstedt et al. (2017). Schlittgen et al. (2016) conclude that PLS-IRRS gives similar quality results in comparison with PLS-GAS, and it is generally applicable to all kinds of PLS path models. Moreover, the PLS-IRRS computations are extremely fast.

In the current literature of PLS-SEM segmentation here examined, the existing methods are almost all model-based segmentation approaches that try to find ho-

mogenous groups in terms of the structural and/or measurement model relations. They do not directly focus on mean differences in the observed of latent variables. In this work, we propose a new approaches named partial least squares K-means (PLS-SEM-KM). PLS-SEM-KM is based on the simultaneous optimization of PLS-SEM and reduced k-means (De Soete & Carroll 1994), where centroids of clusters are located in the reduced space of the LVs, thus, ensuring the optimal partition of the statistical units on the best latent hyper-plane defined by the structural/measurement relations estimated by the pre-specified model. In this way, we segment the population under the analysis and simultaneously identify the structural and measurement relations that have produced that segmentation. These relations, not segment specific, represent a consensus of those that can be obtained by applying PLS-SEM for each cluster. In fact, a relevant issue in marketing is the measurement of the customer satisfaction by using PLS-SEM and at the same time the identification of distinctive customer segments (Ter Hofstede et al. 1999, Wu & DeSarbo 2005, Wedel & Kamakura 2012).

Moreover, a different approach to select the optimal number of segments K is provided. Note that in all the segmentation methods discussed above, researchers must pre-specify a number of segments (clusters) when running the procedure. The optimal number of segments is usually unknown. Ringle et al. (2014) and Schlittgen et al. (2016) propose to firstly run FIMIX-PLS (Hahn et al. 2002, Sarstedt & Ringle 2010) to determine the number of segments and, then, subsequently run PLS-GAS or PLS-IRRS to obtain the final segmentation solution. However, it has been argued that the underlying assumption of a limited number of segments of individuals that are perfectly homogeneous within segments in finite mixture models is too restrictive (Wedel & Kamakura 2012). Whereas, PLS-SEM-KM algorithm includes the optimal K selection through the *gap statistics* proposed by Tibshirani et al. (2001).

The chapter is structured as follows: in the section 3.2 a detailed background on SEM estimated via PLS procedure is provided; in section 3.3 the PLS-SEM-KM model is presented and the PLS algorithm is given; in section 3.4 the performances of PLS-SEM-KM are tested in a detailed simulation study providing a comparison with the FIMIX-PLS approach proposed by Hahn et al. (2002); in section 3.5 the results obtained by an application on real data are shown.

3.2 Structural equation modeling

Before showing the modeling details, the notation and terminology used in this work is here presented to allow the reader to easily follow the subsequent formalizations and algebraic elaborations.

n, J	# of:	observations, MVs
H, L, P	# of:	exogenous LVs, endogenous LVs, LVs ($P = H + L$)
K	# of:	clusters
$\mathbf{\Xi}$	$n \times H$	exogenous LVs matrix
$\mathbf{H}$	$n \times L$	endogenous LVs matrix
$\mathbf{Y}$	$n \times P$	scores matrix ($\mathbf{Y} = [\mathbf{\Xi}, \mathbf{H}]$)
$\mathbf{\Gamma}$	$L \times H$	path coefficients matrix of the exogenous LVs
$\mathbf{B}$	$L \times L$	path coefficients matrix of the endogenous LVs
$\mathbf{Z}$	$n \times L$	errors matrix of the endogenous LVs
$\mathbf{X}$	$n \times J$	data matrix
$\mathbf{E}$	$n \times J$	errors matrix of the data
$\mathbf{\Lambda}_H$	$J \times H$	loadings matrix of the exogenous LVs
$\mathbf{\Lambda}_L$	$J \times L$	loadings matrix of the endogenous LVs
$\mathbf{\Lambda}$	$J \times P$	loadings matrix ($\mathbf{\Lambda} = [\mathbf{\Lambda}_H, \mathbf{\Lambda}_L]$)
$\mathbf{T}$	$n \times H$	errors matrix of the exogenous LVs
$\mathbf{\Delta}$	$n \times L$	errors matrix of the endogenous LVs
$\mathbf{U}$	$n \times K$	membership matrix (binary and row stochastic)

Partial least squares (PLS) methodologies are algorithmic tools with analytic proprieties aiming at solving problems about the stringent assumptions on data, e.g., distributional assumptions that are hard to meet in real life (Tenenhaus et al. 2005, Vinzi et al. 2010). Tenenhaus et al. (2005) try to better clarify the terminology used in the PLS field through an interesting review of the literature, focusing the attention on the structural equation models (SEM) standpoint. Usually, a PLS-SEM (called also PLS-PM, i.e., PLS path model) consists in a combination of two models:

- a *structural model* (or inner model), that specifies the relationships between latent variables (LV). In this context, a LV is an non-observable variable (i.e., a theoretical construct) indirectly described by a block of observable variables which are called manifest variables (MVs);

- a *measurement model* (or outer model), that relates the MVs to their own LVs.

3.2.1 Structural model

Let $\mathbf{X}$ be a $n \times J$ data matrix, summarized by P latent variables ($j = 1, \ldots, J$; $p = 1, \ldots, P$ and $P \leq J$), let $\mathbf{H}$ be the $n \times L$ matrix of the endogenous LVs with generic element $\eta_{i,l}$, and let $\mathbf{\Xi}$ be the $n \times H$ matrix of the exogenous LVs with generic element $\xi_{i,h}$, the structural model is a causality model that relates the P LVs each other through a set of linear equations (Vinzi et al. 2010). In matrix form:

$$\mathbf{H} = \mathbf{H}\mathbf{B}^T + \mathbf{\Xi}\mathbf{\Gamma}^T + \mathbf{Z}, \tag{3.1}$$

where $\mathbf{B}$ is the $L \times L$ matrix of the path coefficients $\beta_{l,l}$ associated to the endogenous latent variables, $\mathbf{\Gamma}$ is the $L \times H$ matrix of the path coefficients $\gamma_{l,h}$ associated to the exogenous latent variables, and $\mathbf{Z}$ is the $n \times L$ matrix of the residual terms $\zeta_{i,l}$.
EXAMPLE 1. An example of structural model is shown in Figure 3.1.

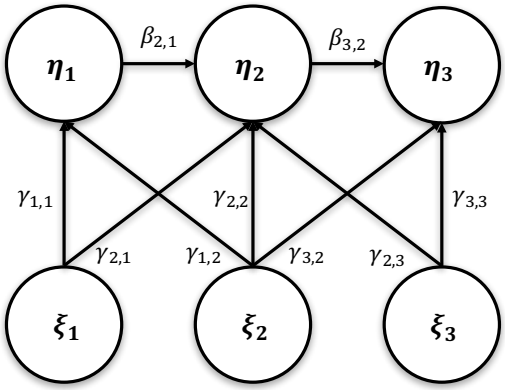


Figure 3.1: Example of structural model with three endogenous LVs and three exogenous LVs

The structural equations related to the path diagram in Figure 3.1 are shown in compact matrix form in Equation (3.2).

$$\begin{bmatrix}\eta_1\\ \eta_2\\ \eta_3\end{bmatrix}^T = \begin{bmatrix}\eta_1\eta_2\eta_3\end{bmatrix}\begin{bmatrix}0 & 0 & 0\\ \beta_{2,1} & 0 & 0\\ 0 & \beta_{3,2} & 0\end{bmatrix}^T + \begin{bmatrix}\xi_1\xi_2\xi_3\end{bmatrix}\begin{bmatrix}\gamma_{1,1} & \gamma_{1,2} & 0\\ \gamma_{2,1} & \gamma_{2,2} & \gamma_{2,3}\\ 0 & \gamma_{3,2} & \gamma_{3,3}\end{bmatrix}^T + \begin{bmatrix}\zeta_1\\ \zeta_2\\ \zeta_3\end{bmatrix}^T . \quad (3.2)$$

3.2.2 Measurement model

In PLS-SEM, unlike traditional SEM approach, there are two ways to relate MVs to their LVs: *reflective* way and *formative* way (Diamantopoulos & Winklhofer 2001, Tenenhaus et al. 2005). In the reflective way it is supposed that each MV reflects its LV, i.e., the observed variables are considered as the effect of the latent construct; a reflective measurement model can be written in matrix form as

$$\begin{aligned}\mathbf{X} &= \mathbf{Y}\mathbf{\Lambda}^T + \mathbf{E}\\ &= \begin{bmatrix}\mathbf{\Xi} & \mathbf{H}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}_H^T\\ \mathbf{\Lambda}_L^T\end{bmatrix} + \mathbf{E}\\ &= \mathbf{\Xi}\mathbf{\Lambda}_H^T + \mathbf{H}\mathbf{\Lambda}_L^T + \mathbf{E},\end{aligned} \quad (3.3)$$

where $\mathbf{\Lambda}_H$ is the $J \times H$ loadings matrix of the exogenous latent constructs with generic element $\lambda_{j,h}$, $\mathbf{\Lambda}_L$ is the $J \times L$ loadings matrix of the endogenous latent constructs with generic element $\lambda_{j,l}$, and $\mathbf{E}$ is the $n \times J$ residuals matrix with element

$\epsilon_{i,j}$, under hypothesis of zero mean and is uncorrelated with $\xi_{i,h}$ and $\eta_{i,l}$. Then, the reflective way implies that each MV is related to its LV by a set of simple regression models with coefficients $\lambda_{j,l}$. Note that, in reflective way it is necessary that the block of MVs is unidimensional with respect to related LV. This condition can be checked through different tools, as well as principal component analysis (PCA), Cronbach's alpha and Dillon-Goldstein's (Vinzi et al. 2010, Sanchez 2013).

Conversely, in the formative way each MV is supposed "*forming*" its LV, i.e., the observed variables are considered as the cause of the latent construct. Formally, in the case of exogenous latent construct the model can be written as

$$\mathbf{\Xi} = \mathbf{X}\mathbf{\Lambda}_H + \mathbf{T}, \tag{3.4}$$

whereas, in the case of endogenous latent construct the model can be written as

$$\mathbf{H} = \mathbf{X}\mathbf{\Lambda}_L + \mathbf{\Delta}, \tag{3.5}$$

where $\mathbf{T}$ and $\mathbf{\Delta}$ are, respectively, the $n \times H$ and $n \times L$ errors matrices with element $\tau_{i,h}$ and $\delta_{i,l}$, under the hypothesis of zero mean and is uncorrelated with $x_{i,j}$. Then, the formative way implies that each MV is related to its LV by a multiple regression model with coefficients λ's.

EXAMPLE 2. In Figure 3.2 are shown two examples of PLS-SEM with three latent constructs (η_1, ξ_1, and ξ_2) and six observed variables (x_1, x_2, x_3, x_4, x_5, and x_6). In particular, there are two exogenous LVs (ξ_1 and ξ_2) and one endogenous LV (η_1). The MVs are related to their LVs in reflective way (left plot) and formative way (right plot).

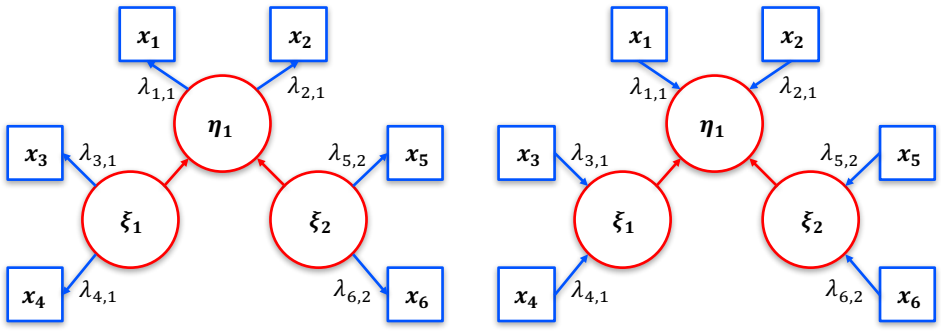


Figure 3.2: Two examples of PLS path model with three LVs and six MVs: reflective measurement models (left) and formative measurement models (right)

Formally, the reflective measurement models represented in the left plot of Figure

3.2 can be written as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix}^T = \begin{bmatrix} \xi_1 & \xi_2 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \lambda_{3,1,} & 0 \\ \lambda_{4,1} & 0 \\ 0 & \lambda_{5,2} \\ 0 & \lambda_{6,2} \end{bmatrix}_H^T + \begin{bmatrix} \eta_1 \end{bmatrix} \begin{bmatrix} \lambda_{1,1} \\ \lambda_{2,1} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}_L^T + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}^T, \tag{3.6}$$

whereas, for the formative measurement models, we can use Equation (3.7) in the case of exogenous LVs, and Equation (3.8) in the case of endogenous LVs.

$$\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}^T = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \lambda_{3,1,} & 0 \\ \lambda_{4,1} & 0 \\ 0 & \lambda_{5,2} \\ 0 & \lambda_{6,2} \end{bmatrix}_H^T + \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix}, \tag{3.7}$$

$$\begin{bmatrix} \eta_1 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{bmatrix} \begin{bmatrix} \lambda_{1,1} \\ \lambda_{2,1} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}_L^T + \begin{bmatrix} \delta_1 \end{bmatrix}. \tag{3.8}$$

3.3 Partial Least Squares K-Means

3.3.1 Model and algorithm

Given the $n \times J$ data matrix $\mathbf{X}$, the $n \times K$ membership matrix $\mathbf{U}$, the $K \times J$ centroids matrix $\mathbf{C}$, the $J \times P$ loadings matrix $\mathbf{\Lambda}$=$[\mathbf{\Lambda}_H, \mathbf{\Lambda}_L]$, and the errors matrices $\mathbf{Z}$ $(n \times L)$ and $\mathbf{E}$ $(n \times J)$, the partial least squares K-means (PLS-SEM-KM) model can be written as follows:

$$\begin{aligned} \mathbf{H} &= \mathbf{H}\mathbf{B}^T + \mathbf{\Xi}\mathbf{\Gamma}^T + \mathbf{Z} \\ \mathbf{X} &= \mathbf{Y}\mathbf{\Lambda}^T + \mathbf{E} = \mathbf{\Xi}\mathbf{\Lambda}_H^T + \mathbf{H}\mathbf{\Lambda}_L^T + \mathbf{E} \\ \mathbf{X} &= \mathbf{U}\mathbf{C}\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{E} = \mathbf{U}\mathbf{C}\mathbf{\Lambda}_H\mathbf{\Lambda}_H^T + \mathbf{U}\mathbf{C}\mathbf{\Lambda}_L\mathbf{\Lambda}_L^T + \mathbf{E}, \end{aligned} \tag{3.9}$$

subject to constraints: (i) $\mathbf{\Lambda}^T\mathbf{\Lambda} = \mathbf{I}$; and (ii) $\mathbf{U} \in \{0, 1\}$, $\mathbf{U}\mathbf{1}_K = \mathbf{1}_n$. Thus, the PLS-SEM-KM model includes the PLS and the clustering equations (i.e., $\mathbf{X} = \mathbf{UC}$

and then, $\mathbf{Y} = \mathbf{X}\mathbf{\Lambda}$ becomes $\mathbf{Y} = \mathbf{U}\mathbf{C}\mathbf{\Lambda}$). In fact, the third set of equations is the reduced K-means model (De Soete & Carroll 1994). The simultaneous estimation of the three sets of equations will produce the estimation of the presupposed SEM describing relations among variables and the corresponding best partitioning of units.

PLS-SEM-KM model belongs to class of methodologies for the simultaneous unsupervised classification and dimensionality reduction in a PLS-SEM context. The method does not born as a segmentation-PLS approach for identifying segment specific relations but could be placed in the clustering field where a PLS analysis is performed.

When applying PLS-SEM-KM, the number of groups K is unknown and the identification of an appropriate number of clusters is not a straightforward task. Several statistical criteria have been proposed. In this work we use the *gap method* discussed in Tibshirani et al. (2001) embedded in the algorithm for estimating simultaneously also the number of clusters, i.e., a *pseudo*-F designed to be applicable to virtually any clustering method. The *gap method* can also may be applicable to any model-based clustering approach without restrictive assumptions on the scores distribution. Given: the $n \times J$ standardized data matrix $\mathbf{X}$; the $J \times P$ design matrix of the measurement model $\mathbf{D}_\Lambda$, with binary elements equal to 1 if a MV is associated to a LV and 0 otherwise; the $P \times P$ path design matrix of the structural model $\mathbf{D}_B$, with binary elements equal to 1 if a latent exogenous or endogenous variable explains a latent endogenous variable and 0 otherwise. Note that the matrix $\mathbf{D}_B$ is symmetrized.
$\mathbf{Y}_h$ is the h-th exogenous latent score and $\mathbf{Y}_l$ is the l-th endogenous latent score; the symbol $\otimes$ indicates the element-wise product of two matrices, while $*$ indicates the adjacent latent scores matrix, i.e., the set of latent scores that are related to the $\mathbf{Y}_h$ or $\mathbf{Y}_l$. The PLS-SEM-KM algorithm is a development of the Wold's original algorithm used to the PLS-SEM estimate in Lohmöller (1989). As you can see from the step 7 of the algorithm (i.e., in the loadings estimation), the method is performed for both reflective measurement models and formative measurement models. $\mathbf{U}$ matrix is optimized row by row solving an assignment problem through the objective function in the step 8 of the algorithm.
Therefore, the algorithm produces a matrix $\mathbf{U}$ of the segments assignment and a matrix $\mathbf{C}$ of centroids with a unique common measurement and structural model coefficients. However, researchers that wish determining segment specific measurement and structural model coefficients can apply group-specific PLS-SEM analysis. The unique measurement and structural model coefficients is interpreted as a consensus of the segment specific coefficients.

Algorithm 2 PLS-SEM-KM algorithm

1: Initialize $\mathbf{\Lambda} = \mathbf{D}_\Lambda$;
Choose K through the *gap method* applied on scores matrix $\mathbf{Y} = \mathbf{X\Lambda}$;
$\omega = 10^{-12}$, iter=0, maxiter=300;
2: Random generate the memberships matrix $\mathbf{U}$;
Compute centers matrix $\mathbf{C} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{X}$;
Compute latent scores matrix $\mathbf{Y} = \mathbf{UC\Lambda}$;
3: iter=iter+1;

Inner approximation

4: Estimate covariance matrix $\mathbf{\Sigma}_Y = n^{-1}\mathbf{Y}^T\mathbf{JY}$ (with $\mathbf{J} = \mathbf{I}n^{-1}\mathbf{1}\mathbf{1}^T$);
5: Compute inner weights $\mathbf{W} = \mathbf{D}_B \otimes \mathbf{\Sigma}_Y$;
6: Estimate new scores $\mathbf{Y}_W = \mathbf{YW}$;

Outer approximation

7: Update $\mathbf{\Lambda} \to \mathbf{\Lambda}_n = \mathbf{C}^T\mathbf{U}^T\mathbf{Y}_W(\mathbf{Y}_W^T\mathbf{Y}_W)^{-1}$; (Reflective way)
$\to \mathbf{\Lambda}_n = (\mathbf{C}^T\mathbf{U}^T\mathbf{UC})^{-1}\mathbf{C}^T\mathbf{U}^T\mathbf{Y}_W$; (Formative way)
8: Update $\mathbf{U} \to \underset{\mathbf{U}}{\operatorname{argmin}} \left\|\mathbf{X} - \mathbf{UC\Lambda}_n\mathbf{\Lambda}_n^T\right\|^2$,
subject to $\mathbf{\Lambda}_n^T\mathbf{\Lambda}_n = \mathbf{1}_P$, $\mathbf{U} = \{0,1\}$, $\mathbf{U1}_K = \mathbf{1}_n$;
9: Compute new centers $\mathbf{C}_n = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{X}$;

Stopping rule

10: Update $K \to Kn$ through the *gap method* applied on scores matrix $\mathbf{Y} = \mathbf{UC_n\Lambda_n}$
11: **if** $K_n \neq K$
go to step 2
12: **else**
13: **if** $\|\mathbf{C\Lambda} - \mathbf{C}_n\mathbf{\Lambda}_n\|^2 > \omega$ & iter<maxiter, $\mathbf{C} = \mathbf{C}_n$, $\mathbf{\Lambda} = \mathbf{\Lambda}_n$;
repeat step 3-12;
14: **else**
exit loop 3-12;
15: **end if**
16: **end if**

Path coefficients estimation

17: **for** $l = 1$ to L **do**
18: **for** $h = 1$ to H **do**
19: Compute $\mathbf{Y}_h = \mathbf{X\Lambda}_h$
20: Compute $\mathbf{Y}_l = \mathbf{X\Lambda}_l$
21: Compute $\mathbf{\Gamma} = (\mathbf{Y}_{h^*}^T\mathbf{Y}_{h^*})^{-1}\mathbf{Y}_{h^*}^T\mathbf{Y}_l$
22: Compute $\mathbf{B} = (\mathbf{Y}_{l^*}^T\mathbf{Y}_{l^*})^{-1}\mathbf{Y}_{l^*}^T\mathbf{Y}_l$
23: **end for**
24: **end for**

Note that, in the PLS-SEM-KM algorithm centroids matrix $\mathbf{C}$ and the loadings matrix $\mathbf{\Lambda}$ simultaneously converge. It is important to remember that the algorithm, given the constraints on $\mathbf{U}$, can be expected to be rather sensitive to *local optima*. For this reasons, it is recommended the use of some randomly started runs to find

the best solution. The algorithm chooses the best solution among the randomly started repetitions through the maximization of the R^{2*} index discussed in the next subsection.

3.3.2 Local and global fit measures

In PLS-SEM context, there is not a well-identified global optimization criterion to assess the goodness of the model, since PLS-SEM models are variance-based models strongly oriented to prediction and its validation mainly is focused on the predictive capability.
According to the PLS-SEM approach, each part of the model needs to be validated: the measurement model, the structural model and the global model (Vinzi et al. 2010). In particular, PLS-SEM provides different fit indices: the *communality* index for the measurement models, the R^2 index for the structural models and the *Goodness of Fit* (GoF) index for the overall model. However, there is some literature criticizing the global GoF for the use in PLS-SEM (Henseler & Sarstedt 2013).

Communalities are simply the squared correlations between MVs and the corresponding LV. Then, communalities measure the part of the covariance between a latent variable and its block of observed variables that is common to both. For the j-th manifest variable of the p-th latent score they are calculated as

$$com(x_{j,p}, y_p) = corr^2(x_{j,p}, y_p). \tag{3.10}$$

Usually, for each p-th block of MVs in the PLS-SEM model, the quality of the entire measurement model is assessed by the mean of the communality indices as in Equation (3.11).

$$com_p(x_{j,p}, y_p) = J_p^{-1} \sum_{j=1}^{J_p} corr^2(x_{j,p}, y_p), \tag{3.11}$$

where J_p is the number of the MVs in the p-th block. Note that, for each endogenous LV of the structural model we have an R^2 interpreted similarly as in any multiple regression model. Then, we use the R^{2*} to indicate the amount mean of variance in the L endogenous latent constructs explained by its independent latent variables (Vinzi et al. 2010, Sanchez 2013):

$$R^{2*} = \bar{R}^2_L. \tag{3.12}$$

3.4 Simulation study

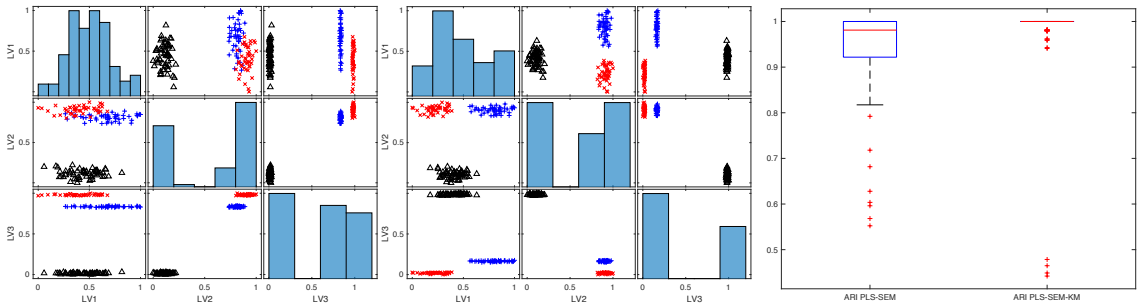
In this section, we have prepared a simulation study for assessing the performances of the partial least squares K-means (PLS-SEM-KM) algorithm through a comparison with finite mixture partial least squares (FIMIX-PLS) proposed by Hahn et al. (2002). Firstly, a motivational example with comparing PLS-SEM-KM and the usual PLS analysis is shown.

3.4.1 Motivational example

We know that for classification aim, the researcher could consider the sequential approach of applying first PLS-SEM in order to determine the LVs and then apply a clustering methodology such as K-means or Gaussian mixture model (GMM) clustering on the latent scores of the PLS-SEM in order to obtain homogeneous clusters. However, Sarstedt & Ringle (2010) empirically illustrate the shortcomings of using a sequential approach. Now, through a simulation example we show that the sequential approach of PLS-SEM and K-means (i.e., the *tandem analysis*) may fail to find the clustering structure in the data, as well as in the cases described also in Vichi & Kiers (2001), where the researcher applies a dimensionality reduction technique, such as principal component analysis (PCA) or factor analysis (FA), and then applies a cluster analysis methodology on the factor scores. The simulated data set is formed by two exogenous LVs, having a clustering structure into three groups. Then, an endogenous LV has been generated by a Normal distribution. This could be interpreted as a synthetic marketing data set where customers are split in "satisfied", "neither satisfied nor unsatisfied" and "unsatisfied" according to the "perceived value" and "perceived quality" constructs (exogenous LVs) and the "satisfaction" construct (endogenous LV).

In Figure 3.3a the scatterplot matrix of the three LVs shows the performance of the sequential application of PLS-SEM and K-means. The clusters are not well separated; in fact, the adjusted Rand index (ARI) (Rand 1971, Hubert & Arabie 1985) between the generated partition and the partition obtained by K-means computed on the LVs of the PLS-SEM is equal to 0.64. Note that ARI is equal to 0 when two random partitions are compared and it is equal to 1 when two identical partitions are compared. In Figure 3.3b the scatterplot matrix of the LVs shows the simultaneous application of PLS-SEM and K-means as proposed in this work with a specific methodology. This time the clusters are well separated and the partition is clearly detected. ARI is equal to 1, i.e., the new methodology exactly identifies the generated partition. We might think that this is only a simulated example. Thus,

we have repeated the analysis on other 300 data sets generated as above described. The boxplot of the distribution of the 300 adjusted Rand index (ARI) evaluations (Figure 3.3c) confirms that, in almost all cases the new methodology exactly finds the generated partition (mean of ARI equal to 0.98), while the sequential application of PLS-SEM and K-means finds the true partition in 15% of cases and the mean of ARI is equal to 0.65.



(a) (b) (c)

Figure 3.3: Left figure represents the scatterplot-matrix of the LVs estimated by the sequential application of PLS-SEM and K-means; center figure represents the scatterplot-matrix of the LVs estimated by the simultaneous application of PLS-SEM and K-means; right figure represents the boxplot of ARI distribution between the true and estimated partition obtained by the sequential and simultaneous approaches on 300 data sets.

Thus, we can conclude that the sequential application of PLS-SEM and K-means is an unreliable approach to obtain the best partition of the units and the best structural equation modeling analysis when data are heterogeneous as those simulated. This example clarify that the researcher is interested in finding simultaneously the measurement and structural relations that are useful to identify homogeneous segments present in the population. This result strongly motivates the study of a new model that identifies simultaneously the best clustering and the best manifest variables reconstructed by a unique common set of measurement/structural relationships.

3.4.2 Simulation scheme

We have simulated data matrices formed by different samples of statistical units and 9 MVs ($n \times 9$ data matrices) with a structural of K groups. The 9 generated variables are split in three blocks related to 3 LVs according the path diagram shown in Figure 3.4. On the two exogenous LVs the measurement model is both in reflective way (left) and formative way (right).

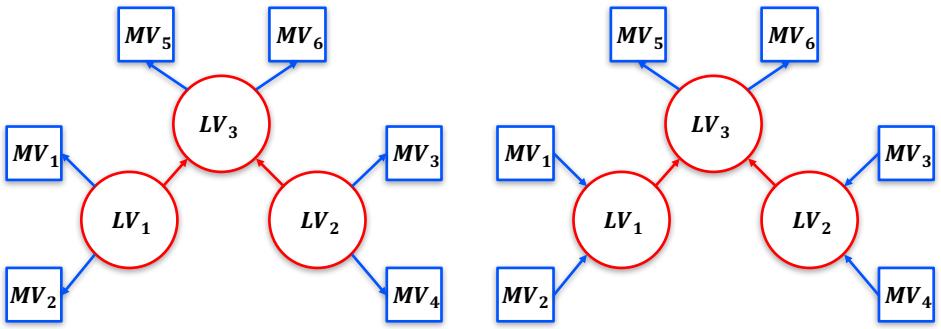


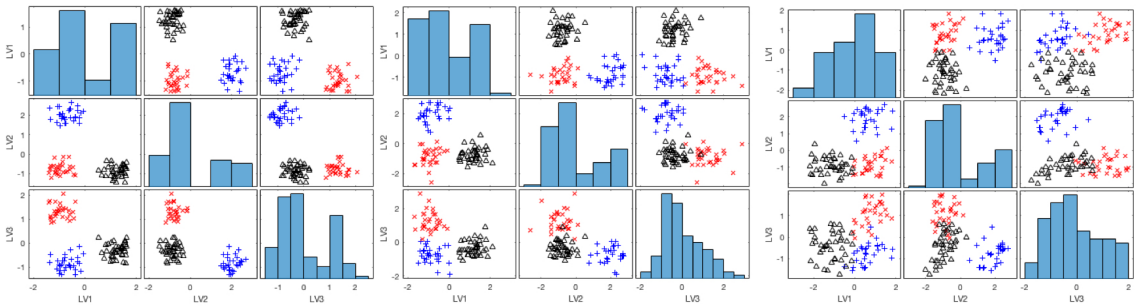
Figure 3.4: Path diagrams of the measurement models specified by the simulation scheme

Then, data matrices have been generated by a mixture of K different Normal distributions with group-specific mean vectors and covariance matrices for $K = 3$ and $K = 4$ as reported below:

3 Groups		4 Groups	
$\mu_1 = [\ 0,\ 10]^T;$	$\boldsymbol{\Sigma}_1 = \sigma^2\mathbf{I}_3$	$\mu_1 = [-10,\ 10]^T;$	$\boldsymbol{\Sigma}_1 = \sigma^2\mathbf{I}_4$
$\mu_2 = [-10, -10]^T;$	$\boldsymbol{\Sigma}_2 = \sigma^2\mathbf{I}_3$	$\mu_2 = [\ 10,\ 10]^T;$	$\boldsymbol{\Sigma}_1 = \sigma^2\mathbf{I}_4$
$\mu_3 = [\ 10, -10]^T;$	$\boldsymbol{\Sigma}_3 = \sigma^2\mathbf{I}_3$	$\mu_3 = [-10, -10]^T;$	$\boldsymbol{\Sigma}_1 = \sigma^2\mathbf{I}_4$
		$\mu_4 = [\ 10, -10]^T;$	$\boldsymbol{\Sigma}_1 = \sigma^2\mathbf{I}_4$

Moreover, in order to mask the groups structure in the data, we have added an errors matrix $\mathbf{E}$ generated by a multivariate Normal distribution (9 uncorrelated dimensions) with means equal to zero (i.e., noise) and standard deviation fixed as: $\sigma = 1.5$ (*low error*), $\sigma = 2.5$ (*medium error*), $\sigma = 3.5$ (*high error*). In Figure 3.5a the scatterplot-matrix of a random generation of the latent scores with low error is shown. The three generated groups with three different colors (30 points blue, 30 points red and 40 points black) and three different symbols ($+$, $\times$, and $\triangle$) are very well-separated and homogenous. Different results are shown with medium error (Figure 3.5b), where the three groups not well-separated, mostly between the latent scores ξ_2 and η_1. Finally, Figure 3.5c shows the results obtained with high error that, obviously, correspond to the most confused situation: the groups are not separated and homogeneous, and there is an overlap in all the three couple of latent dimensions.

Moreover, for better investigating the performance of both PLS-SEM-KM and FIMIX-PLS we have realized a simulation study that represents different data constellations that could occur in empirical applications. According to recent simulation studies on PLS segmentation (Becker et al. 2013, Ringle et al. 2014, Schlittgen et al. 2016), we have selected the following experimental factors:



(a) (b) (c)

Figure 3.5: Scatterplot-matrix of (standardized) generated data with low, medium and high error.

- *Number of observations*: **small sample size** ($n = 150$); **large sample size** ($n = 300$).
- *Number of segments (clusters)*: $K = 3$; $K = 4$.
- *Segments size*: **balanced** (mixture proportion when $K = 3$: $p_1 = 0.33$, $p_2 = 0.33$, $p_3 = 0.34$; mixture proportion when $K = 4$: $p_1 = p_2 = p_3 = p_4 = 0.25$); **unbalanced 1** (mixture proportion when $K = 3$: $p_1 = 0.66$, $p_2 = 0.17$, $p_3 = 0.17$; mixture proportion when $K = 4$: $p_1 = 0.40$, $p_2 = 0.20$, $p_3 = 0.20$, $p_4 = 0.20$); **unbalanced 2** (mixture proportion when $K = 3$: $p_1 = 0.15$, $p_2 = 0.42$, $p_3 = 0.43$; mixture proportion when $K = 4$: $p_1 = 0.10$, $p_2 = 0.30$, $p_3 = 0.30$, $p_4 = 0.30$).
- *Standard deviation of data generation error*: **low error** ($\sigma = 1.5$); **medium error** ($\sigma = 2.5$); **high error** ($\sigma = 3.5$).
- *PLS measurement model*: **Model 1**: (reflective-Mode A) shown in left plot of Figure 3.4; **Model 2**: (formative-Mode B) shown in right plot of Figure 3.4.

In order to have more stable results, we have randomly generated 100 datasets for each factor level combination. Then, in particular we have $2 \times 2 \times 3 \times 3 \times 2 \times 100 = 7200$ generated datasets.

3.4.3 Results

We have separated the simulation results in 4 different contexts, each of them with 18 different experimental cases random repeated 100 times. Table 3.1 shown the 18 different experimental cases.
In particular, we have context 1: *path model 1 and* $K = 3$, context 2: *path model 2*

Table 3.1: Experimental cases list of the simulation study

Exp. Case	Sample Size	Segments size	Error level
1	small	balanced	low
2	small	balanced	medium
3	small	balanced	high
4	small	unbalanced 1	low
5	small	unbalanced 1	medium
6	small	unbalanced 1	high
7	small	unbalanced 2	low
8	small	unbalanced 2	meidum
9	small	unbalanced 2	high
10	large	balanced	low
11	large	balanced	medium
12	large	balanced	high
13	large	unbalanced 1	low
14	large	unbalanced 1	medium
15	large	unbalanced 1	high
16	large	unbalanced 2	low
17	large	unbalanced 2	medium
18	large	unbalanced 2	high

and $K = 3$, context 3: *path model 1 and* $K = 4$, and context 4: *path model 2 and* $K = 4$.

For evaluating the performance of the models we have used the R^{2*} index shown in Equation 3.12. In Table 3.2 we can see the arithmetic mean and the standard deviation of the R^{2*} values obtained for each experimental case of the first and second simulated context by PLS-SEM-KM and FIMIX-PLS, respectively.
Similarly, in Table 3.3 we can see the arithmetic mean and the standard deviation of each R^{2*} distribution obtained for each experimental case of the third and fourth simulated context by PLS-SEM-KM and FIMIX-PLS, respectively. Tables 3.2 and 3.3 show that the results obtained by PLS-SEM-KM are in almost all cases better than model FIMIX-PLS. In Context 1, where the path model 1 has been considered, the difference between PLS-SEM-KM and FIMIX-PLS are more relevant for cases from 4 to 9 (from 22% to 29% better) and for cases from 13 to 18 (from 23% to 31% better), corresponding to the unbalanced cases.
In the Context 2 differences are still in favor of the PLS-SEM-KM, but with a less relevant magnitude (no more than 14%), this time in the balanced cases. Also in contexts 3 and 4 the performance obtained by PLS-SEM-KM are almost always better than that obtained by FIMIX-PLS. Furthermore, in terms of statistical significance in almost cases the variability of the R^{2*} distribution in FIMIX-PLS is bigger than that shown in PLS-SEM-KM.

Moreover, the results show also that R^{2*} index for both PLS-SEM-KM and FIMIX-PLS reduces with the increase of the number of segments. This is expected because the probability of misclassification increases. Note that in FIMIX-PLS, as

Table 3.2: Mean and standard deviation of R^{2*} obtained by of PLS-SEM-KM and FIMIX-PLS for all experimental cases of the first and second simulated context

	Context 1				Context 2			
	PLS-SEM-KM		FIMIX-PLS		PLS-SEM-KM		FIMIX-PLS	
Case	$\mu_{R^{2*}}$	$\sigma_{R^{2*}}$	$\mu_{R^{2*}}$	$\sigma_{R^{2*}}$	$\mu_{R^{2*}}$	$\sigma_{R^{2*}}$	$\mu_{R^{2*}}$	$\sigma_{R^{2*}}$
1	0.982	0.028	0.975	0.117	0.979	0.061	0.961	0.134
2	0.969	0.027	0.956	0.121	0.971	0.022	0.970	0.136
3	0.951	0.029	0.911	0.299	0.924	0.131	0.922	0.153
4	0.979	0.019	0.873	0.322	0.982	0.020	0.877	0.342
5	0.955	0.049	0.899	0.332	0.963	0.037	0.821	0.422
6	0.936	0.062	0.900	0.345	0.939	0.064	0.800	0.452
7	0.978	0.023	0.897	0.356	0.982	0.023	0.854	0.358
8	0.952	0.036	0.888	0.346	0.945	0.082	0.821	0.367
9	0.938	0.032	0.856	0.398	0.939	0.090	0.810	0.379
10	0.984	0.019	0.977	0.021	0.984	0.029	0.910	0.116
11	0.964	0.029	0.966	0.018	0.947	0.098	0.899	0.118
12	0.950	0.029	0.949	0.020	0.939	0.120	0.934	0.125
13	0.978	0.024	0.874	0.312	0.972	0.076	0.844	0.314
14	0.958	0.043	0.896	0.299	0.958	0.041	0.831	0.333
15	0.938	0.047	0.877	0.333	0.936	0.050	0.897	0.334
16	0.982	0.016	0.853	0.278	0.981	0.019	0.855	0.278
17	0.954	0.044	0.855	0.299	0.962	0.029	0.819	0.299
18	0.913	0.055	0.800	0.310	0.937	0.039	0.821	0.299

such as in other segmentation models, the correct identification of the number of clusters (segments) is not easy when the number of segments increases. This because FIMIX-PLS follows a mixture regression concept that allows the estimation of separate linear regression functions, and in this way the number of parameters exponentially increases when the number of segments increase, and the usual criteria based on likelihood function, such as AIC and BIC become not very reliable (Bulteel et al. 2013).

Furthermore, it is useful recall that we have generated data from normal mixture model; thus FIMIX-PLS is advantaged since the data for the simulation study are generated according to the FIMIX-PLS hypotheses, by assuming that each endogenous latent variable η_l is distributed as a finite mixture of conditional multivariate normal densities (Ringle et al. 2010). Conversely, in PLS-SEM-KM there are not particular assumption on the distribution of data.

In order to understand the performance of PLS-SEM-KM algorithm we have also

Table 3.3: Mean and standard deviation of the R^{2*} obtained by of PLS-SEM-KM and FIMIX-PLS for all experimental cases of the third and fourth simulated context

	Context 3				Context 4			
	PLS-SEM-KM		FIMIX-PLS		PLS-SEM-KM		FIMIX-PLS	
Case	$\mu_{R^{2*}}$	$\sigma_{R^{2*}}$	$\mu_{R^{2*}}$	$\sigma_{R^{2*}}$	$\mu_{R^{2*}}$	$\sigma_{R^{2*}}$	$\mu_{R^{2*}}$	$\sigma_{R^{2*}}$
1	0.951	0.062	0.944	0.213	0.986	0.014	0.911	0.322
2	0.933	0.098	0.932	0.230	0.968	0.032	0.898	0.342
3	0.924	0.054	0.900	0.244	0.947	0.035	0.834	0.365
4	0.947	0.053	0.900	0.344	0.984	0.019	0.899	0.347
5	0.935	0.067	0.879	0.368	0.973	0.024	0.874	0.365
6	0.916	0.057	0.877	0.377	0.943	0.033	0.832	0.388
7	0.963	0.041	0.851	0.346	0.988	0.013	0.823	0.384
8	0.939	0.065	0.846	0.375	0.962	0.056	0.810	0.399
9	0.921	0.055	0.842	0.385	0.937	0.038	0.800	0.399
10	0.972	0.031	0.893	0.210	0.988	0.017	0.892	0.313
11	0.947	0.057	0.866	0.231	0.962	0.072	0.881	0.333
12	0.923	0.059	0.842	0.265	0.928	0.095	0.890	0.373
13	0.961	0.049	0.893	0.398	0.987	0.015	0.834	0.372
14	0.945	0.044	0.821	0.397	0.965	0.039	0.831	0.389
15	0.911	0.070	0.811	0.399	0.933	0.070	0.810	0.399
16	0.959	0.040	0.864	0.213	0.982	0.018	0.852	0.342
17	0.932	0.071	0.833	0.364	0.953	0.086	0.822	0.355
18	0.920	0.066	0.821	0.388	0.938	0.055	0.814	0.387

studied the presence of *local minima* and situations of *overfitting*. Then, once established that there are cases where the adjusted Rand index (ARI) is lower than 1 (i.e., the real partition is not identified), it is useful to analyze the single case where the real partition has not been found by PLS-SEM-KM algorithm. Table 3.4 shows the performance of the model in terms of clustering capability for 100 randomly chosen experimental conditions. The second column of the table shows the percentage of times the *gap method*, discussed in Section 3.3, identifies the real number of clusters ($K = 3$ or $K = 4$). The third column shows the percentage of times the algorithm finds the true partition ($ARI = 1$), while the fourth and the fifth columns show the percentage of *local minima* and of *overfitting*, respectively (i.e., when $ARI < 1$). In particular, we have *local minima* when the performance (in terms of R^2) obtained through the partition identified by the model is better than the performance obtained through the generated real partition. Otherwise, we have *overfitting*.

Table 3.4: Performance of the PLS-SEM-KM algorithm using a single random start in the three different error levels for 100 randomly chosen experimental conditions (percentage values)

Sd. Error	Optimal K	Model is *true*	Local minima	Overfitting
$\sigma = 1.5$	100.00	99.40	0.10	0.50
$\sigma = 2.5$	100.00	77.30	7.20	15.50
$\sigma = 3.5$	100.00	75.40	9.40	16.00

We can try to reduce the number of *local minima* by increasing the number of initial random starts. In these cases, the use of 15 random starts usually suffices, when the error is not very high ($\sigma = 1.5$), and the groups structure is not masked. Indeed, the algorithm finds the optimal solution in 99.40%; while there are 0.10% of *local minima* cases and 0.50% of *overfitting*. However, in the cases where the groups structure is masked as in the case of medium and high level of error, the algorithm cannot completely eliminates the number of *local minima*. In these two cases the algorithm finds the optimal solution in 77.30% and 75.40% of cases, respectively. Thus, it is advisable to increase initial random starts when the clustering of the data is not clear. Then, the algorithm chooses the best solution among the 15 repetitions through the maximization of the R^2 index.

3.5 Application on real data

In this section an application on real data of the partial least squares K-means (PLS-SEM-KM) model is presented. For this application the European Consumer Satisfaction Index (ECSI) has been considered analyzing the ECSI approach in mobile phone industry (Bayol et al. 2000, Tenenhaus et al. 2005).

3.5.1 ECSI model for the mobile phone industry

The dataset consists in 24 observed variables that represent the answers of 250 consumers of a mobile phone provider. We have chosen this data set which represents a very well-known benchmark used to show many new methodologies in PLS-SEM. In Figure 3.6 is represented the complete ECSI model for the mobile phone industry. For underlining the good results obtained by PLS-SEM-KM a comparison with a normal PLS-SEM analysis has been done. In this way, we prove that the PLS-SEM-KM algorithm add the clustering aim to the simple PLS-SEM without change the causal relationships among latent constructs and manifest variables.

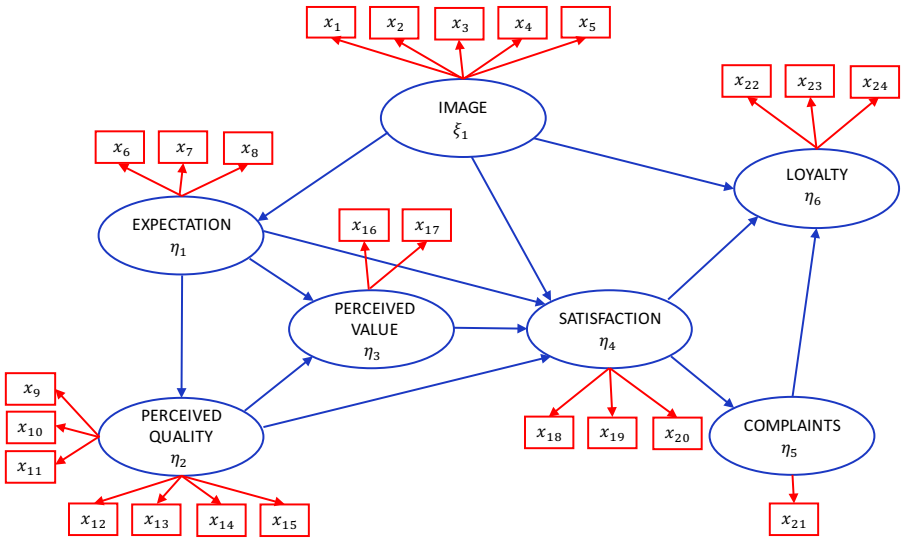


Figure 3.6: ECSI model for the mobile phone industry

The MVs included in dataset are the following:

x_1: It can be trusted in what it says and does

x_2: It is stable and firmly established

x_3: It has a social contribution for the society

x_4: It is concerned with customers

x_5: It is innovative and forward looking

x_6: Expectations for the overall quality of "your mobile phone provider" at the moment you became customer of this provider

x_7: Expectations for "your mobile phone provider" to provide products and services to meet your personal need

x_8: How often did you expect that things could go wrong at "your mobile phone provider"

x_9: Overall perceived quality

x_{10}: Technical quality of the network

x_{11}: Customer service and personal advice offered

x_{12}: Quality of the services you use

x_{13}: Range of services and products offered

x_{14}: Reliability and accuracy of the products and services provided

x_{15}: Clarity and transparency of information provided

x_{16}: Given the quality of the products and services offered by "your mobile phone provider" how would you rate the fees and prices that you pay for them?

x_{17}: Given the fees and prices that you pay for "your mobile phone provider" how would you rate the quality of the products and services offered by "your mobile

phone provider"?
$\mathbf{x_{18}}$: Overall satisfaction
$\mathbf{x_{19}}$: Fulfillment of expectations
$\mathbf{x_{20}}$: How well do you think "your mobile phone provider" compares with your ideal mobile phone provider?
$\mathbf{x_{21}}$: You complained about "your mobile phone provider" last year.
How well, or poorly, was your most recent complaint handled; or
You did not complain about "your mobile phone provider" last year.
Imagine you have to complain to "your mobile phone provider" because of a bad quality of service or product. To what extent do you think that "your mobile phone provider" will care about your complaint?
$\mathbf{x_{22}}$: If you would need to choose a new mobile phone provider how likely is it that you would choose "your provider" again?
$\mathbf{x_{23}}$: Let us now suppose that other mobile phone providers decide to lower their fees and prices, but "your mobile phone provider" stays at the same level as today. At which level of difference (in %) would you choose another mobile phone provider?
$\mathbf{x_{24}}$: If a friend or colleague asks you for advice, how likely is it that you would recommend "your mobile phone provider"?

3.5.2 Results

With applying the PLS-SEM-KM algorithm on the ECSI data, we have identified a number of clusters $K = 3$, with the corresponding value of *pseudo*-F equal to 1.3994 as shown in Figure 3.7.

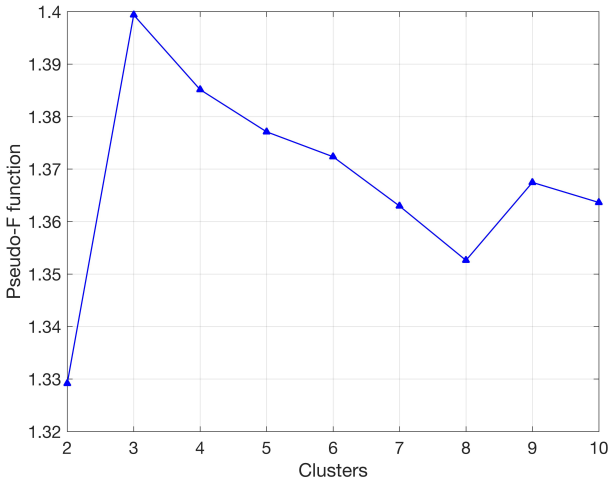


Figure 3.7: *Pseudo*-F function obtained via gap method in PLS-SEM-KM algorithm from 2 to 10 clusters

Table 3.5 includes the loading values obtained by PLS-SEM-KM and PLS-SEM, respectively. Note that in the PLS-SEM-KM model, the loading matrix $\mathbf{\Lambda}$ is normalized (see Section 3 for details), then we need to normalize also the loading matrix obtained by the PLS-SEM analysis to compare the results. From Table 3.5 we can note that the models obtain very similar results, only in some particular case the PLS-SEM loadings are slightly bigger than that obtained by PLS-SEM-KM.

Table 3.5: Loading values estimated by PLS-SEM-KM and PLS-SEM

Measurement model	PLS-SEM-KM	PLS-SEM
Image $\rightarrow x_{01}$	0.449	0.482
Image $\rightarrow x_{02}$	0.398	0.388
Image $\rightarrow x_{03}$	0.355	0.373
Image $\rightarrow x_{04}$	0.528	0.497
Image $\rightarrow x_{05}$	0.486	0.481
Expectation $\rightarrow x_{06}$	0.615	0.642
Expectation $\rightarrow x_{07}$	0.607	0.576
Expectation $\rightarrow x_{08}$	0.503	0.506
Perceived quality $\rightarrow x_{09}$	0.419	0.400
Perceived quality $\rightarrow x_{10}$	0.284	0.318
Perceived quality $\rightarrow x_{11}$	0.399	0.390
Perceived quality $\rightarrow x_{12}$	0.377	0.383
Perceived quality $\rightarrow x_{13}$	0.375	0.376
Perceived quality $\rightarrow x_{14}$	0.381	0.386
Perceived quality $\rightarrow x_{15}$	0.397	0.388
Perceived value $\rightarrow x_{16}$	0.624	0.694
Perceived value $\rightarrow x_{17}$	0.781	0.720
Satisfaction $\rightarrow x_{18}$	0.558	0.554
Satisfaction $\rightarrow x_{19}$	0.563	0.587
Satisfaction $\rightarrow x_{20}$	0.609	0.590
Complaints $\rightarrow x_{21}$	1.000	1.000
Loyalty $\rightarrow x_{22}$	0.585	0.656
Loyalty $\rightarrow x_{23}$	0.099	0.171
Loyalty $\rightarrow x_{24}$	0.805	0.735

Whereas, In Table 3.6 the path coefficients obtained by PLS-SEM-KM and PLS-SEM are shown, respectively. Note that the latent scores used for the path coefficients estimation are standardized (i.e., the path coefficients are correlations). From the structural models comparison, we note that, like to the measurement model, the estimation results are very similar between the two approaches.
Table 3.6 shows that for both methods the *Image* construct has a positive relationship with all its endogenous LVs, though it has a stronger effect on the *Expectations* construct (0.51 and 0.49, respectively) than *Satisfaction* (0.18 and 0.15, respectively)

Table 3.6: Path coefficients estimated by PLS-SEM-KM and PLS-SEM

Structural model	PLS-SEM-KM	PLS-SEM
Image → Expectation	0.507	0.493
Image → Satisfaction	0.177	0.153
Image → Loyalty	0.201	0.212
Expectation → Perceived quality	0.554	0.545
Expectation → Perceived value	0.048	0.066
Expectation → Satisfaction	0.071	0.037
Perceived quality → Perceived value	0.557	0.540
Perceived quality → Satisfaction	0.509	0.544
Perceived value → Satisfaction	0.191	0.200
Satisfaction → Complaints	0.523	0.540
Satisfaction → Loyalty	0.479	0.466
Complaints → Loyalty	0.067	0.050

and *Loyalty* (0.20 and 0.21, respectively). The *Expectations* construct has a significant effect on the *Perceived Quality* only (0.55 and 0.54, respectively), while it has very low effect on the *Perceived Value* (0.05 and 0.07, respectively) and *Satisfaction* (0.07 and 0.04, respectively). The *Perceived Quality* block has effect on *Perceived Value* (0.56 and 0.54, respectively) and *Satisfaction* (0.51 and 0.54, respectively). The *Perceived Value* construct has an effect equal to 0.19 for PLS-SEM-KM and 0.20 for PLS-SEM, respectively, on the *Satisfaction*, which has an effect equal to 0.52 and 0.54 on the *Complaints*. Finally, the *Complaints* construct has effect on the *Loyalty* only, with a correlation level equal to 0.07 and 0.05, respectively.

Now, we show the results obtained on the model assessment. In Table 3.7 we can see a comparison of the fit measures obtained on each latent construct by PLS-SEM-KM and PLS-SEM, respectively.

Table 3.7: Fit measures computed on each block of MVs in PLS-SEM-KM and PLS-SEM

	PLS-SEM-KM		PLS-SEM	
	Communality	R-Squared	Communality	R-Squared
Image	0.200	-	0.476	-
Expectations	0.333	0.257	0.471	0.243
Perceived quality	0.143	0.307	0.574	0.297
Perceived value	0.500	0.342	0.849	0.335
Satisfaction	0.333	0.677	0.682	0.672
Complaints	1.000	0.274	1.000	0.292
Loyalty	0.333	0.454	0.520	0.432
Average	0.592	0.385	0.570	0.378

From these results, we can say that the PLS-SEM-KM model shows performances

slightly better than PLS-SEM in terms of both communalities and R^2. In particular, for the PLS-SEM-KM model we have obtained the communality average equal to 0.5916 and the R^2 average equal to 0.3852. Then, in summary we can say that the PLS-SEM-KM model does not change the quality of the causal relationships estimation. In other words, our proposed model keeps the PLS structure adding the clustering aim to the usual analysis.

The last step of the analysis is the description of the groups defined by PLS-SEM-KM model. Table 3.8 shows the summary statistics of the three found groups computed on the seven normalized latent scores.

Table 3.8: Summary statistics of the three groups of mobile phone customers

	Group 1 ($n = 92$)						
	ξ_1	η_1	η_2	η_3	η_4	η_5	η_6
Min	0.460	0.180	0.660	0.000	0.537	0.000	0.019
$Q1$	0.722	0.652	0.775	0.688	0.710	0.778	0.824
$Median$	0.802	0.773	0.837	0.778	0.787	0.889	0.898
$Mean$	0.796	0.752	0.840	0.763	0.794	0.832	0.862
$Q3$	0.861	0.849	0.905	0.878	0.875	1.000	0.956
Max	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Group 2 ($n = 112$)						
	ξ_1	η_1	η_2	η_3	η_4	η_5	η_6
Min	0.225	0.145	0.483	0.000	0.273	0.000	0.190
$Q1$	0.541	0.481	0.594	0.511	0.526	0.556	0.594
$Median$	0.600	0.584	0.648	0.622	0.599	0.667	0.698
$Mean$	0.607	0.584	0.643	0.591	0.589	0.638	0.696
$Q3$	0.681	0.664	0.687	0.667	0.647	0.778	0.804
Max	0.845	1.000	0.831	0.889	1.000	1.000	1.000
	Group 3 ($n = 46$)						
	ξ_1	η_1	η_2	η_3	η_4	η_5	η_6
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$Q1$	0.306	0.359	0.284	0.333	0.272	0.333	0.263
$Median$	0.440	0.497	0.414	0.444	0.353	0.444	0.467
$Mean$	0.392	0.471	0.398	0.423	0.345	0.447	0.460
$Q3$	0.494	0.599	0.486	0.556	0.445	0.667	0.626
Max	0.676	0.820	0.704	1.000	0.691	1.000	1.000

The first group, formed by 92 observations, indicates a *highly satisfied* profile of customers (central values around the 0.8); the second group, formed by 112 observations, indicates a *medially satisfied* profile of customers (central values around 0.6); the third group, formed by 46 observations, indicates a *lowly satisfied* profile of customers (central values around the 0.4).

Finally, to show that our proposal could be a useful tool also for group-specific

segmentation, in Table 3.9 a comparison between the group-specific structural models estimated by PLS-SEM-KM and FIMIX- PLS is shown.

Table 3.9: Group-specific structural models estimated by PLS-SEM-KM and FIMIX-PLS

Structural model	PLS-SEM-KM			FIMIX-PLS		
	k_1	k_2	k_3	k_1	k_2	k_3
Image → Expectation	0.305	0.011	-0.231	0.289	0.419	0.437
Image → Satisfaction	0.158	0.157	0.227	0.211	0.073	0.281
Image → Loyalty	0.019	0.217	0.145	0.046	0.300	0.121
Expectation → Perceived quality	0.349	0.137	0.357	0.470	0.316	0.508
Expectation → Perceived value	0.082	-0.159	0.059	0.077	0.084	0.172
Expectation → Satisfaction	-0.006	0.133	0.005	0.046	0.112	0.016
Perceived quality → Perceived value	0.169	0.191	0.314	0.391	0.443	0.439
Perceived quality → Satisfaction	0.409	0.309	0.210	0.524	0.522	0.313
Perceived value → Satisfaction	0.081	0.281	0.379	0.174	0.266	0.207
Satisfaction → Complaints	0.289	0.057	0.349	0.349	0.324	0.293
Satisfaction → Loyalty	0.303	0.397	0.338	0.536	0.477	0.527
Complaints → Loyalty	0.016	-0.078	0.172	-0.300	0.113	-0.011
R^2 Expectations	0.099	0.171	0.188	0.084	0.175	0.191
R^2 Perceived quality	0.241	0.015	0.243	0.221	0.010	0.258
R^2 Perceived value	0.179	0.233	0.305	0.187	0.227	0.300
R^2 Satisfaction	0.645	0.644	0.457	0.618	0.617	0.445
R^2 Complaints	0.145	0.112	0.112	0.122	0.105	0.086
R^2 Loyalty	0.298	0.489	0.355	0.308	0.543	0.362
Average	0.268	0.277	0.276	0.220	0.253	0.235
Segment size	36.8%	44.8%	18.4%	34.8%	43.6%	21.6%

From the results we can see that PLS-SEM-KM obtains R^2 almost always better than FIMIX-PLS, even if only slightly. Very different are the estimated path coefficients by both approaches. In particular, seems that in FIMIX-PLS the three identified segments do not particularly discriminate the structural relationships.

3.6 Concluding remarks

In a wide range of applications, the assumption that data are collected from a single homogeneous population, is often unrealistic, and the identification of different groups (clusters) of observations constitutes a critical issue in many fields.

This work is focused on the structural equation modeling (SEM) in the PLS-SEM context, (i.e., SEM estimated via partial least squares (PLS) method), when the data are heterogeneous and tend to form clustering structures. We know that the traditional approach to clustering in SEM consists in estimating separate models for each cluster, where the partition is *a priori* specified by the researcher or obtained via clustering methods. Conversely, the partial least squares K-means (PLS-SEM-KM) approach, provides a single model that guarantees the best partition of objects

represented by the best causal relationship in the reduced latent space. Moreover, our proposal, unlike the recent proposed methods, does not mainly focus on the heterogeneous structural or measurement model relations (i.e., the group-specific structural and measurement models identification) but on the mean differences of individuals profile definition (i.e., prototypes) of segments, based on the isolation (i.e., between cluster variance) and homogeneity (i.e., between cluster variance) criteria of the groups.

The simulation study has highlighted a good reliability of the model, which guarantees good results in different experimental cases, when the data have a clustering structure; conversely, the sequential approach to use PLS- SEM followed by clustering on the latent variables may fail to identify the correct clusters. The simulation study shows that in almost all experimental cases PLS-SEM-KM achieves better than finite mixture partial least squares (FIMIX-PLS) model proposed by Hahn et al. (2002). Moreover, we recall FIMIX-PLS in the simulation study has been advantaged since it is based on the assumption that each endogenous latent construct is distributed as a finite mixture of multivariate normal densities, and we have generated data from mixtures of normal distributions. However, imposition of a distributional assumption on the endogenous latent variables may prove to be problematic. This criticism gains force when one considers that PLS path modeling is generally preferred to covariance structure analysis (CSA) in circumstances where assumptions of multivariate normality cannot be made (Ringle et al. 2012). Conversely, in PLS-SEM-KM there are not distributional assumptions. Another problem that was found for FIMIX-PLS, as such as for other segmentation models, is the correct identification of the number of clusters (segments) when it increases since the approach follows a mixture of regressions; concept that needs the estimation of separate linear regression functions. In this way the number of parameters exponentially increases at the increasing of the number of segments, and the usual criteria based on likelihood function, as such as AIC and BIC are not very reliable (Bulteel et al. 2013). In the PLS-SEM-KM algorithm the gap-method proposed by Tibshirani et al. (2001) is used, and the simulation study shows that the real number of clusters is identified in 100% of cases in all the simulated contexts.

On the other hand, in the application on real data we can say that PLS-SEM-KM, in the optimal case (i.e., when the causal structure of the model well-represents the partition that characterizes the data), does not particularly modify the results on the structural and measurement models obtained by the simple PLS-SEM as shown in literature (Bayol et al. 2000, Tenenhaus et al. 2005). Also in comparison with FIMIX-PLS the results obtained by PLS-SEM-KM are generally better. Moreover, the PLS-SEM-KM results in Table 3.9 show that it could be a useful tool also for group-specific segments identification.

However, in future research could be interesting to evaluate the PLS-SEM-KM performance against the more recent approaches, as prediction oriented segmentation in PLS path models (PLS-POS) proposed by Becker et al. (2013), genetic algorithm segmentation in partial least squares path modeling (PLS-GAS) proposed by Ringle et al. (2014), and particularly segmentation of PLS path models through iterative reweighted regressions (PLS-IRRS) proposed by Schlittgen et al. (2016).

Bibliography

Agrawal, R., Gehrke, J., Gunopulos, D. & Raghavan, P. (1998), *Automatic subspace clustering of high dimensional data for data mining applications*, Vol. 27, ACM.

Arabie, P. & Hubert, L. (1996), Advances in cluster analysis relevant to marketing research, *in* 'From Data to Knowledge', Springer, pp. 3–19.

Balabin, R. M., Safieva, R. Z. & Lomakina, E. I. (2010), 'Gasoline classification using near infrared (nir) spectroscopy data: Comparison of multivariate techniques', *Analytica Chimica Acta* **671**(1-2), 27–35.

Bayol, M.-P., de la Foye, A., Tellier, C. & Tenenhaus, M. (2000), 'Use of pls path modelling to estimate the european consumer satisfaction index (ecsi) model', *Statistica Applicata* **12**(3), 361–375.

Becker, J.-M., Rai, A., Ringle, C. M. & Völckner, F. (2013), 'Discovering unobserved heterogeneity in structural equation models to avert validity threats', *Mis Quarterly* pp. 665–694.

Bellincontro, A., Taticchi, A., Servili, M., Esposto, S., Farinelli, D. & Mencarelli, F. (2012), 'Feasible application of a portable nir-aotf tool for on-field prediction of phenolic compounds during the ripening of olives for oil production', *Journal of agricultural and food chemistry* **60**(10), 2665–2673.

Benzécri, J.-P. (1979), 'Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [bin. mult.]', *Cahiers de l'Analyse des Données* **4**(3), 377–378.

Bolton, R. & Krzanowski, W. (2003), 'Projection pursuit clustering for exploratory data analysis', *Journal of Computational and Graphical Statistics* **12**(1), 121–142.

Brereton, R. G. & Lloyd, G. R. (2014), 'Partial least squares discriminant analysis: taking the magic away', *Journal of Chemometrics* **28**(4), 213–225.

Bühlmann, P. & Van De Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Bulteel, K., Wilderjans, T. F., Tuerlinckx, F. & Ceulemans, E. (2013), 'Chull as an alternative to aic and bic in the context of mixtures of factor analyzers', *Behavior Research Methods* **45**(3), 782–791.

Campbell, N. A. (1980), 'Shrunken estimators in discriminant and canonical variate analysis', *Applied Statistics* pp. 5–24.

Cawley, G. C. & Talbot, N. L. (2003), 'Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers', *Pattern Recognition* **36**(11), 2585–2592.

Cayuela, J. A. & Camino, M. d. C. P. (2010), 'Prediction of quality of intact olives by near infrared spectroscopy', *European journal of lipid science and technology* **112**(11), 1209–1217.

Clemmensen, L., Hastie, T., Witten, D. & Ersbøll, B. (2011), 'Sparse discriminant analysis', *Technometrics* **53**(4), 406–413.

Cortes, C. & Vapnik, V. (1995), 'Machine learning', *Support vector networks* **20**, 273–297.

Davies, D. L. & Bouldin, D. W. (1979), 'A cluster separation measure', *IEEE transactions on pattern analysis and machine intelligence* (2), 224–227.

De Soete, G. & Carroll, J. D. (1994), K-means clustering in a low-dimensional euclidean space, *in* 'New approaches in classification and data analysis', Springer, pp. 212–219.

De Soete, G. & Heiser, W. J. (1993), 'A latent class unfolding model for analyzing single stimulus preference ratings', *Psychometrika* **58**(4), 545–565.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22.

Desarbo, W., Jedidi, K., Cool, K. & Schendel, D. (1991), 'Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups', *Marketing Letters* **2**(2), 129–146.

Devijver, P. A. & Kittler, J. (1982), *Pattern recognition: A statistical approach*, Prentice hall.

Diamantopoulos, A. & Winklhofer, H. M. (2001), 'Index construction with formative indicators: An alternative to scale development', *Journal of marketing research* **38**(2), 269–277.

Ding, B. & Gentleman, R. (2005), 'Classification using generalized partial least squares', *Journal of Computational and Graphical Statistics* **14**(2), 280–298.

Dudoit, S., Fridlyand, J. & Speed, T. P. (2002), 'Comparison of discrimination methods for the classification of tumors using gene expression data', *Journal of the American statistical association* **97**(457), 77–87.

Fisher, T. J. & Sun, X. (2011), 'Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix', *Computational Statistics & Data Analysis* **55**(5), 1909–1918.

Fix, E. & Hodges, J. L. (1989), 'Discriminatory analysis. nonparametric discrimination: consistency properties', *International Statistical Review/Revue Internationale de Statistique* **57**(3), 238–247.

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

Gallardo, L., Osorio, E. & Sanchez, J. (2005), 'Application of near infrared spectroscopy (nirs) for the real-time determination of moisture and fat contents in olive pastes and wastes of oil extraction', *Alimentación Equipos y Tecnologia* **24**(206), 85–89.

Galvan, V., Gorostiza, O. F., Banwait, S., Ataie, M., Logvinova, A. V., Sitaraman, S., Carlson, E., Sagi, S. A., Chevallier, N., Jin, K. et al. (2006), 'Reversal of alzheimer's-like pathology and behavior in human app transgenic mice by mutation of asp664', *Proceedings of the National Academy of Sciences* **103**(18), 7130–7135.

Garcia, J. M., Seller, S. & Perez-Camino, M. C. (1996), 'Influence of fruit ripening on olive oil quality', *Journal of agricultural and food chemistry* **44**(11), 3516–3520.

Greenacre, M. J. (1984), 'Theory and applications of correspondence analysis. 1984'.

Guo, Y., Hastie, T. & Tibshirani, R. (2006), 'Regularized linear discriminant analysis and its application in microarrays', *Biostatistics* **8**(1), 86–100.

Guyon, I., Makhoul, J., Schwartz, R. & Vapnik, V. (1998), 'What size test set gives good error rate estimates?', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(1), 52–64.

Hahn, C., Johnson, M. D., Herrmann, A. & Huber, F. (2002), 'Capturing customer heterogeneity using a finite mixture pls approach', *Schmalenbach Business Review* **54**(3), 243–269.

Hardy, O. J., Maggia, L., Bandou, E., Breyne, P., Caron, H., CHEVALLIER, M.-H., Doligez, A., Dutech, C., Kremer, A., LATOUCHE-HALLÉ, C. et al. (2006), 'Fine-scale genetic structure and gene dispersal inferences in 10 neotropical tree species', *Molecular ecology* **15**(2), 559–571.

Hastie, T., Buja, A. & Tibshirani, R. (1995), 'Penalized discriminant analysis', *The Annals of Statistics* pp. 73–102.

Heiser, W. J. (1993), Clustering in low-dimensional space, *in* 'Information and classification', Springer, pp. 162–173.

Henseler, J. & Sarstedt, M. (2013), 'Goodness-of-fit indices for partial least squares path modeling', *Computational Statistics* **28**(2), 565–580.

Hubert, L. & Arabie, P. (1985), 'Comparing partitions', *Journal of classification* **2**(1), 193–218.

Hwang, H., Montréal, H., Dillon, W. R. & Takane, Y. (2006), 'An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents', *Psychometrika* **71**(1), 161–171.

Izenman, A. J. (2013), Linear discriminant analysis, *in* 'Modern multivariate statistical techniques', Springer, pp. 237–280.

Jedidi, K., Jagpal, H. S. & DeSarbo, W. S. (1997), 'Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity', *Marketing Science* **16**(1), 39–59.

Jimenez, L. O. & Landgrebe, D. A. (1998), 'Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data', *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **28**(1), 39–54.

Joachims, T. (2005), A support vector method for multivariate performance measures, *in* 'Proceedings of the 22nd international conference on Machine learning', ACM, pp. 377–384.

Jöreskog, K. G. (1978), 'Structural analysis of covariance and correlation matrices', *Psychometrika* **43**(4), 443–477.

Kemsley, E. (1996), 'Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods', *Chemometrics and intelligent laboratory systems* **33**(1), 47–61.

Kriegel, H.-P., Kröger, P. & Zimek, A. (2009), 'Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering', *ACM Transactions on Knowledge Discovery from Data (TKDD)* **3**(1), 1.

León, L., Garrido-Varo, A. & Downey, G. (2004), 'Parent and harvest year effects on near-infrared reflectance spectroscopic analysis of olive (olea europaea l.) fruit traits', *Journal of agricultural and food chemistry* **52**(16), 4957–4962.

Lohmöller, J.-B. (1989), Predictive vs. structural modeling: Pls vs. ml, *in* 'Latent variable path modeling with partial least squares', Springer, pp. 199–226.

MacQueen, J. et al. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, Oakland, CA, USA, pp. 281–297.

Martens, H. (2001), 'Reliable and relevant modelling of real world data: a personal account of the development of pls regression', *Chemometrics and intelligent laboratory systems* **58**(2), 85–95.

McLachlan, G., Krishnan, T. & Ng, S. (2004), 'The em algorithm (no. 2004, 24)', *Papers/Humboldt-Universtat. Berlin: Center for Applied Statistics and Economics* .

Misaki, M., Kim, Y., Bandettini, P. A. & Kriegeskorte, N. (2010), 'Comparison of multivariate classifiers and response normalizations for pattern-information fmri', *Neuroimage* **53**(1), 103–118.

Peck, R. & Van Ness, J. (1982), 'The use of shrinkage estimators in linear discriminant analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence* (5), 530–537.

Pérez-Enciso, M. & Tenenhaus, M. (2003), 'Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (pls-da) approach', *Human genetics* **112**(5-6), 581–592.

Rand, W. M. (1971), 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical association* **66**(336), 846–850.

Rao, C. R. (1948), 'The utilization of multiple measurements in problems of biological classification', *Journal of the Royal Statistical Society. Series B (Methodological)* **10**(2), 159–203.

Richter, N. F., Cepeda, G., Roldán, J. L. & Ringle, C. M. (2016), 'European management research using partial least squares structural equation modeling (pls-sem)', *European Management Journal* **34**(6), 589–597.

Rigdon, E. E. (2016), 'Choosing pls path modeling as analytical method in european management research: A realist perspective', *European Management Journal* **34**(6), 598–605.

Rigdon, E. E., Sarstedt, M. & Ringle, C. M. (2017), 'On comparing results from cb-sem and pls-sem: Five perspectives and five recommendations', *Marketing Zfp* **39**(3), 4–16.

Ringle, C. M., Sarstedt, M. & Schlittgen, R. (2014), 'Genetic algorithm segmentation in partial least squares structural equation modeling', *OR spectrum* **36**(1), 251–276.

Ringle, C. M., Sarstedt, M. & Straub, D. (2012), 'A critical look at the use of pls-sem in mis quarterly'.

Ringle, C. M., Wende, S. & Will, A. (2010), Finite mixture partial least squares analysis: Methodology and numerical examples, *in* 'Handbook of partial least squares', Springer, pp. 195–218.

Ringle, C., Wende, S. & Will, A. (2005), 'Smartpls 2.0 m3. university of hamburg, hamburg, germany'.

Sanchez, G. (2013), 'Pls path modeling with r', *Berkeley: Trowchez Editions* **383**, 2013.

Sarstedt, M. (2008), 'A review of recent approaches for capturing heterogeneity in partial least squares path modelling', *Journal of Modelling in Management* **3**(2), 140–161.

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O. & Gudergan, S. P. (2016), 'Estimation issues with pls and cbsem: Where the bias lies!', *Journal of Business Research* **69**(10), 3998–4010.

Sarstedt, M. & Ringle, C. M. (2010), 'Treating unobserved heterogeneity in pls path modeling: a comparison of fimix-pls with different data analysis strategies', *Journal of Applied Statistics* **37**(8), 1299–1318.

Sarstedt, M., Ringle, C. M. & Hair, J. F. (2017), Treating unobserved heterogeneity in pls-sem: a multi-method approach, *in* 'Partial Least Squares Path Modeling', Springer, pp. 197–217.

Schlittgen, R., Ringle, C. M., Sarstedt, M. & Becker, J.-M. (2016), 'Segmentation of pls path models by iterative reweighted regressions', *Journal of Business Research* **69**(10), 4583–4592.

Squillacciotti, S. (2010), Prediction oriented classification in pls path modeling, *in* 'Handbook of partial least squares', Springer, pp. 219–233.

Steenkamp, J.-B. E. & Baumgartner, H. (2000), 'On the use of structural equation models for marketing modeling', *International Journal of Research in Marketing* **17**(2-3), 195–202.

Suykens, J. A. & Vandewalle, J. (1999), 'Least squares support vector machine classifiers', *Neural processing letters* **9**(3), 293–300.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M. & Lauro, C. (2005), 'Pls path modeling', *Computational statistics & data analysis* **48**(1), 159–205.

Ter Hofstede, F., Steenkamp, J.-B. E. & Wedel, M. (1999), 'International market segmentation based on consumer–product relations', *Journal of Marketing Research* **36**(1), 1–17.

Thomaz, C. E., Kitani, E. C. & Gillies, D. F. (2006), 'A maximum uncertainty lda-based approach for limited sample size problems-with application to face recognition', *Journal of the Brazilian Computer Society* **12**(2), 7–18.

Tibshirani, R., Walther, G. & Hastie, T. (2001), 'Estimating the number of clusters in a data set via the gap statistic', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2), 411–423.

Timmerman, M. E., Ceulemans, E., Kiers, H. A. & Vichi, M. (2010), 'Factorial and reduced k-means reconsidered', *Computational Statistics & Data Analysis* **54**(7), 1858–1871.

Tran, T. N., Wehrens, R. & Buydens, L. M. (2006), 'Knn-kernel density-based clustering for high-dimensional multivariate data', *Computational Statistics & Data Analysis* **51**(2), 513–525.

Vichi, M. & Kiers, H. A. (2001), 'Factorial k-means analysis for two-way data', *Computational Statistics & Data Analysis* **37**(1), 49–64.

Vinzi, V., Chin, W. W., Henseler, J. & Wang, H. (2010), *Handbook of partial least squares: Concepts, methods and applications*, Heidelberg, Dordrecht, London, New York: Springer.

Vinzi, V., Trinchera, L., Squillacciotti, S. & Tenenhaus, M. (2008), 'Rebus-pls: A response-based procedure for detecting unit segments in pls path modelling', *Applied Stochastic Models in Business and Industry* **24**(5), 439–458.

Wang, Z. & Xue, X. (2014), Multi-class support vector machine, *in* 'Support Vector Machines Applications', Springer, pp. 23–48.

Wedel, M. & Kamakura, W. A. (2012), *Market segmentation: Conceptual and methodological foundations*, Vol. 8, Springer Science & Business Media.

Wehrens, R. & Mevik, B.-H. (2007), 'The pls package: principal component and partial least squares regression in r', *Journal of Statistical Software* **18**(2).

Weston, J. & Watkins, C. (1998), Multi-class support vector machines, Technical report, Citeseer.

Wichern, D. W. & Johnson, R. A. (1992), *Applied multivariate statistical analysis*, Vol. 4, Prentice Hall New Jersey.

Wold, S. (2001), 'Personal memories of the early pls development', *Chemometrics and Intelligent Laboratory Systems* **58**(2), 83–84.

Wu, J. & DeSarbo, W. S. (2005), 'Market segmentation for customer satisfaction studies via a new latent structure multidimensional scaling model', *Applied Stochastic Models in Business and Industry* **21**(4-5), 303–309.

Yu, H. & Yang, J. (2001), 'A direct lda algorithm for high-dimensional data Äîwith application to face recognition', *Pattern recognition* **34**(10), 2067–2070.

Zhang, H., Berg, A. C., Maire, M. & Malik, J. (2006), Svm-knn: Discriminative nearest neighbor classification for visual category recognition, *in* 'Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on', Vol. 2, IEEE, pp. 2126–2136.