4

Exploratory Data Analysis with Latent Subspace Models

José Camacho
Department of Signal Theory, Telematics and Communication, University of Granada, Granada Spain

1. Introduction

Exploratory Data Analysis (EDA) has been employed for decades in many research fields, including social sciences, psychology, education, medicine, chemometrics and related fields (1) (2). EDA is both a data analysis philosophy and a set of tools (3). Nevertheless, while the philosophy has essentially remained the same, the tools are in constant evolution. The application of EDA to current problems is challenging due to the large scale of the data sets involved. For instance, genomics data sets can have up to a million of variables (5). There is a clear interest in developing EDA methods to manage these scales of data while taking advantage of *the basic importance of simply looking at data* (3).

In data sets with a large number of variables, collinear data and missing values, projection models based on latent structures, such as Principal Component Analysis (PCA) (6) (7) (1) and Partial Least Squares (PLS) (8) (9) (10), are valuable tools within EDA. Projection models and the set of tools used in combination simplify the analysis of complex data sets, pointing out to special observations (outliers), clusters of similar observations, groups of related variables, and crossed relationships between specific observations and variables. All this information is of paramount importance to improve data knowledge.

EDA based on projection models has been successfully applied in the area of chemometrics and industrial process analysis. In this chapter, several standard tools for EDA with projection models, namely score plots, loading plots and biplots, are revised and their limitations are elucidated. Two recently proposed tools are introduced to overcome these limitations. The first of them, named Missing-data methods for Exploratory Data Analysis or MEDA for short (11), is used to investigate the relationships between variables in projection subspaces. The second one is an extension of MEDA, named observation-based MEDA or *o*MEDA (33), to discover the relationships between observations and variables. The EDA approach based on PCA/PLS with scores and loading plots, MEDA and *o*MEDA is illustrated with several real examples from the chemometrics field.

This chapter is organized as follows. Section 2 briefly discusses the importance of subspace models and score plots to explore the data distribution. Section 3 is devoted to the investigation of the relationship among variables in a data set. Section 4 studies the relationship between observations and variables in latent subspaces. Section 5 presents a EDA case study of Quantitative Structure-Activity Relationship (QSAR) modelling and

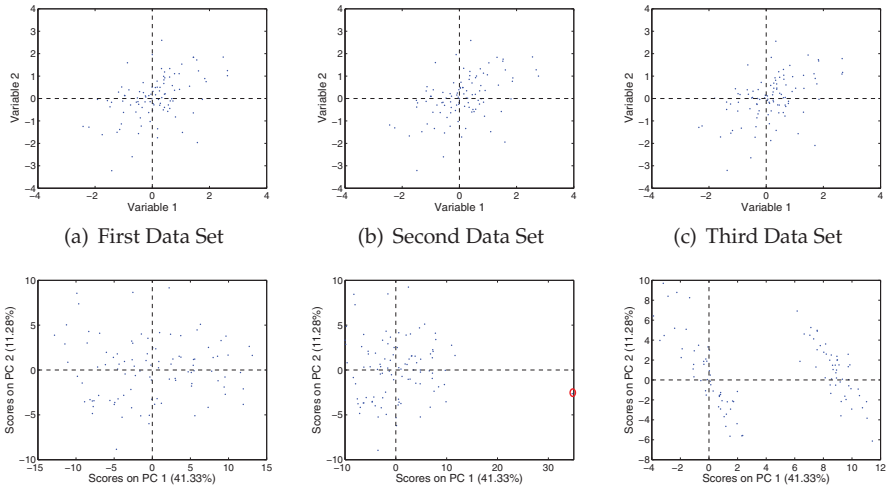
section 6 proposes some concluding remarks. Examples and Figures were computed using the MATLAB programming environment, with the PLS-Toolbox (32) and home-made software. A MATLAB toolbox with the tools employed in this chapter is available at http://wdb.ugr.es/ josecamacho/.

2. Patterns in the data distribution

The distribution of the observations in a data set contains relevant information for data understanding. For instance, in an industrial process, one outlier may represent an abnormal situation which affects the process variables to a large extent. Studying this observation with more detail, one may be able to identify if it is the result of a process upset or, very commonly, a sensor failure. Also, clusters of observations may represent different operation points. Outliers, clusters and trends in the data may be indicative of the degree of control in the process and of assignable sources of variation. The identification of these sources of variation may lead to the reduction of the variance in the process with the consequent reduction of costs.

The distribution of the observations can be visualized using scatter plots. For obvious reasons, scatter plots are limited to three dimensions at most, and typically to two dimensions. Therefore, the direct observation of the data distribution in data sets with several tens, hundreds or even thousands of variables is not possible. One can always construct scatter plots for selective pairs or thirds of variables, but this is an overwhelming and often misleading approach. Projection models overcome this problem. PCA and PLS can be used straightforwardly to visualize the distribution of the data in the latent subspace, considering only a few latent variables (LVs) which contain most of the variability of interest. Scatter plots of the scores corresponding to the LVs, the so-called score plots, are used for this purpose.

Score plots are well known and accepted in the chemometric field. Although simple to understand, score plots are paramount for EDA. The following example may be illustrative of this. In Figure 1, three simulated data sets of the same size (100×100) are compared. Data simulation was performed using the technique named Approximation of a DIstribution for a given COVariance matrix (15), or ADICOV for short. Using this technique, the same covariance structure was simulated for the three data sets but with different distributions: the first data set presents a multi-normal distribution in the latent subspace, the second one presents a severe outlier and the third one presents a pair of clusters. If the scatter plot of the observations in the plane spanned by the first two variables is depicted (first row of Figure 1), the data sets seem to be almost identical. Therefore, unless an extensive exploration is performed, the three data sets may be though to come from a similar data generation procedure. However, if a PCA model for each data set is fitted and the score plots corresponding to the first 2 PCs are shown (second row of Figure 1), differences among the three data sets are made apparent: in the second data set there is one outlier (right side of Figure 1(e)) and in the third data set there are two clusters of observations. As already discussed, the capability to find these details is paramount for data understanding, since outliers and clusters are very informative of the underlaying phenomena. Most of the times these details are also apparent in the original variables, but finding them may be a tedious work. Score plots after PCA modelling are perfectly suited to discover large deviations among the observations, avoiding the overwhelming task of visualizing each possible pair of original variables. Also, score plots in regression models such as PLS are paramount for model interpretation prior to prediction.



(a) First Data Set (b) Second Data Set (c) Third Data Set

(d) First Data Set (e) Second Data Set (f) Third Data Set

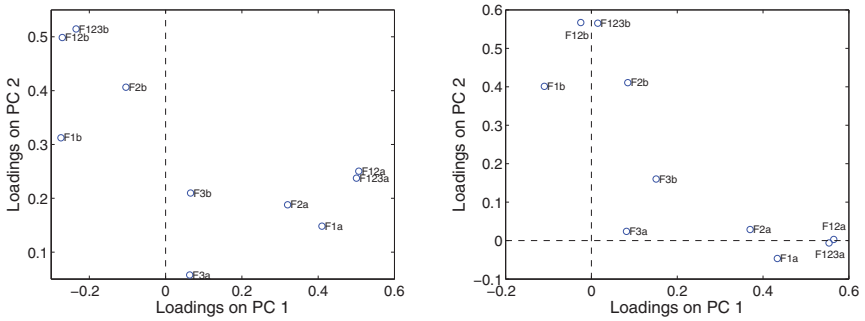
Fig. 1. Experiment with three simulated data sets of dimension 100×100. Data simulation was performed using the ADICOV technique (15). In the first row of figures, the scatter plots corresponding to the first two variables in the data sets are shown. In the second row of figures, the scatter plots (score plots) corresponding to the first two PCs in the data sets are shown.

3. Relationships among variables

PCA has been often employed to explore the relationships among variables in a data set (19; 20). Nevertheless, it is generally accepted that Factor Analysis (FA) is better suited than PCA to study these relationships (1; 7). This is because FA algorithms are designed to distinguish between shared and unique variability. The shared variability, the so-called communalities in the FA community, reflect the common factors–common variability–among observable variables. The unique variability is only present in one observable variable. The common factors make up the relationship structure in the data. PCA makes no distinction between shared and unique variability and therefore it is not suited to find the structure in the data.

When either PCA or FA are used for data understanding, a two step procedure is typically followed (1; 7). Firstly, the model is calibrated from the available data. Secondly, the model is rotated to obtain a so-called simple structure. The second step is aimed at obtaining loading vectors with as much loadings close to 0 as possible. That way, the loading vectors are easier to interpret. It is generally accepted that oblique transformations are preferred to the more simple orthogonal transformations (19; 20), although in many situations the results are similar (1).

The limitation of PCA to detect common factors and the application of rotation methods will be illustrated using the pipelines artificial examples (14). Data for each pipeline are simulated according to the following equalities:



(a) PCA loadings (b) Varimax rotation

Fig. 2. Loading plots of the PCA model fitted from the data in the two pipelines example: (a) original loadings and (b) loadings after varimax rotation.

$$F12 = F1 + F2$$

$$F123 = F1 + F2 + F3$$

where F1, F2 and F3 represent liquid flows which are generated independently at random following a normal distribution of 0 mean and standard deviation 1. A 30% of measurement noise is generated for each of the five variables in a pipeline at random, following a normal distribution of 0 mean:

$$\tilde{x}_i = (x_i + \sqrt{0.3} \cdot n)/(\sqrt{1.3})$$

where $\tilde{x}_i$ is the contaminated variable, x_i the noise-free variable and n the noise generated. This simulation structure generates blocks of five variables with three common factors: the common factor F1, present in the observed variables F1, F12 and F123; the common factor F2, present in the observed variables F2, F12 and F123; and the common factor F3, present in F3 and F123. Data sets of any size can be obtained by combining the variables from different pipelines. In this present example, a data set with 100 observations from two pipelines for which data are independently generated is considered. Thus, the size of the data set is 100×10 and the variability is built from 6 common factors.

Figure 2 shows the loading plots of the PCA model of the data before and after rotation. Loading plots are interpreted so that close variables, provided they are far enough from the origin of coordinates, are considered to be correlated. This interpretation is not always correct. In Figure 2(a), the first component separates the variables corresponding to the two pipelines. The second component captures variance of most variables, specially of those in the second pipeline. The two PCs capture variability corresponding to most common factors in the data at the same time, which complicates the interpretation. As already discussed, PCA is focused on variance, without making the distinction between unique and shared variance. The result is that the common factors are not aligned with the PCs. Thus, one single component reflects several common factors and the same common factor may be reflected in several components. As a consequence, variables with high and similar loadings in the same subset of components do not necessarily need to be correlated, since they may present very different loadings

in others components. Because of this, inspecting only a pair of components may lead to incorrect conclusions. A good interpretation would require inspecting and interrelating all pairs of components with relevant information, something which may be challenging in many situations. This problem affects the interpretation and it is the reason why FA is generally preferred to PCA.

Figure 2(b) shows the resulting loadings after applying one of the most used rotation methods: the varimax rotation. Now, the variables corresponding to each pipeline are grouped towards one of the loading vectors. This highlights the fact that there are two main and orthogonal sources of variability, each one representing the variability in a pipeline. Also, in the first component variables collected from pipeline 2 present low loadings whereas in the second component variables collected from pipeline 1 present low loadings. This is the result of applying the notion of simple structure, with most of the loadings rotated towards 0. The interpretation is simplified as a consequence of improving the alignment of components with common factors. This is especially useful in data sets with many variables.

Although FA and rotation methods may improve the interpretation, they still present severe limitations. The derivation of the structure in the data from a loading plot is not straightforward. On the other hand, the rotated model depends greatly on the normalization of the data and the number of PCs (1; 21). To avoid this, several alternative approaches to rotation have been suggested. The point in common of these approaches is that they find a trade-off between variance explained and model simplicity (1). Nevertheless, imposing a simple structure has also drawbacks. Reference (11) shows that, when simplicity is pursued, there is a potential risk of simplifying even the true relationships in the data set, missing part of the data structure. Thus, the indirect improvement of data interpretation by imposing a simple structure may also report misleading results in certain situations.

3.1 MEDA

MEDA is designed to find the true relationships in the data. Therefore, it is an alternative to rotation methods or in general to the simple structure approach. A main advantage of MEDA is that, unlike rotation or FA methods, it is applied over any projection subspace without actually modifying it. The benefit is twofold. Firstly, MEDA is straightforwardly applied in any subspace of interest: PCA (maximizing variance), PLS (maximizing correlations) and any other. On the contrary, FA methods are typically based on complicated algorithms, several of which have not been extended to regression. Secondly, MEDA is also useful for model interpretation, since common factors and components are easily interrelated. This is quite useful, for instance, in the selection of the number of components.

MEDA is based on the capability of missing values estimation of projection models (22–27). The MEDA approach is depicted in Figure 3. Firstly, a projection model is fitted from the calibration $N \times M$ matrix $\mathbf{X}$ (and optionally $\mathbf{Y}$). Then, for each variable m, matrix $\mathbf{X}_m$ is constructed, which is a $N \times M$ matrix full with zeros except in the m-th column where it contains the m-th column of matrix $\mathbf{X}$. Using $\mathbf{X}_m$ and the model, the scores are estimated with a missing data method. The known data regression (KDR) method (22; 25) is suggested at this point. From the scores, the original data is reconstructed and the estimation error computed. The variability of the estimation error is compared to that of the original data according to the following index of goodness of prediction:

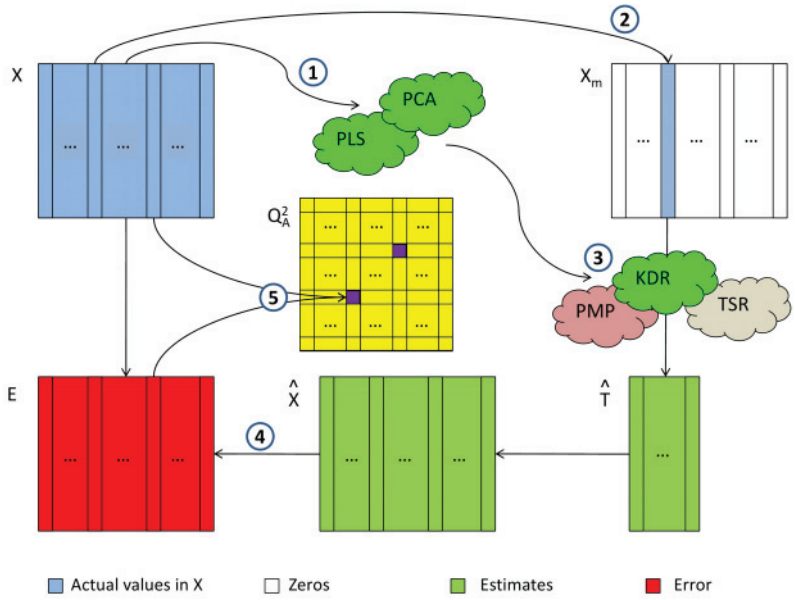


Fig. 3. MEDA technique: (1) model calibration, (2) introduction of missing data, (3) missing data imputation, (4) error computation, (5) computation of matrix $\mathbf{Q}^2_A$.

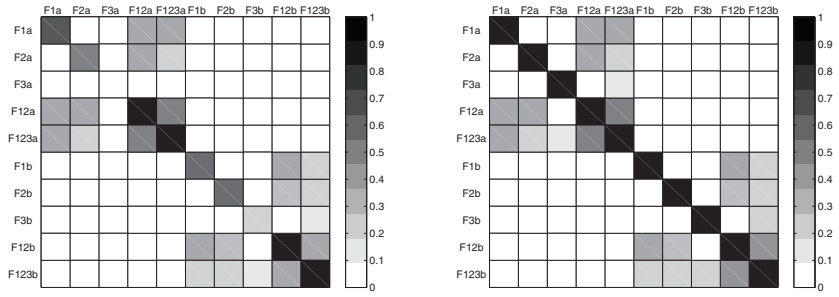
$$q^2_{A,(m,l)} = 1 - \frac{\|\hat{\mathbf{e}}_{A,(l)}\|^2}{\|\mathbf{x}_{(l)}\|^2}, \quad \forall l \neq m. \tag{1}$$

where $\hat{\mathbf{e}}_{A,(l)}$ corresponds to the estimation error for the l-th variable and $\mathbf{x}_{(l)}$ is its actual value. The closer the value of the index is to 1, the more related variables m and l are. After all the indices corresponding to each pair of variables are computed, matrix $\mathbf{Q}^2_A$ is formed so that $q^2_{A,(m,l)}$ is located at row m and column l. For interpretation, when the number of variables is large, matrix $\mathbf{Q}^2_A$ can be shown as a color map. Also, a threshold can be applied to $\mathbf{Q}^2_A$ so that elements over this threshold are set to 1 and elements below the threshold are set to 0.

The procedure depicted in Figure 3 is the original and more general MEDA algorithm. Nevertheless, provided KDR is the missing data estimation technique, matrix Q^2_A can be computed from cross-product matrices following a more direct procedure. The value corresponding to the element in the i-th row and j-th column of matrix $\mathbf{Q}^2_A$ in MEDA is equal to:

$$q^2_{A,(m,l)} = \frac{2 \cdot S_{ml} \cdot S_{ml^A} - (S_{ml^A})^2}{S_{mm} \cdot S_{ll}}. \tag{2}$$

where S_{lm} stands for the cross-product of variables $\mathbf{x}_l$ and $\mathbf{x}_m$, i.e. $S_{lm} = \mathbf{x}_l^T \cdot \mathbf{x}_m$, and S_{lm^A} stands for the cross-product of variables $\mathbf{x}_l$ and $\mathbf{x}^A_m$, being $\mathbf{x}^A_m$ the projection of $\mathbf{x}_m$ in the model sub-space in coordinates of the original space. Thus, S_{lm} is the element in the l-th row



(a) 2 PCs (b) 6 PCs

Fig. 4. MEDA matrix from the PCA model fitted from the data in the two pipelines example.

and m-th column of the cross-product matrix $\mathbf{XX} = \mathbf{X}^T \cdot \mathbf{X}$ and S_{lm^A} corresponds to the element in the l-th row and m-th column of matrix $\mathbf{XX} \cdot \mathbf{P}_A \cdot \mathbf{P}_A^t$ in PCA and of matrix $\mathbf{XX} \cdot \mathbf{R}_A \cdot \mathbf{P}_A^t$ in PLS, with:

$$\mathbf{R}_A = \mathbf{W}_A \cdot (\mathbf{P}_A^T \cdot \mathbf{W}_A)^{-1}. \tag{3}$$

The relationship of the MEDA algorithm and cross-product matrices was firstly pointed out by Arteaga (28) and it can also be derived from the original MEDA paper (11). Equation (2) represents a direct and fast procedure to compute MEDA, similar in nature to the algorithms for model fitting from cross-product matrices, namely the eigendecomposition (ED) for PCA and the kernel algorithms (29) (30) (31) for PLS.

In Figure 4(a), the MEDA matrix corresponding to the 2 PCs PCA model of the example in the previous section, the two independent pipelines, is shown. The structure in the data is elucidated from this matrix. The separation between the two pipelines is shown in the fact that upper-right and lower-left quadrants are close to zero. The relationship among variables corresponding to factors F1 and F2 are also apparent in both pipelines. Since the variability corresponding to factors F3 is barely captured by the first 2 PCs, these are not reflected in the matrix. Nevertheless, if 6 PCs are selected, (Figure 4(b)) the complete structure in the data is clearly found.

MEDA improves the interpretation of both the data set and the model fitted without actually pursuing a simple structure. The result is that MEDA has better properties than rotation methods: it is more accurate and its performance is not deteriorated when the number of PCs is overestimated. Also, the output of MEDA does not depend on the normalization of the loadings, like rotated models do, an it is not limited to subspaces with two or three components at most. A comparison of MEDA with rotation methods is out of the scope of this chapter. Please refer to (11) for it and also for a more algorithmic description of MEDA.

3.2 Loading plots and MEDA

The limitations of loading plots and the application of MEDA were introduced with the pipelines artificial data set. This is further illustrated in this section with two examples provided with the PLS-toolbox (32): the Wine data set, which is used in the documentation of the cited software to show the capability of PCA for improving data understanding, and the

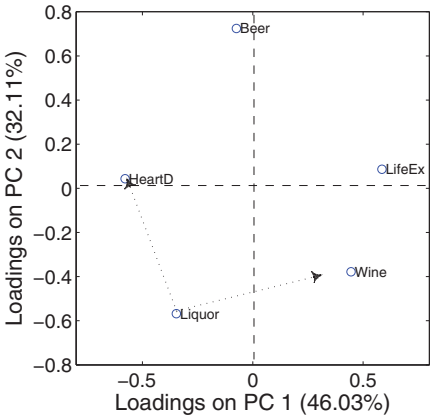


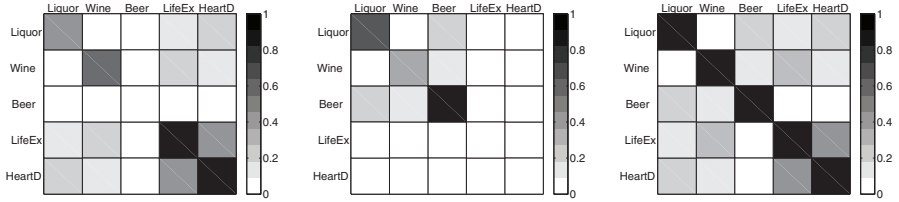
Fig. 5. Loading plot of the first 2 PCs from the Wine Data set provided with the PLS-toolbox (32).

PLSdata data set, which is used to introduce regression models, including PLS. The reading of the analysis and discussion of both data sets in (32) is recommended.

As suggested in (32), two PCs are selected for the PCA model of the Wine data set. The corresponding loading plot is shown in Figure 5. According to the reference, this plot shows that variables HeartD and LifeEx are negatively correlated, being this correlation captured in the first component. Also, "wine is somewhat positively correlated with life expectancy, likewise, liquor is somewhat positively correlated with heart disease". Finally, bear, wine and liquor form a triangle in the figure, which "suggests that countries tend to trade one of these vices for others, but the sum of the three tends to remain constant". Notice that although these conclusions are interesting, some of them are not so evidently shown by the plot. For instance, Liquor is almost as close to HeartD than to Wine. Is Liquor correlated to Wine as it is to HeartD?

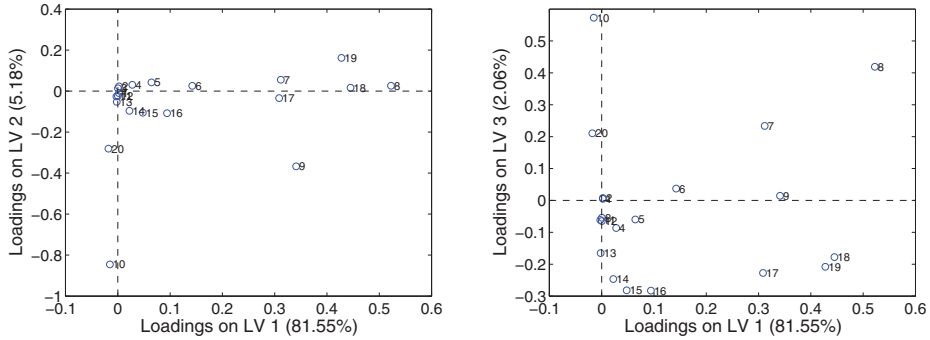
MEDA can be used to improve the interpretation of loading plots. In Figure 6(a), the MEDA matrix for the first PC is shown. It confirms the negative correlation between HeartD and LifeEx, and the–lower–positive correlation between HeartD and Liqour and LifeEx and Wine. Notice that these three relationships are three different common factors. Nevertheless, they all manifest in the same component, making the interpretation with loading plots more complex. The MEDA matrix for the second PC in Figure 6(b) shows the relationship between the three types of drinks. The fact that the second PC captures this relationship was not clear in the loading plot. Furthermore, the MEDA matrix shows that Wine and Liquor are not correlated, answering to the question in the previous paragraph. Finally, this absence of correlation refutes that countries tend to trade wine for liquor or viceversa, although this effect may be true for bear.

In the PLSdata data set, the aim is to obtain a regression model that relates 20 temperatures measured in a Slurry-Fed Ceramic Melter (SFCM) with the level of molten glass. The x-block contains 20 variables which correspond to temperatures collected in two vertical thermowells. Variables 1 to 10 are taken from the bottom to the top in thermowell 1, and variables 11 to 20 from the bottom to the top in thermowell 2. The data set includes 300 training observations

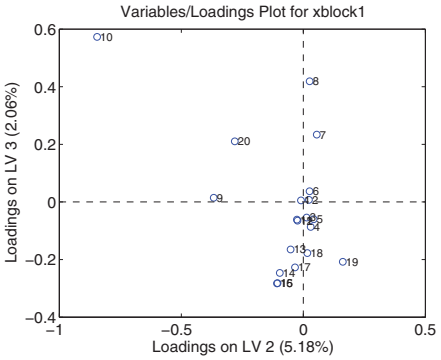


(a) First PC (b) Second PC (c) First 2 PCs

Fig. 6. MEDA matrices of the first PCs from the Wine data set provided with the PLS-toolbox (32).



(a) LV 1 vs LV2 (b) LV1 vs LV3

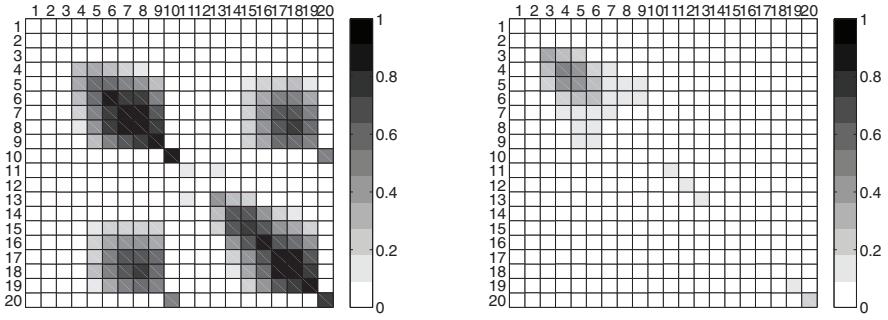


(c) LV2 vs LV3

Fig. 7. Loading plots from the PLS model in the Slurry-Fed Ceramic Melter data set.

and 200 test observations. This same data set was used to illustrate MEDA with PCA in (11) with the temperatures and the level of molten glass together in the same block of data [1].

[1] There are some results of the analysis in (11) which are not coherent with those reported here, since the data sets used do not contain the same observations.



(a) First three LVs (b) Fourth LV

Fig. 8. MEDA matrices from the PLS model in the Slurry-Fed Ceramic Melter data set.

Following recommendations in (32), a 3 LVs PLS model from a mean-centered x-block was fitted. The loading plots corresponding to the three possible pairs of LVs are shown in Figure 7. The first LV captures the predictive variance in the temperatures, with higher loadings for higher indices of the temperatures in each thermowell. This holds exception made on temperatures 10 and 20, which present their predictive variance in the second and third LVs, being the variance in 10 between three and four times higher than that in 20. The third LVs also seems to discriminate between both thermowells. For instance, in Figure 7(c) most temperatures of the first thermowell are in the upper middle of the Figure, whereas most temperatures of the second thermowell are in the lower middle. Nevertheless, it should be noted that the third LV only captures 2% of the variability in the x-block and 1% of the variability in the y-block.

The information gained with the loading plots can be complemented with that in MEDA, which brings a much clearer picture of the structure in the data. The MEDA matrix for the PLS model with 3 LVs is shown in Figure 8(a). There is a clear auto-correlation effect in the temperatures, so that closer sensors are more correlated. This holds exception made on temperatures 10 and 20. Also, the corresponding temperatures in both thermowells are correlated, including 10 and 20. Finally, the temperatures at the bottom do not contain almost any predictive information of the level of molten glass. In (32), the predictive error by cross-validation is used to identify the number of LVs. Four LVs attain the minimum predictive error, but 3 LVs are selected since the fourth LV does not contribute much to the reduction of this error. In Figure 8(b), the contribution of this fourth LV is shown. It is capturing the predictive variability from the third to the fifth temperature sensors in the first thermowell, which are correlated. The corresponding variability in the second thermowell is already captured by the first 3 LVs. This is an example of the capability of MEDA for model interpretation, which can be very useful in the determination of the number of LVs. In this case and depending on the application of the model, the fourth LV may be added to the model in order to compensate the variance captured in both thermowells, even if the improvement in prediction performance is not high.

The information provided by MEDA can also be useful for variable selection. In this example, temperatures 1-2 and 11-12 do not contribute to a relevant degree to the regression model. As shown in Table 1, if those variables are not used in the model, its prediction performance

Variables	Complete model	[3 : 10 13 : 20]	[3 : 10]	[13 : 20]
LVs	3	3	3	3
X-block variance	88.79	89.84	96.36	96.93
Y-block variance	87.89	87.78	84.61	83.76
RMSEC	0.1035	0.1039	0.1166	0.1198
RMSECV	0.1098	0.1098	0.1253	0.1271
RMSEP	0.1396	0.1394	0.1522	0.1631

Table 1. Comparison of three PLS models in the Slurry-Fed Ceramic Melter data set. The variance in both blocks of data and the Root Mean Square Error of Calibration (RMSEC), Cross-validation (RMSECV) and Prediction (RMSEP) are compared.

remains the same. Also, considering the correlation among thermowells, one may be tempted to use only one of the thermowells for prediction, reducing the associated costs of maintaining two thermowells. If this is done, only 8 predictor variables are used and the prediction performance is reduced, but not to a large extent. Correlated variables in a prediction model help to better discriminate between true structure and noise. For instance, in this example, when only the sensors of one thermowell are used, the PLS model captures more x-block variance and less y-block variance. This is showing that more specific–noisy–variance in the x-block is being captured. Using both thermowells reduces this effect. Another example of variable selection with MEDA will be presented in Section 5.

3.3 correlation matrices and MEDA

There is a close similarity between MEDA and correlation matrices. To this regard, equation (2) simplifies the interpretation of the MEDA procedure. The MEDA index combines the original variance with the model subspace variance in S_{ml} and S_{ml^A}. Also, the denominator of the index in eq. (2) is the original variance. Thus, those pairs of variables where a high amount of the total variance of one of them can be recovered from the other are highlighted. This is convenient for data interpretation, since only factors of high variance are highlighted. On the other hand, it is easy to see that when the number of LVs, A, equals the rank of $\mathbf{X}$, then Q^2_A is equal to the element-wise squared correlation matrix of $\mathbf{X}$, C^2 (11). This can be observed in the following element-wise equality:

$$q^2_{Rank(X),(m,l)} = \frac{S^2_{ml}}{S_{mm} \cdot S_{ll}} = c^2_{(m,l)}. \tag{4}$$

This equivalence shows that matrix Q^2_A has a similar structure than the–element-wise squared–correlation matrix. To elaborate this similarity, a correlation matrix can be easily extended to the notion of latent subspace. The correlation matrix in the latent subspace, C_A, can de defined as the correlation matrix of the reconstruction of $\mathbf{X}$ with the first A LVs. Thus, $C_A = \mathbf{P}_A \cdot \mathbf{P}^t_A \cdot \mathbf{C} \cdot \mathbf{P}_A \cdot \mathbf{P}^t_A$ in PCA and $C_A = \mathbf{P}_A \cdot \mathbf{R}^t_A \cdot \mathbf{C} \cdot \mathbf{R}_A \cdot \mathbf{P}^t_A$ in PLS. If the elements of C_A are then squared, the element-wise squared correlation in the latent subspace, noted as C^2_A, is obtained. Strictly speaking, each element of C^2_A is defined as:

$$c^2_{A,(m,l)} = \frac{S^2_{m^A l^A}}{S_{m^A m^A} \cdot S_{l^A l^A}}. \tag{5}$$

However, for the same reason explained before, if C_A^2 is aimed at data interpretation, the denominator should be original variance:

$$c^2_{A,(m,l)} = \frac{S^2_{m^A l^A}}{S_{mm} \cdot S_{ll}}. \tag{6}$$

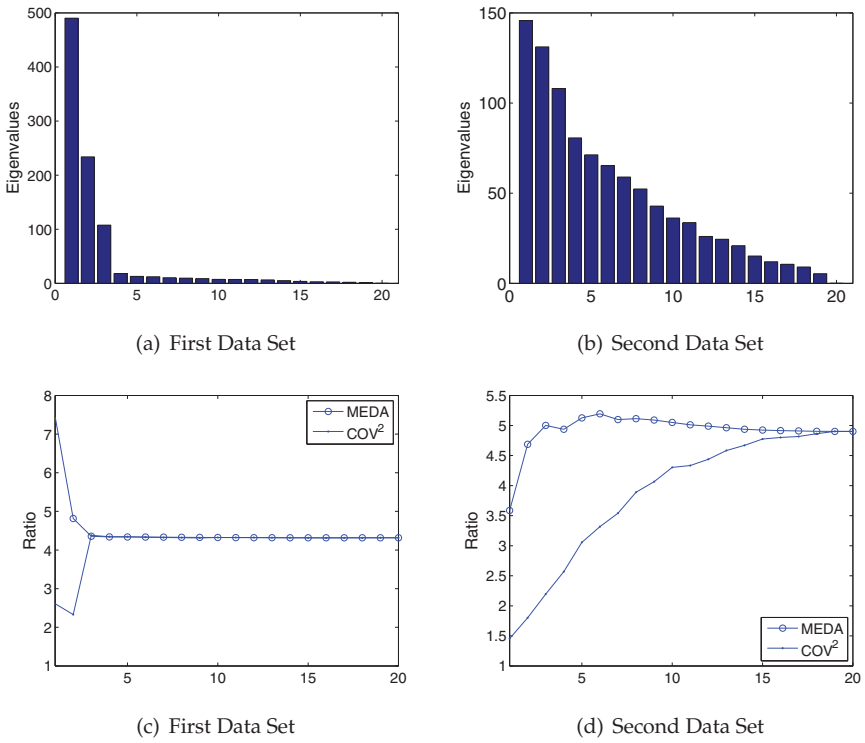
If this equation is compared to equation (2), we can see that the main difference between MEDA and the–projected and element-wise squared–correlation matrix is the combination of original and projected variance in the numerator of the former. This combination is paramount for interpretation. Figure 9 illustrates this. The example of the pipelines is used again, but in this case ten pipelines and only 20 observations are considered, yielding a dimension of 20×50 in the data. Two data sets are simulated. In the first one, the pipelines are correlated. As a consequence, the data present three common factors represented by the three biggest eigenvalues in Figure 9(a). In the second one, each pipeline is independently generated, yielding a more distributed variance in the eigenvalues (Figure 9(b)). For matrices Q_A^2 and C_A^2 to infer the structure in the data, they should have large values in the elements which represent real structural information (common factors) and low values in the rest of the elements. Since in both data sets it is known a-priori which elements in the matrices represent actual common factors and which not, the mean values for the two groups of elements in matrices Q_A^2 and C_A^2 can be computed. The ratio of these means, computed by dividing the mean of the elements with common factors by the mean of the elements without common factors, is a measure of the discrimination capability between structure and noise of each matrix. The higher this index is, the better the discrimination capability is. This ratio is shown in Figures 9(c) and 9(d) for different numbers of PCs. Q_A^2 outperforms C_A^2 until all relevant eigenvalues are incorporated to the model. Also, Q_A^2 presents maximum discrimination capability for a reduced number of components. Notice that both alternative definitions of C_A^2 in equations (5) and (6) give exactly the same result, though equation (6) is preferred for visual interpretation.

4. Connection between observations and variables

The most relevant issue for data understanding is probably the connection between observations and variables. It is almost useless to detect certain details in the data distribution, such as outliers or clusters, if the set of variables related to these details are not identified. Traditionally, biplots (12) have been used for this purpose. In biplots, the scatter plots of loadings and scores are combined in a single plot. Apart from relevant considerations regarding the comparability of the axes in the plot, which is also important for any scatter plots, and of the scales in scores and loadings (18), biplots may be misleading just because of the loading plot included. In this point, a variant of MEDA, named observation-based MEDA or *o*MEDA, can be used to unveil the connection between observations and variables without the limitations of biplots.

4.1 *o*MEDA

*o*MEDA is a variant of MEDA to connect observations and variables. Basically, *o*MEDA is a MEDA algorithm applied over a combination of the original data and a dummy variable designed to cover the observations of interest. Take the following example: a number of subsets of observations $\{\mathbf{C}_1, ..., \mathbf{C}_N\}$ form different clusters in the scores plot which are located



(a) First Data Set (b) Second Data Set

(c) First Data Set (d) Second Data Set

Fig. 9. Comparison of MEDA and the projected and element-wise squared covariance matrix in the identification of the data structure from the PCA model fitted from the data in the ten pipelines example: (a) and (b) show the eigenvalues when the pipelines are correlated and independent, respectively, and (c) and (d) show the ratio between the mean of the elements with common factors and the mean of the elements without common factors in the matrices.

far from the bulk of the data, **L**. One may be interested in identifying, for instance, the variables related to the deviation of $\mathbf{C}_1$ from **L** without considering the rest of clusters. For that, a dummy variable **d** is created so that observations in $\mathbf{C}_1$ are set to 1, observations in **L** are set to -1, while the remaining observations are left to 0. Also, values other than 1 and -1 can be included in the dummy variable if desired. *o*MEDA is then performed using this dummy variable.

The *o*MEDA technique is illustrated in Figure 10. Firstly, the dummy variable is designed and combined with the data set. Then, a MEDA run is performed by predicting the original variables from the dummy variable. The result is a single vector, $\mathbf{d}_A^2$, of dimension $M \times 1$, being M the number of original variables. In practice, the *o*MEDA index is slightly different to that used in MEDA. Being **d** the dummy variable, designed to compare a set of observations with value 1 (or in general positive values) with another set with value -1 (or in general negative values), then the *o*MEDA index follows:

$$d^2_{A,(l)} = \|\mathbf{x}^d_{(l)}\|^2 - \|\hat{\mathbf{e}}^d_{A,(l)}\|^2, \quad \forall l. \tag{7}$$

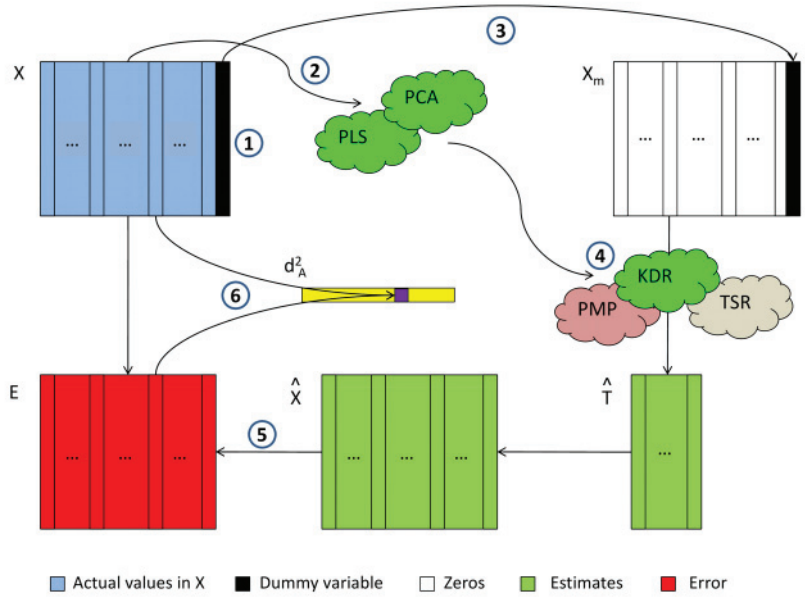


Fig. 10. *o*MEDA technique: (1) introduction of the dummy variable, (2) model calibration, (3) introduction of missing data, (4) missing data imputation, (5) error computation, (6) computation of vector $\mathbf{d}^2_A$.

where $\mathbf{x}^d_{(l)}$ represents the values of the l-th variable in the original observations different to 0 in $\mathbf{d}$ and $\hat{\mathbf{e}}^d_{A,(l)}$ is the corresponding estimation error. The main difference between the computation of index $d^2_{A,(l)}$ in *o*MEDA and that of MEDA is the absence of the denominator in the former. This modification is convenient to avoid high values in $d^2_{A,(l)}$ when the amount of variance of a variable in the reduced set of observations of interest is very low. Once $\mathbf{d}^2_A$ is computed for a given dummy variable, sign information can be added from the mean vectors of the two groups of observations considered (33).

In practice, in order to avoid any modification in the PCA or PLS subspace due to the introduction of the dummy variable, the *o*MEDA algorithm is slightly more complex than the procedure shown in Figure 10. For a description of this algorithm refer to (33). However, like in MEDA, the *o*MEDA vector can be computed in a more direct way by assuming KDR (26) is used as the missing data estimation procedure. If this holds, the *o*MEDA vector follows:

$$d^2_{A,(l)} = 2 \cdot \mathbf{x}^t_{(l)} \cdot \mathbf{D} \cdot \mathbf{x}_{A,(l)} - \mathbf{x}^t_{A,(l)} \cdot \mathbf{D} \cdot \mathbf{x}_{A,(l)}, \tag{8}$$

where $\mathbf{x}_{(l)}$ represents the l-th variable in the–complete–set of original observations and $\mathbf{x}_{A,(l)}$ its projection in the latent subspace in coordinates of the original space and:

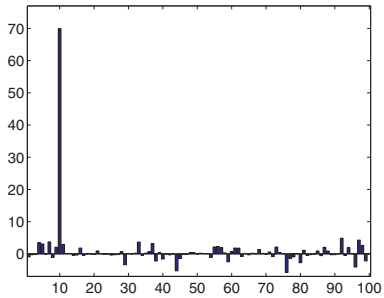


Fig. 11. *o*MEDA vector of two clusters of data from the 10 PCs PCA model of a simulated data set of dimension 100×100. This model captures 30% of the variability. Data present two clusters in variable 10.

$$\mathbf{D} = \frac{\mathbf{d} \cdot (\mathbf{d})^T}{\|\mathbf{d}\|^2}. \tag{9}$$

Finally, the equation can also be reexpressed as follows:

$$d^2_{A,(l)} = \frac{1}{N} \cdot (2 \cdot \Sigma^d_{(l)} - \Sigma^d_{A,(l)}) \cdot |\Sigma^d_{A,(l)}| \tag{10}$$

with $\Sigma^d_{(l)}$ and $\Sigma^d_{A,(l)}$ being the weighted sum of elements in $\mathbf{x}(l)$ and $\mathbf{x}_{A,(l)}$ according to the weights in $\mathbf{d}$, respectively. Equation (10) has two advantages. Firstly, it presents the *o*MEDA vector as a weighted sum of values, which is easier to understand. Secondly, it has the sign computation built in, due to the absolute value in the last element. Notice also that *o*MEDA inherits the combination of total and projected variance present in MEDA.

In Figure 11 an example of *o*MEDA is shown. For this, a 100×100 data set with two clusters of data was simulated. The distribution of the observations was designed so that both clusters had significantly different values only in variable 10 and then data was auto-scaled. The *o*MEDA vector clearly highlights variable 10 as the main difference between both clusters.

4.2 Biplots vs oMEDA

Let us return to the discussion regarding the relationship between the common factors and the components. As already commented, several common factors can be captured by the same component in a projection model. As a result, a group of variables may be located close in a loading plot without the need to be correlated. This is also true for the observations. Thus, two observations closely located in a score plot may be quite similar or quite different depending on their scores in the remaining LVs. However, score plots are typically employed to observe a general distribution of the observations. This exploration is more aimed at finding differences among observations rather than similarities. Because of this, the problem described for loading plots is not so relevant for the typical use of score plots. However, this is a problem when interpreting biplots. In biplots, deviations in the observations are related to deviations in the variables. Like loading plots, biplots may be useful to perform a fast view on the data, but any conclusion should be confirmed with another technique. *o*MEDA is perfectly suited for this.

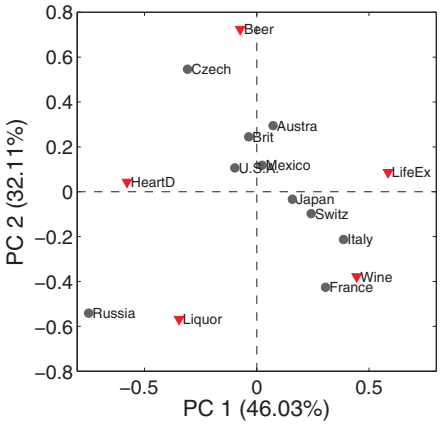


Fig. 12. Biplot of the first 2 PCs from the Wine Data set provided with the PLS-toolbox (32).

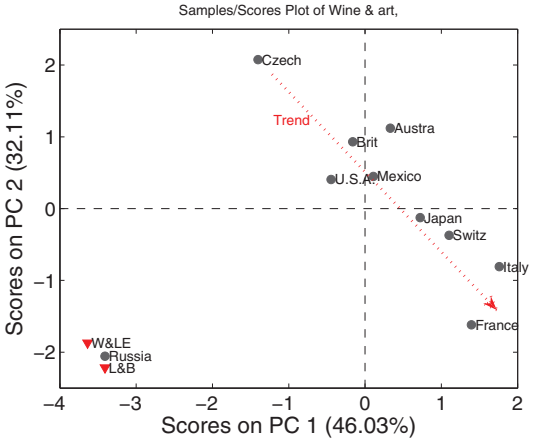
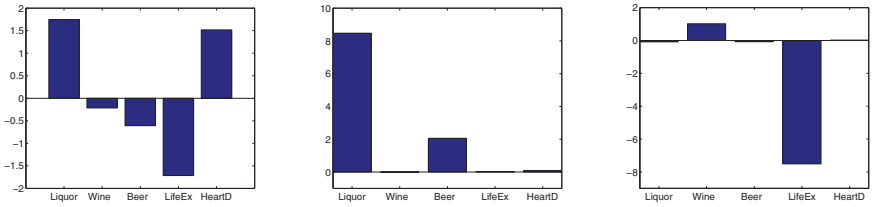


Fig. 13. Score plot of the first 2 PCs from the Wine Data set provided with the PLS-toolbox (32) and two artificial observations.

In Figure 12, the biplot of the Wine data set is presented. The distribution of the scores show that all countries but Russia follow a trend from the Czech Republic to France, while Russia is located far from this trend. The biplot may be useful to make hypothesis on the variables related to the special nature of Russia or to the trend in the rest of countries. Nevertheless, this hypothesis making is not straightforward. To illustrate this, in Figure 13 the scores are shown together with two artificial observations. The artificial observations were designed to lay close to Russia in the score plot of the first two PCs. In both cases, three of the five variables were left to their average value in the Wine data set and only two variables are set to approach Russia. Thus, observation W&LE only uses variables Wine and LifeEx to yield a point close to Russia in the score plot while the other variables are set to the average. Observation L&B only uses Liquor and Beer. With this example, it can be concluded that very little can be said about Russia only by looking at the biplot.



(a) Russia Vs The rest (b) L&B Vs The rest (c) W&LE Vs The rest

Fig. 14. oMEDA vectors of the first 2 PCs from the Wine Data set provided with the PLS-toolbox (32). Russia (a), L&B (b) and W&LE (c) compared to the rest of countries.

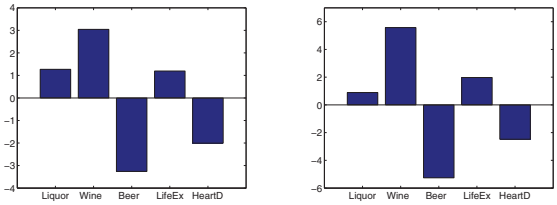
In Figure 14(a), the oMEDA vector to discover the differences between Russia and the rest of countries is shown. For this, a dummy variable is built where all observations except Russia are set to -1 and Russia is set to 1. oMEDA shows that Russia has in general less life expectancy and more heart disease and liquor consumption than the rest of countries. The same experiment is repeated for artificial observations L&B and W&LE in Figures 14(b) and 14(c). oMEDA clearly distinguishes among the three observations, while in the biplot they seem to be very similar.

To analyze the trend shown by all countries except Russia in Figure 12, the simplest approach is to compare the most separated observations, in this case France and the Czech Republic. The oMEDA vector is shown in Figure 15(a). In this case, the dummy variable is built so that France has value 1, the Czech Republic has value -1 and the rest of the countries have 0 value. Thus, positive values in the oMEDA vector identify variables with higher value in France than in the Czech Republic and negative values the opposite. oMEDA shows that the French consume more wine and less beer than Czech people. Also, according to the data, the former seem to be more healthy.

Comparing the two most separated observations may be misleading in certain situations. Another choice is to use the capability of oMEDA to unveil the variables related to any direction in a score plot. For instance, let us analyze the trend of the countries incorporating the information in all the countries. For this, different weights are considered in the dummy variable. We can think of these weights as-approximate–projections of the observations in the direction of interest. Following this approach, the weights listed in Table 2 are assigned, which approximate the projection of the countries in the imaginary line depicted by the arrow in Figure 13. Since Russia is not in the trend, it is left to 0. Using these weights, the resulting oMEDA vector is shown in Figure 15(b). In this case, the analysis of the complete set of observations in the trend resembles the conclusions in the analysis of the two most separated observations.

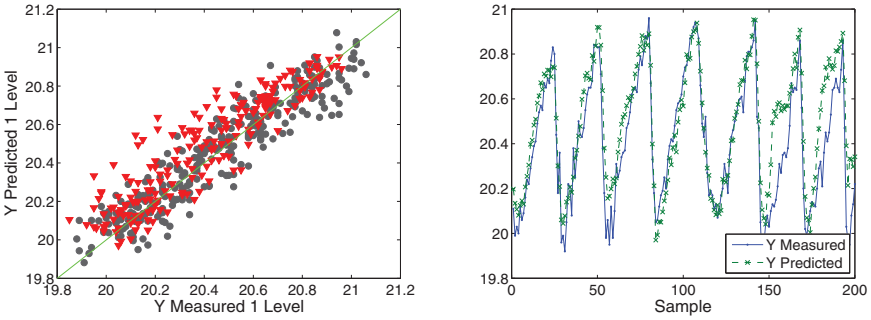
Country	**Weight**	**Country**	**Weight**
France	3	Mexico	-1
Italy	2	U.S.A.	-1
Switz	1	Austra	-1
Japan	1	Brit	-1
Russia	0	Czech	-3

Table 2. Weights used in the dummy variable for the oMEDA vector in Figure 15(b).



(a) France Vs Czech (b) All except Russia

Fig. 15. oMEDA vectors of the first 2 PCs from the Wine Data set provided with the PLS-toolbox (32). In (a), France and Czech Republic are compared. In (b), the trend shown in the score plot by all countries except Russia is analyzed.



(a) Training & Test (b) Test

Fig. 16. Measured vs predicted values of molten glass level in the PLS model with 3 LVs fitted from the Slurry-Fed Ceramic Melter (SFCM) data set in (32).

Let us return to the PLSdata data set. A PLS model relating temperatures with the level of molten glass was previously fitted. As already discussed, the data set includes 300 training observations and 200 test observations. The measured and predicted values of both sets of observations are compared in Figure 16(a). The predicted values in the test observations (inverted triangles) tend to be higher than true values. This is also observed in Figure 16(b). The cause for this seems to be that the process has slightly moved from the operation point where training data was collected. oMEDA can be used to identify this change of operation point by simply comparing training and test observations in the model subspace. Thus, training (value 1) and test observations (value -1) are compared in the subspace spanned by the first 3 LVs of the PLS model fitted only from training data. The resulting oMEDA vector is shown in Figure 17. According to the result, considering the test observations have value -1 in the dummy variable, it can be concluded that the process has moved to a situation in which top temperatures are higher than during model calibration.

5. Case study: Selwood data set

In this section, an exploratory data analysis of the Selwood data set (34) is carried out. The data set was downloaded from http://michem.disat.unimib.it/chm/download/datasets.htm. It

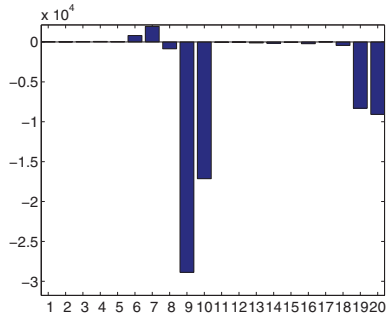


Fig. 17. oMEDA vector comparing training and test observations in the PLS model with 3 LVs fitted from the Slurry-Fed Ceramic Melter (SFCM) data set in (32).

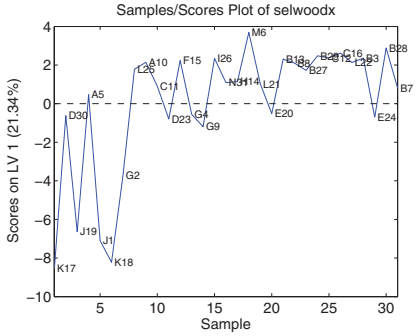
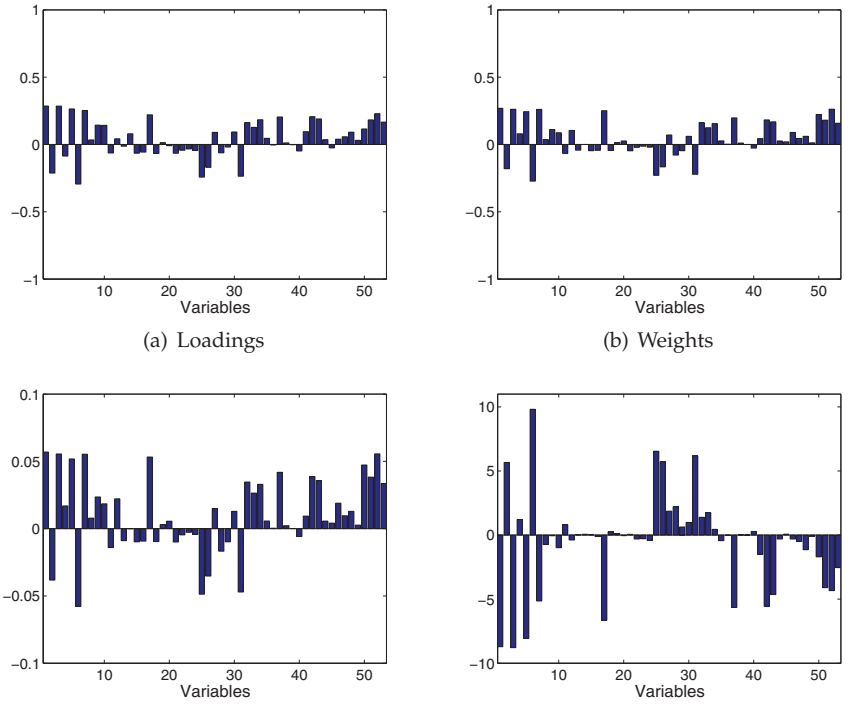


Fig. 18. Scores corresponding to the first LV in the PLS model with the complete set of descriptors of the Selwood dataset.

consists of 31 antifilarial antimycin A_1 analogues for which 53 physicochemical descriptors were calculated for Quantitative Structure-Activity Relationship (QSAR) modelling. The set of descriptors is listed in Table 3. These descriptors are used for predicting in vitro antifilarial activity (-LOGEC50). This data set has been employed for testing variables selection methods, for instance in (35; 36), in order to find a reduced number of descriptors with best prediction performance. Generally speaking, these variable selection methods are based on complex optimization algorithms which make use of heuristics to reduce the search space.

Indices	**Descriptors**
1:10	ATCH1 ATCH2 ATCH3 ATCH4 ATCH5 ATCH6 ATCH7 ATCH8 ATCH9 ATCH10
11:20	DIPV_X DIPV_Y DIPV_Z DIPMOM ESDL1 ESDL2 ESDL3 ESDL4 ESDL5 ESDL6
21:30	ESDL7 ESDL8 ESDL9 ESDL10 NSDL1 NSDL2 NSDL3 NSDL4 NSDL5 NSDL6
31:40	NSDL7 NSDL8 NSDL9 NSDL10 VDWVOL SURF_A MOFI_X MOFI_Y MOFI_Z PEAX_X
41:50	PEAX_Y PEAX_Z MOL_WT S8_1DX S8_1DY S8_1DZ S8_1CX S8_1CY S8_1CZ LOGP
51:53	M_PNT SUM_F SUM_R

Table 3. Physicochemical descriptors of the Selwood dataset.

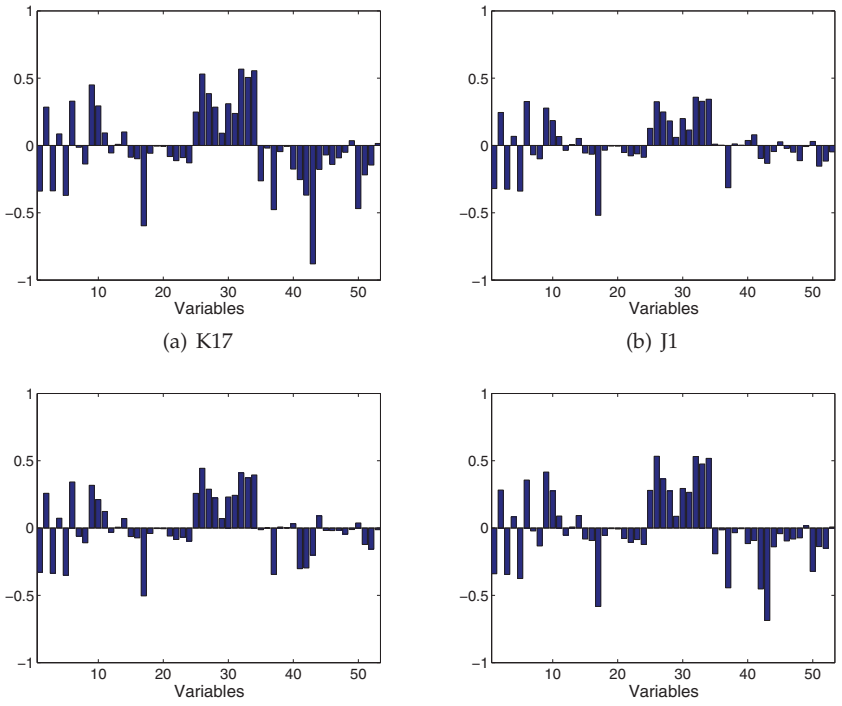


(a) Loadings (b) Weights

(c) Regression Coefficients (d) *o*Meda (K17, J1, J19 and K18 vs the rest)

Fig. 19. Several vectors corresponding to the first LV in the PLS model with the complete set of descriptors of the Selwood dataset.

First of all, a PLS model is calibrated between the complete set of descriptors and -LOGEC50. Leave-one-out cross-validation suggests one LV. The score plot corresponding to the 31 analogues in that LV are shown in Figure 18. Four of the compounds, namely K17, J1, J19 and K18, present an abnormally low score. This deviation is highly contributing to the variance in the first LV and the reason for it should be investigated. Two of these compounds, K17 and K18, were catalogued as outliers by (34), where the authors stated that "Chemically, these compounds are distinct from the bulk of the training set in that they have an n-alkyl side chain as opposed to a side chain of the phenoxy ether type". Since the four compounds present an abnormally low score in the first LV, typically the analyst may interpret the coefficients of that LV to try to explain this abnormality. In Figure 19, the loadings, weights and regression coefficients of the PLS model are presented together with the *o*MEDA vector. The latter identifies those variables related to the deviation of the four compounds from the rest. The *o*MEDA vector is similar, but with opposite sign, to the other vectors in several descriptors, but quite different in others. Therefore, the loadings, weights or coefficient vectors should not be used in this case for the investigation of the deviation, or otherwise one may arrive to incorrect conclusions. On the other hand, it may be worth to check whether the *o*MEDA vector is representative of the deviation in the four compounds. Performing *o*MEDA individually on each of the compounds confirm this fact (see Figure 20)



(a) K17 (b) J1

(c) J19 (d) K18

Fig. 20. *o*MEDA vectors corresponding to the first LV in the PLS model with the complete set of descriptors of the Selwood dataset. To compute each of the vectors, one of the four compounds K17, J1, J19 and K18 are set to 1 in the dummy variable and the other three are set to 0, while the rest of compounds in the data set are set to -1.

The subsequent step is to search for relevant descriptors (variable selection) For this, MEDA will be employed. In this point, there are two choices. The four compounds with low score in the first LV may be treated as outliers and separated from the rest of the data (34) or the complete data set may be modelled with a single QSAR model (35; 36). It should be noted that differences among observations in one model may not be found in a different model, so that the same observation may be an outlier or a normal observation depending on the model. Furthermore, as discussed in (35), the more general the QSAR model is, so that it models a wider set of compounds, the better. Therefore, the complete set of compounds will be considered in the remaining of the example. On the other hand, regarding the analysis tools used, there are different possibilities. MEDA may be applied over the PLS model relating the descriptors in the x-block and -LOGEC50 in the y-block. Alternatively, both blocks may be joined together in a single block of data and MEDA with PCA be applied. The second choice will be generally preferred to avoid over-fitting, but typically both approaches may lead to the same conclusions, like it happens in the present example.

The application of MEDA requires to select the number of PCs in the PCA model. Considering that the aim is to understand how the variability in -LOGEC50 is related to the descriptors in

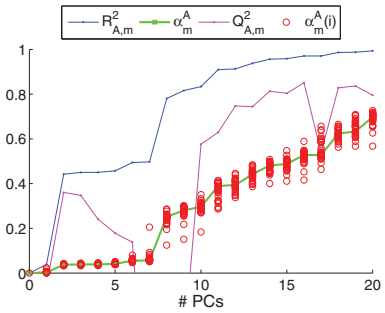


Fig. 21. Structural and Variance Information (SVI) plot of in vitro antifilarial activity (-LOGEC50). The data set considered combines -LOGEC50 with the complete set of descriptors of the Selwood dataset.

the data set, the Structural and Variance Information (SVI) plots are the adequate analysis tool (14). The SVI plots combine variance information with structural information to elucidate how a PCA model captures the variance of a single variable. The SVI plot of a variable v reveals how the following indices evolve with the addition of PCs in the PCA model:

- The R^2 statistic, which measures the variance of v.
- The Q^2 statistic, which measures the performance of the missing data imputation of v, or otherwise stated its prediction performance.
- The α statistic, which measures the portion of variance of v which is identified as unique variance, i.e. variance not shared with other variables.
- The stability of α, as an indicator of the stability of the model calibration.

Figure 21 shows the SVI plot of -LOGEC50 in the PCA model with the complete set of descriptors. The plot shows that the model remains quite stable until 5-6 PCs are included. This is seen in the closeness of the circles which represents the different instances of α computed on a leave-one-out cross-validation run. The main portion of variability in -LOGEC50 is captured in the second and eighth PCs. Nevertheless, is not until the tenth PC that the missing data imputation (Q^2) yields a high value. For more PCs, the captured variability is only slightly augmented. Since MEDA makes use of the missing data imputation of a PCA model, Q^2 is a relevant index. At the same time, from equation (2) is clear that MEDA is also influenced by captured variance. Thus, 10 PCs are selected. In any case, it should be noted that MEDA is quite robust to the overestimation in the number of PCs (11) and very similar MEDA matrices are obtained for 3 or more PCs in this example.

The MEDA matrix corresponding to the PCA model with 10 PCs from the data set which combines -LOGEC50 with the complete set of descriptors of the Selwood dataset is presented in Figure 22. For variable selection, the most relevant part of this matrix is the last column (or row), which corresponds to -LOGEC50. This vector is shown in Figure 23(a). Those descriptors with high value in this vector are the ones from which -LOGEC50 can be better predicted. Nevertheless, the selection of, say, the first n variables with higher value is not an adequate strategy because the relationship among the descriptors should also be considered. Let us select the descriptor with better prediction performance, in this case ATCH6, though

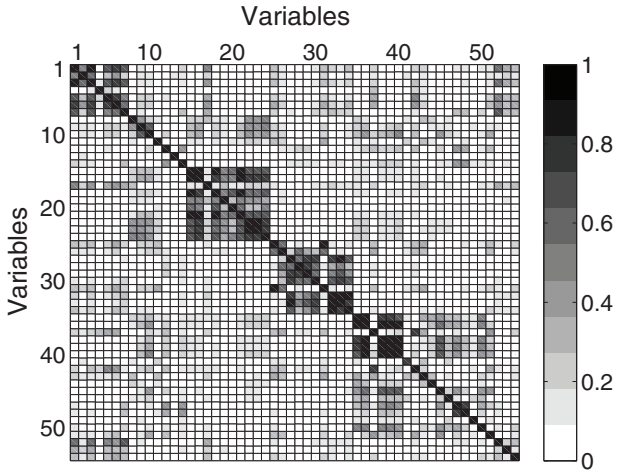
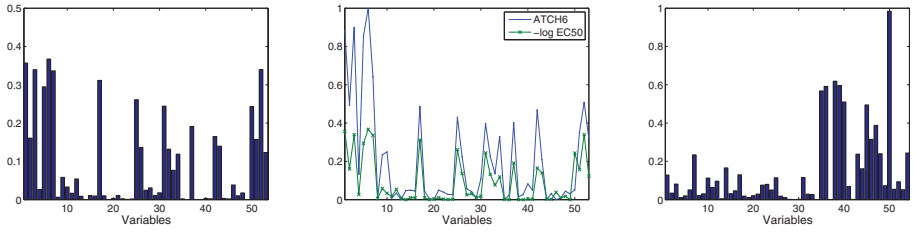


Fig. 22. MEDA matrix of the PCA model with 10 PCs from the data set which combines the in vitro antifilarial activity (-LOGEC50) with the complete set of descriptors of the Selwood dataset.



(a) (b) (c)

Fig. 23. MEDA vector corresponding to the in vitro antifilarial activity (-LOGEC50) (a) comparison between this vector and that corresponding to ATCH6 (b) and MEDA vector corresponding to LOGP (c) in the PCA model with 10 PCs from the data set which combines -LOGEC50 with the complete set of descriptors of the Selwood dataset.

ATCH1, ATCH3, ATCH7 or SUM_F have a very similar prediction performance. The least-squares regression model with ATCH6 as regressor attains a Q^2 equal to 0.30, more than the Q^2 attained by any number of LVs in the PLS model with the complete set of descriptors. If for instance ATCH6 and ATCH1 are used as regressors, $Q^2 = 0.26$ is obtained for least squares regression and $Q^2 = 0.31$ for PLS with 1 LV, which means an almost negligible improvement with the addition of ATCH1. The facts that the improvement is low and that the 1 LV PLS model outperforms the least squares model are caused by the correlation between ATCH6 and ATCH1, correlation clearly pointed out in the MEDA matrix (see the element at the sixth column and first row or the first column and sixth row) Clearly, both ATCH6 and ATCH1 are related to the same common factor in -LOGEC50. However, the variability in -LOGEC50 is the result of several sources of variability, which may be common factors with other descriptors.

Therefore, in order to introduce a new common factor in the model other than that in ATCH6, we need to find a descriptor related to -LOGEC50 but not to ATCH6. Also, the model may be improved by introducing a descriptor related to ATCH6 but not to -LOGEC50. For this, Figure 23(b) compares the columns in the MEDA matrix corresponding to ATCH6 and -LOGEC50. The comparison should not be performed in terms of direct differences between values. For instance, ATCH1 and ATCH6 are much more correlated than ATCH1 and -LOGEC50. It is the difference in shape which is informative. Thus, we find that -LOGEC50 present a high correlation with LOGP (variable 50) which is not found in ATCH6. Thus, LOGP presents a common factor with -LOGEC50 which is not present in ATCH6. Using LOGP and ATCH6 as regressors, the least squares model presents $Q^2 = 0.37$.

If an additional descriptor is to be added to the model, again it should present a different common factor with any of the variables in the model. The MEDA vector corresponding to LOGP is shown in Figure 23(c). This descriptor is related to a number of variables which are not related to -LOGEC50. This relationship represents a common factor in LOGP but not in -LOGEC50. The inclusion of a descriptor containing this common factor, for instance MOFI_Y (variable 38) may improve prediction because it may help to distinguish the portion of variability in LOGP which is useful to predict -LOGEC50 from the portion which is not. Using LOGP, ATCH6 and MOFI_Y as regressors yields $Q^2 = 0.56$, illustrating that the addition of a descriptor which is not related to the predicted variable may be useful for prediction.

In Figure 24, the two common factors described before, the one present in ATCH6 and -LOGEC50 and the one present in LOGP and MOFI_Y, are approximately highlighted in the MEDA matrix. If variables ATCH6 and MOFI_Y are replaced by others with the same common factors, the prediction performance of the model remains similar. However, LOGP is utmost for the model since is the only descriptor which relates the second common factor and -LOGEC50. These results are coherent with findings in the literature. Both (35) and (36) highlight the relevance of LOGP, and justify it with the results in several more publications. Furthermore, the top 10 models found in (35), presented in Table 4, follow the same patter of the solution found here. The models with three descriptors contain LOGP with one descriptor from the first and second common factors. The models with two descriptors contain LOGP and a variable with the second common factor.

Descriptors	Q^2
SUM_F (52) LOGP (50) MOFI_Y (38)	0.647
ESDL3 (17) LOGP (50) SURF_A (36)	0.645
SUM_F (52) LOGP (50) MOFI_Z (39)	0.644
LOGP (50) MOFI_Z (39)	0.534
ESDL3 (17) LOGP (50) MOFI_Y (38)	0.605
ESDL3 (17) LOGP (50) MOFI_Z (39)	0.601
LOGP (50) MOFI_Y (38)	0.524
LOGP (50) PEAX_X (40)	0.518
LOGP (50) SURF_A (36)	0.501
SUM_F (52) LOGP (50) PEAX_X (40)	0.599

Table 4. Top 10 models obtained after variable selection of the Selwood data set in (35)

Finally, in Figure 25 the plot of measured vs predicted values of -LOGEC50 in the model resulting from the exploration is shown. No outliers are identified, though the four

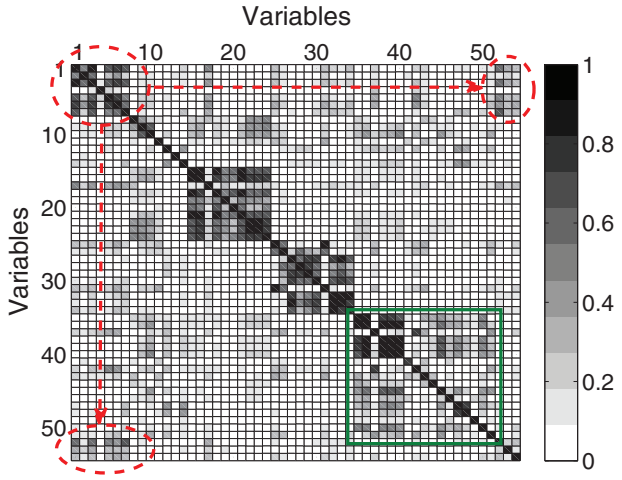


Fig. 24. MEDA matrix of the PCA model with 10 PCs from the data set which combines the in vitro antifilarial activity (-LOGEC50) with the complete set of descriptors of the Selwood dataset. Two common factors are highlighted. The first one is mainly found in descriptors 1 to 3, 5 to 7, 17, 52 and 53. The second one is mainly found in descriptors 35, 36, 38 to 40, 45, 47 and 50. Though the second common factor is not present in -LOGEC50, it is in LOGP.

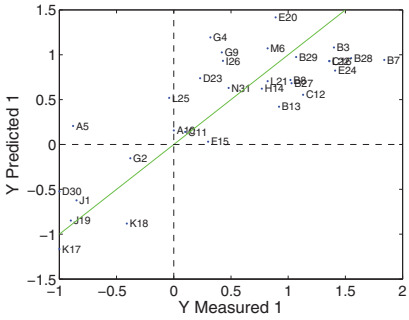


Fig. 25. Plot of measured vs predicted values of -LOGEC50 in the model with regressors LOGP, ATCH6 and MOFI_Y of the Selwood dataset.

compounds previously highlighted are found at the bottom left corner. This result support the non-convenience of isolating these compounds.

Notice that MEDA is not a variable selection technique per se and therefore other methods may be more powerful for this purpose. Nevertheless, being an exploratory method, the benefit of using MEDA for variable selection is that the general solution can be identified and understood, like in the present example. On the contrary, most variable selection approaches are based on highly complex algorithms which can only report a set of possible alternative solutions (e.g. Table 4).

6. Conclusion

In this chapter, new tools for exploratory data analysis are presented and combined with already well known techniques in the chemometrics field, such as projection models, score and loading plots. The shortcomings and potential pitfalls in the application of common tools are elucidated and illustrated with examples. Then, the new techniques are introduced to overcome these problems.

The Missing-data methods for Exploratory Data Analysis technique, MEDA for short, studies the relationships among variables. As it is discussed in the chapter, while chemometric models such as PCA and PLS are quite useful for data understanding, they have a main problem which complicates its interpretation: a single component captures several sources of variability or common factors and at the same time a single common factor is captured in several components. MEDA, like rotation methods or Factor Analysis (FA), is a tool for the identification of the common factors in subspace models, in order to elucidate the structure in the data. The output of MEDA is similar to a correlation matrix but with better properties associated. MEDA is the perfect complement of loading plots. It gives a different picture of the relationships among variables which is especially useful to find groups of related variables. Using a Quantitative Structure-Activity Relationship (QSAR) example, it was shown that the understanding of the relationships among variables in the data may lead to perform variable selection with similar performance of highly sophisticated algorithms, with the extra benefit that the global solution is not only found but also understood.

The second technique introduced in this chapter is a variant of MEDA, named observation-based MEDA or *o*MEDA. *o*MEDA was designed to identify the variables which differ between two groups of observations in a latent subspace, but it can be used for the more general problem of identifying the variables related to a given direction in the score plot. Thus, when a number of observations are located in a specific direction in the score plot, *o*MEDA gives the variables related to that distribution. *o*MEDA is the perfect complement of score plots and much more reliable than biplots. It can also be seen as an extension of contribution plots to groups of observations. It may be especially useful to check whether the distribution of a new set of observations agree with a calibration model.

Though MEDA and *o*MEDA are grounded on missing-data imputation methods and their original algorithms are complex to a certain extent, both tools can be computed with very simple equations. A MATLAB toolbox with the tools employed in this chapter, including MEDA, *o*MEDA, ADICOV and SVI plots, is available at http://wdb.ugr.es/ josecamacho/.

7. Acknowledgement

Research in this work is partially supported by the Spanish Ministry of Science and Technology through grant CEI BioTIC GENIL (CEB09-0010).

8. References

[1] Jolliffe I.T.. *Principal component analysis*. EEUU: Springer Verlag Inc. 2002.

[2] Han J., Kamber M.. *Data Mining: Concepts and Techniques*. agora.cs.illinois.edu: Morgan Kaufmann Publishers, Elsevier 2006.

[3] Keren Gideon, Lewis Charles. *A Handbook for data analysis in the behavioral sciences: statistical issues*. Hillsdale, NJ: Lawrence Erlbaum Associates 1993.

[4] Tukey John W. *Exploratory data analysis*. Addison-Wesley Series in Behavioral ScienceReading, MA: Addison-Wesley 1977.
[5] Teo Yik Y.. Exploratory data analysis in large-scale genetic studies *Biostatistics.* 2010;11:70-81.
[6] Pearson K.. On Lines and Planes of Closest Fit to Systems of Points in Space *Philosophical Magazine.* 1901;2:559–572.
[7] Jackson J.E.. *A User's Guide to Principal Components.* England: Wiley-Interscience 2003.
[8] Wold H., Lyttkens E.. Nonlinear iterative partial least squares (NIPALS) estimation procedures in *Bull. Intern. Statist. Inst. Proc., 37th session, London*:1–15 1969.
[9] Geladi P., Kowalski B.R.. Partial Least-Squares Regression: a tutorial *Analytica Chimica Acta.* 1986;185:1–17.
[10] Wold S., om M. Sj Eriksson L.. PLS-regression: a basic tool of chemometrics *Chemometrics and Intelligent Laboratory Systems.* 2001;58:109–130.
[11] Camacho J.. Missing-data theory in the context of exploratory data analysis *Chemometrics and Intelligent Laboratory Systems.* 2010;103:8–18.
[12] Gabriel K.R.. The biplot graphic display of matrices with application to principal component analysis *Biometrika.* 1971;58:453–467.
[13] Westerhuis J.A., Gurden S.P., Smilde A.K.. Generalized contribution plots in multivariate statistical process monitoring *Chemometrics and Intelligent Laboratory Systems.* 2000;51:95–114.
[14] Camacho J., Picó J., Ferrer A.. Data understanding with PCA: Structural and Variance Information plots *Chemometrics and Intelligent Laboratory Systems.* 2010;100:48–56.
[15] Camacho J., Padilla P., Díaz-Verdejo J., Smith K., Lovett D.. Least-squares approximation of a space distribution for a given covariance and latent sub-space *Chemometrics and Intelligent Laboratory Systems.* 2011;105:171–180.
[16] Kosanovich K.A., Dahl K.S., Piovoso M.J.. Improved Process Understanding Using Multiway Principal Component Analysis *Engineering Chemical Research.* 1996;35:138–146.
[17] Ferrer A.. Multivariate Statistical Process Control based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process *Quality Engineering.* 2007;19:311–325.
[18] Kjeldahl K., Bro R.. Some common misunderstandings in chemometrics *Journal of Chemometrics.* 2010;24:558–564.
[19] L. Fabrigar, D. Wegener, R. MacCallum, E. Strahan, Evaluating the use of exploratory factor analysis in psychological research, Psycological Methods 4 (3) (1999) 272–299.
[20] A. Costello, J. Osborne, Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis, Practical Assessment, Research & Evaluation 10 (7) (2005) 1–9.
[21] I. Jolliffe, Rotation of principal components: choice of normalization constraints, Journal of Applied Statistics 22 (1) (1995) 29–35.
[22] P. Nelson, P. Taylor, J. MacGregor, Missing data methods in pca and pls: score calculations with incomplete observations, Chemometrics and Intelligent Laboratory Systems 35 (1996) 45–65.
[23] D. Andrews, P. Wentzell, Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer, Analytica Chimica Acta 350 (1997) 341Ű–352.
[24] B. Walczak, D. Massart, Dealing with missing data: Part i, Chemometrics and Intelligent Laboratoy Systems 58 (2001) 15–27.

[25] F. Arteaga, A. Ferrer, Dealing with missing data in mspc: several methods, different interpretations, some examples, Journal of Chemometrics 16 (2002) 408–418.
[26] F. Arteaga, A. Ferrer, Framework for regression-based missing data imputation methods in on-line mspc, Journal of Chemometrics 19 (2005) 439–447.
[27] M. Reis, P. Saraiva, Heteroscedastic latent variable modelling with applications to multivariate statistical process control, Chemometrics and Intelligent Laboratoy Systems 80 (2006) 57–66.
[28] F. Arteaga, Unpublished results.
[29] F. Lindgren, P. Geladi, S. Wold, The kernel algorithm for pls, Journal of Chemometrics 7 (1993) 45–59.
[30] S. de Jong, C. ter Braak, Comments on the pls kernel algorithm, Journal of Chemometrics 8 (1994) 169–174.
[31] B. Dayal, J. MacGregor, Improved pls algorithms, Journal of Chemometrics 11 (1997) 73–85.
[32] B. Wise, N. Gallagher, R. Bro, J. Shaver, W. Windig, R. Koch, PLSToolbox 3.5 for use with Matlab, Eigenvector Research Inc., 2005.
[33] Camacho J. Observation-based missing data methods for exploratory data analysis to unveil the connection between observations and variables in latent subspace models *Journal of Chemometrics* 25 (2011) 592 - 600.
[34] D.L. Selwood, D.J. Livingstone, J.C.W. Comley, A.B. O'Dowd, A.T. Hudson, P. Jackson, K.S. Jandu, V.S. Rose, J.N. Stables Structure-Activity Relationships of Antifiral Antimycin Analogues: A Multivariate Pattern Recognition Study, Journal of Medical Chemistry 33 (1990) 136–142.
[35] S.J. Cho, M.A. Hermsmeier, Genetic Algorithm Guided Selection: Variable Selection and Subset Selection, J. Chem. Inf. Comput. Sci. 42 (2002) 927–936.
[36] S.S. Liu, H.L. Liu, C.S. Yin, L.S. Wang, VSMP: A Novel Variable Selection and Modelling Method Based on the Prediction, J. Chem. Inf. Comput. Sci. 43 (2003) 964–969.