Efficient PCA-exploration of high-dimensional datasets

Oxana Rodionova[1], Sergey Kucheryavskiy[2], Alexey Pomerantsev[1]

[1]N.N. Semenov Federal Research Center for Chemical Physics RAS, Moscow, Russia

[2]Department of Chemistry and Bioscience, Aalborg University, Denmark

Abstract

Basic tools for exploration and interpretation of Principal Component Analysis (PCA) results are well-known and thoroughly described in many comprehensive tutorials. However, in the recent decade, several new tools have been developed. Some of them were originally created for solving authentication and classification tasks. In this paper we demonstrate that they can also be useful for the exploratory data analysis.

We discuss several important aspects of the PCA exploration of high dimensional datasets, such as estimation of a proper complexity of PCA model, dependence on the data structure, presence of outliers, etc. We introduce new tools for the assessment of the PCA model complexity such as the plots of the degrees of freedom developed for the orthogonal and score distances, as well as the Extreme and Distance plots, which present a new look at the features of the training and test (new) data. These tools are simple and fast in computation. In some cases, they are more efficient than the conventional PCA tools. A simulated example provides an intuitive illustration of their application. Three real-world examples originated from various fields are employed to demonstrate capabilities of the new tools and ways they can be used. The first example considers the reproducibility of a handheld spectrometer using a dataset that is presented for the first time. The other two datasets, which describe the authentication of olives in brine and classification of wines by their geographical origin, are already known and are often used for the illustrative purposes.

The paper does not touch upon the well-known things, such as the algorithms for the PCA decomposition, or interpretation of scores and loadings. Instead, we pay attention primarily to more advanced topics, such as exploration of data homogeneity, understanding and evaluation of an optimal model complexity. The examples are accompanied by links to free software that implements the tools.

Introduction

Principal component analysis (PCA) [1,2] is a primary method used for analysis of multivariate signals, spectra of various origin, physicochemical data, hyperspectral images, and other high-dimensional datasets which contain hidden information. PCA is available in almost all chemometric software packages and often applied automatically as a routine technical procedure. In this paper, we offer new and easy-to-use tools that help to get a better understanding of the data under study.

PCA can be applied either solely, for the exploratory data analysis, or it can be utilized as the first step of classification (e.g. in SIMCA), or of calibration (e.g. in principal component regression, PCR) [3–5]. PCA is also often applied as a data compression algorithm [6].

When PCA is used for the exploratory analysis, the following conventional tools are employed:

1. Scores (score plot) — to explore the relationship between individual measurements or observations, e.g. for revealing of trends, groups, extreme objects, etc.
2. Loadings (loading plot) — to explore the relationship between variables and to find the influence of variables on the principal components (PCs).
3. Distances (distance plot) — to find extreme objects and outliers in the PCA model with a given number of components.
4. Residual or explained variance (variance plot) – to find an optimal number of components in the PCA model.

The first two tools are pretty straightforward while the last two are rather ambiguous. In particular, the distances (the score and orthogonal distances to the model) are highly dependent on the number of components used in the model. The same object can appear as an outlier or as a regular object depending on the number of PCs. The statistical analysis of the distances can be done in different ways [7] but analysts often use the one that is available by default in a software at hand.

Estimation of the model complexity is also not a straightforward issue [1,2,8–11]. Often, PCA is viewed primarily as a dimensionality reduction method, so the model complexity is defined as the number of PCs that explain most of the systematic variation in the data. This variation is assessed using various characteristics, for example the total explained variance (TEV). In this case, the complexity is selected as the number of PCs that explains at least 80% (90%, or any other predefined value) of the total data variance [1]. There are other rules, e.g. based on eigenvalues (so called the Kaiser rule) [12], or on the analysis of a scree plot, in which either eigenvalues or the explained variance are presented for each PC (so called "elbow" rule) [13].

This leads to a vicious circle in the PCA analysis — the correct interpretation of the distances depends on the optimal number of PCs, which, in turn, depends on the data complexity that can be revealed by analysis of the distances. In addition, the complexity of a PCA model also depends on the purpose for which the PCA results are used.

If we consider a data set obtained from a single population, it usually fits one of the following cases:

1. Noise or fully random data (in this case any number of PCs may be considered as optimal)
2. Structure + noise (so only the PCs explaining structural part are important)
3. Structure + noise + outliers

Sometimes, a data set consists of individual measurements associated with different populations. An example is a set of spectra obtained in different experimental rounds (e.g. the same instrument, but different days). Another example is a case of two mixed populations that results in two structural and two noise parts: Structure 1 + Noise 1 + Structure 2 + Noise 2. Moreover, the structural parts can also overlap, that means they contain shared information which is common for both groups. Those can be the spectra of the same sample collection acquired using two similar spectrometers. In such cases, the selection of the optimal number of components is not straightforward.

Recently, several new tools have been developed in the frame of DD-SIMCA (Data Driven Soft Independent Modelling of Class Analogy) method [14] aimed at solving the authentication and classification tasks. These tools rely mainly on the analysis of distribution of the object distances (the orthogonal and the score distances), assuming that their parameters should be estimated from the data (hence the name). In this paper, we show that these tools can also be useful for the exploratory PCA analysis, especially for complicated cases, in which the conventional methods are not very effective.

We confirm the efficiency of these tools using several case studies based on the data of different nature and various modelling objectives (e.g. detection of extreme objects and outliers, working with data comprised two or more populations, etc.). All examples and plots are developed in R (v. 4.0.2) supplemented with *mdatools* package (v. 0.11.2) [15], so readers can easily reproduce them and apply the tools to their own data. A brief description of *mdatools* functionality related to the topic of the paper is given in supplementary materials (S4). Most of the tools are also available in the DD-SIMCA GUI toolbox for MATLAB [16]. It was also decided to share the simulated spectral dataset used in Section 2 and the related R code; both are available via GitHub repository: https://github.com/svkucheryavski/newpcatools.

1. Theory

1.1 Principal component analysis

PCA operates with matrix $\mathbf{X} = \{x_{ij}\}$ of size $I{\times}J$ (for example I spectra with J wavelength), which is obtained from the original data matrix $\mathbf{X}_{\text{raw}}$ by some preprocessing – centering, scaling, etc. The matrix is decomposed using a formula

$$\mathbf{X} = \mathbf{T}_A \mathbf{P}_A^{\mathrm{t}} + \mathbf{E} \tag{1}$$

where A is the number of principal components (PCs), $\mathbf{T}_A = \{t_{ia}\}$ ($I \times A$) and $\mathbf{P}_A = \{p_{ja}\}$ ($J \times A$) are the matrices of *scores* and *loadings* respectively, and $\mathbf{E} = \{e_{ij}\}$ is the *residuals* matrix. The columns of the loadings matrix are the unit vectors which define the direction of PCs. The rows of the scores matrix are the coordinates of the projections of data points on the PC space. The residual matrix contains a part of the data that is not explained by the PCs.

The PCA model is the first term in Eq. (1) which explains the data using the selected number of PCs, A. The relative residual variance:

$$R_A = \frac{SS_{err}}{SS_{tot}} = \sum_{i,j} e_{ij}^2 / \sum_{i,j} x_{ij}^2 \tag{2}$$

is used as a measure of the PCA model performance. The complementary explained variance is computed as $1 - R_{\mathrm{A}}$. Both values can be obtained for the entire PCA model (giving correspondingly *total residual variance*, TRV, and *total explained variance*, TEV) as well as for the individual contribution of each PC. The values are usually shown in a plot in dependence on the number of PCs.

The relationship between the PCA model and each object can be characterized by two distances: the orthogonal distance and the score distance. The orthogonal distance (OD), q (often denoted as Q),

$$q_i = \sum_{j=1}^{J} e_{ij}^2 \tag{3}$$

is the squared Euclidian distance between a data point, corresponding to the object, and the PC space computed in the original variable space.

The score distance (SD), h (also known as Hotelling's T^2 distance),

$$h_i = \sum_{a=1}^{A} \frac{t_{ia}^2}{\lambda_a} \tag{4}$$

is calculated in the PCA score subspace as the Mahalanobis distance between the projection of the point and the subspace origin.

It was shown[17] that both distances are well approximated by the scaled chi-squared distributions:

$$N_q \frac{q}{q_0} \propto \chi^2(N_q), \qquad N_h \frac{h}{h_0} \propto \chi^2(N_h) \tag{5}$$

where h_0 and q_0 are the scaling factors, whereas N_h and N_q are the numbers of the degrees of freedom (DoF). The scaling factors and DoFs are the distribution parameters which are unknown and estimated using a data driven approach. In case of a regular (no outliers) data, the estimates are based on the conventional mean and variance values calculated for ODs and SDs. Explicit formulae are presented in *Supplementary materials* (Eq. (S14-S15)).

1.2 Irregular data, robust estimates

The term “irregular data” is used to refer to datasets which contain outliers and/or comprise samples from several populations. In these cases, the classic estimators, which are based on the conventional mean and variance values (Eq.(S14-S15)), are not appropriate.

For irregular data, a robust approach has been proposed [14,17], in which the mean and variance of the corresponding distance are replaced with their robust analogues, namely median (M_u) and interquartile range statistics (S_u). All calculation details are presented in Supplementary materials (Eqs. (S16-S17)).

In practice, it is useful to compare the classic and robust estimates of DOFs N_h, and N_q. If the corresponding values (e.g., classic $\hat{N}_h$ and robust $\widetilde{N}_h$) differ considerably, this indicates that the data set is irregular and testifies that the robust method is preferable. Below it is shown that behavior of the DoFs in dependence on the number of PCs, both in regular and irregular cases, is an important characteristic of the data complexity.

1.3 Full distance

To estimate how well an object is fitted by the PCA model a *full distance (FD)*, *f*, is introduced [18]. This is a weighted sum of the OD and SD statistics computed as:

$$f = N_q \frac{q}{q_0} + N_h \frac{h}{h_0} \tag{6}$$

FD also follows the chi-squared distribution with DoF equals

$$N_f = N_q + N_h \tag{7}$$

Using distance *f* we can split the samples into three groups:

Regular samples are those for which the full distance is smaller than a critical value (*regular limit*) computed for a significance level, α. The limit can be obtained using the inverse cumulative distribution function (ICDF, or quantile function) for the chi-squared distribution with a given DoF and probability $p = 1 - \alpha$. The threshold can be shown in the distance plot as a line:

$$f_\alpha = \chi^{-2}(1-\alpha, N_q + N_h) \tag{8}$$

Outliers are the data objects which are significantly different from the regular ones. The outliers can be detected by comparing the full distances with another critical value, an *outlier limit*, computed using ICDF function for probability $p = (1 - \gamma)^{1/I}$.

$$f_{\gamma} = \chi^{-2}(1 - p, N_q + N_h) \tag{9}$$

Here γ is the outlier significance level and I is the number of samples in the training set. Outliers are always harmful for a PCA model and therefore they must be identified and removed.

Extreme samples are the samples which full distance is located between the regular limit and the outlier limit. These samples always exist in a data set and their amount depends on the number of samples and levels α and γ.

2. New tools for exploratory PCA analysis. Simulated example

We introduce the capabilities of the new PCA tools using a simulated dataset. The aim of this dataset is to demonstrate the proposed tools for the data with known complexity. Moreover, this data is close in structure to real spectroscopic data. It is based on a real world example of the NIR spectra obtained in the previously studied of Amlodipine tablets [5].

The simulated data is prepared using the following procedure:

1. The first six loading vectors are taken from the PCA decomposition of the NIR spectra of Amlodipine. This results in the (200×6) orthonormal matrix **V**.
2. The (100×6) orthonormal score matrix **U** is obtained as the PCA scores of the normally distributed random numbers.
3. The diagonal matrix of singular values **S** is designed using a set of values: 7.5; 5; 2.5; 1; 0.5; 0.0001.
4. The (100×200) clean data matrix $\mathbf{X}_b$ is calculated as $\mathbf{USV}^t + \mathbf{D}_{mean},$ where $\mathbf{D}_{mean}$ is a row vector that contains values (individual for each column) that simulate a spectral baseline. Thus, the exact complexity of X_b is equal to six.
5. Finally, the (100×200) data matrix **X** is calculated as $\mathbf{X}_b + \mathbf{G}_\sigma$. Here the (100×200) matrix $\mathbf{G}_\sigma$ contains the Gaussian noise, $N(0; \sigma^2)$.
6. Matrix **X** is divided into two subsets, the (80×200) training set $\mathbf{X}_c$ and the (20×200) test set $\mathbf{X}_t$.

The noise level σ (in step 5) can be varied to track how the actual model complexity is changed and then estimated using the proposed tool. In particular, $\sigma = 0.005$ masks the systematic variations originated from the 5th and 6th components and in this way, reduces an effective data rank up to $A = 4$. In case of $\sigma = 0.025$, the effective rank decreases up to $A = 2$

2.1 Distance plot

The distance plot, in which the score distances are plotted versus the orthogonal distances, is a well-known tool. We propose several improvements, which can potentially extend the use of the plot and make it more efficient. They include the following points:

1. The distance values are normalized ($q/q_0, h/h_0$) using either the classic or robust estimates of the scaling factors, q_0 and h_0. If the distances are spread, a simple log transformation of the axes can improve the plot visibility.
2. The critical limits (for the extreme objects and outliers) should be computed using the distance values in accordance with the selected approach. If outliers are present, the classic estimate often does not reveal them clearly, therefore it is important to study the plot using both the classic and robust estimates.
3. In many cases, the Distance plot can be used complementary to the conventional tools. For example, the scores plot with the Hotelling ellipse can be used for detection of the extreme objects, or for estimation of the optimal model complexity when observations are grouped.

Figure 1 demonstrates the use of the Distance plot as a tool complementary to the score plot. Both plots represent different ‘points of view’ on the same data, which are the spectral data simulated without noise, both the training (80 x 200) and test (20 x 200) sets. Using the same $\alpha = 0.05$ for the critical levels in both plots, we can see three samples located outside the Hotelling ellipse in the scores plot (Figure 1, left), and six objects beyond the regular threshold in the distance plot (Figure 1, right).

The Scores plot is developed using the score values only, and it presents only two selected PCs. The Distance plot is created compositionally using both the SD and OD values, and it provides a cumulative result up to the selected number of PCs.

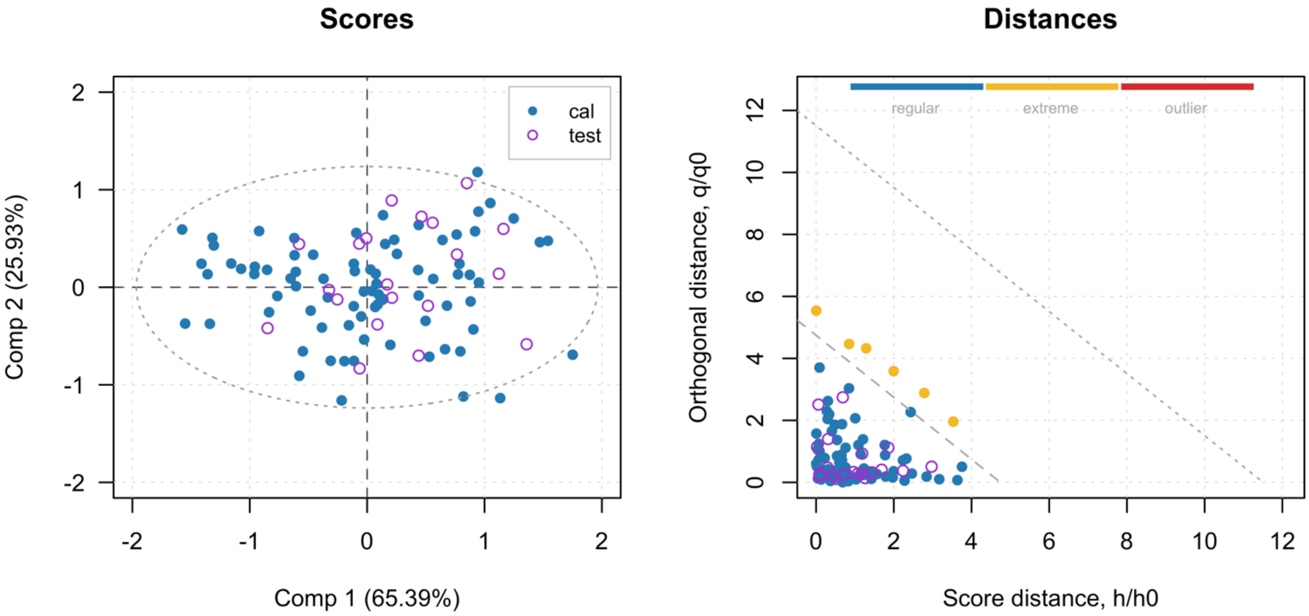

Figure 1. The scores (left) and Distance (right) plots for the first two PCs of decomposition of the simulated dataset without noise.

The critical thresholds are important members of the Distance plot. Their position depends on the proper estimation of DoFs for the OD and SD distributions. In ideal case, the number of samples, which are located inside the regular area, corresponds to the selected significance α for any number of PCs. The choice of estimators (robust or classic) is also important when dataset is contaminated with outliers.

Let us contaminate the simulated training set with six moderate outliers (Figure 2). In this case, the thresholds should be calculated using the robust approach.

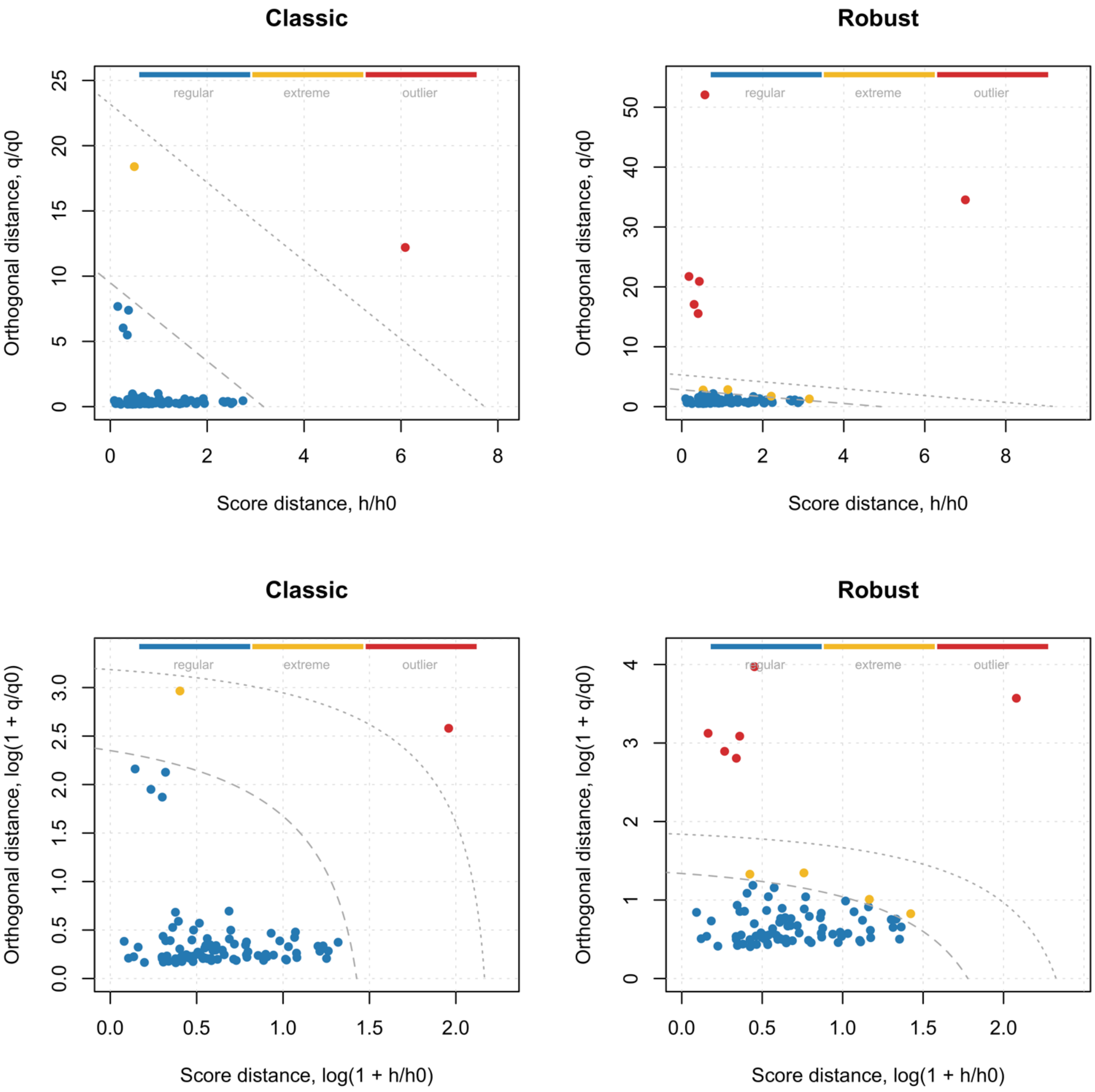

Figure 2. Distance plots for the simulated dataset contaminated with six moderate outliers. The left-side plots are developed using the classic approach for estimation of the distance distribution parameters. The right-side plots are developed using the robust approach.

Plots in the left part of Figure 2 are developed using the classic estimates, while the right plots are based on the robust approach. As one can notice, the classic estimates fail to detect all outliers, and this leads to a wrong evaluation of the extreme observations. Application of the robust approach solves this issue — all six outliers are detected correctly, and the number of extremes, four objects, corresponds to the expected number (5% of 80, as α = 0.05 and I = 80) precisely. The bottom plots show the same results after the log transformation of the axes. This modification makes the plots more convenient for exploration of outliers and extremes.

2.2 DoF plot

The robust, $\tilde{N}_q$, and classic $\hat{N}_q$estimates of the Degrees of Freedom for the orthogonal distance can be utilized for evaluation of the optimal number of components, A, as well as for early indication whether a dataset is contaminated with outliers.

The DoF parameter refers to a statistic, which is the sum of the squared random values. In classical mathematical theory, these random variables are independent, so DoF is equal to the number of terms in the sum. In PCA, this can only happen with a data without structure, which contains only noise. Corresponding plots are shown in Supplementary materials (S2).

Any data structure implies the internal connections and links, which reduce the effective number of independent variables, and thus DoF. By increasing the number of PCs in PCA, we obtain the residuals that are progressively depleted in structures up to the limiting state of the white noise. This leads to an increase of the DoF value. Therefore, by examining the plot of DoF versus PC (which we will call the *DoF plot*), we can notice changes and jumps, which reflect the data complexity.

In order to demonstrate this effect, five datasets are simulated using the procedure described at the beginning of this section. The datasets have the identical structure but various noise levels, σ= {0, 0.005, 0.01, 0.025, 0.05}, which effectively mask a part of the structure that exists in the "clean" data. In this way, the actual complexity is decreased. This mimics real cases in which measurements are always accompanied with noise. Depending on the level of the noise, the part of data structure, which can be explained by the PCA model, is changed.

The left plot in Figure 3 shows the DoF plot developed for each dataset. As one can see, the dependence of N_q (in this case the classic estimate is shown) on the number of PCs reflects the real data complexity very well. For example, in case of the noise with σ = 0.005, we can see a clear break between A = 4 and A = 5, where N_q jumps from 2 to 185. This indicates that for A = 5 there is no structure left in the residuals but only noise. The greater the noise level the less data structure can be explained by the PCA model. For comparison, the corresponding total residual variance plots have no evident signs that indicate the model complexity (Figure 3, right).

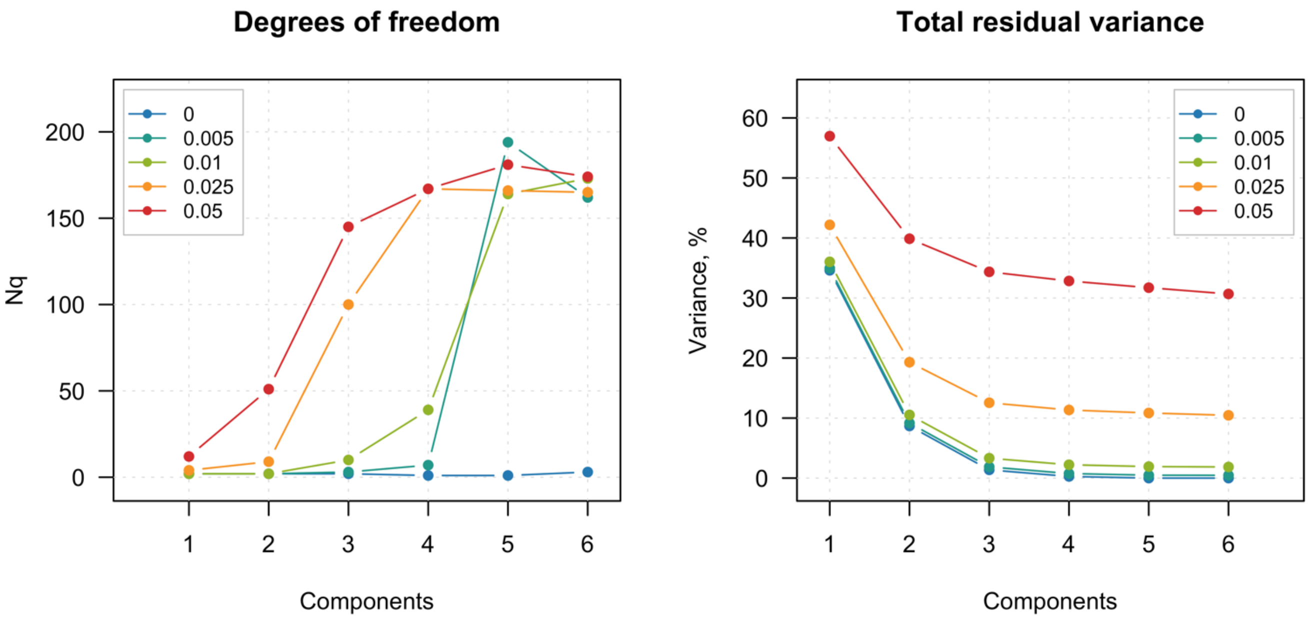

Figure 3. Dependence of DoF for orthogonal distance on the number of PCs. Calculated for data simulated with various level of introduced noise $\sigma = 0; 0.005; 0.010; 0.025; 0.05$. Left plot: Nq vs PC; right plot TRV vs. PC

As it was already mentioned, the conventional way to estimate $\widehat{N}_q$ is sensitive to possible outliers and thus can lead to a wrong estimation. To overcome this problem the robust approach should be used. This gives an opportunity of employing the DoF plot for indication of possible contamination. Therefore, it is necessary to investigate the behavior of both estimates — classic and robust as it is shown in Figure 4. Both plots are built for the dataset simulated using $\sigma = 0.005$, so $A = 4$ explains the structural part of the data. The left plot is developed using the PCA decomposition of the clean, outlier free data. The right plot is built using the PCA model for the data contaminated with six outliers (same as used in Figure 2). Both classic (blue color) and robust (shown in red) estimates for N_q are plotted.

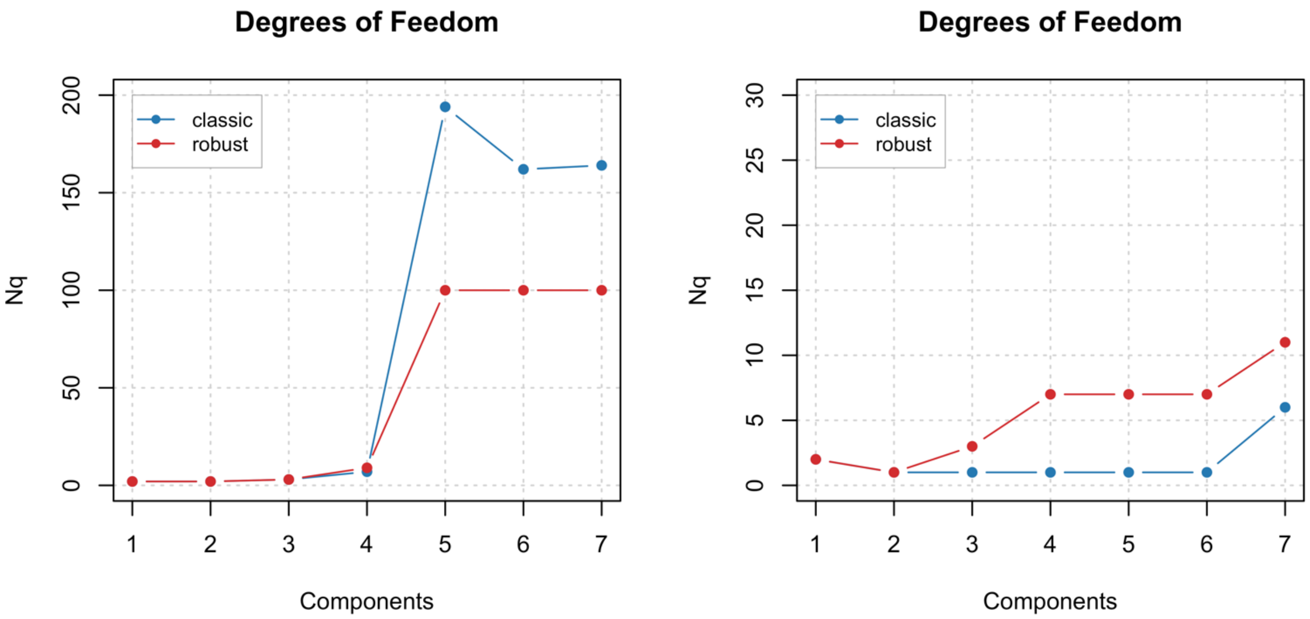

Figure 4. Dependence of DoF for orthogonal distance on the number of PCs. Left plot is made using 'clean' data, right plot — for data with six moderate outliers.

As one can see, there is a very a good agreement between the classic and the robust estimates for all important components in case of the outlier free data (left plot). Also, both statistics show a clear jump between $A = 4$ and $A = 5$ indicating that the optimal number of components is four. It should be noted that starting from $N_q > 50$ there is no point in evaluating and explaining the exact DoF value, so the difference between $N_q = 100$ and $N_q = 150$ is insignificant here.

In case of the contaminated data (right plot), the classic estimate does not show any clear changes up to $A = 7$. In case the robust approach is used, the DoF values increase after $A = 2$. Thus, we observe a significant discrepancy in the behavior of these curves, which signals the presence of outliers. We recommend removing outliers for establishing a correct cutoff level.

2.3 Extreme plot

The Extreme plot is another effective tool, which helps to analyze the model complexity or to assess the model performance. The idea behind this plot is to verify the PCA performance for various α-values at a fixed number of PCs. Since $n = \alpha I$ is the theoretically expected number of extremes, it is possible to track the relationship between the observed number of extremes and the expected one. This can be done graphically by means of the Extreme plot, which shows the empirical number of extremes vs. the theoretical number together with the corresponding tolerant intervals. The abscissa axis shows the expected number of extremes. For example, if α = 0.25 (Figure 5, top plots), we can expect that 25% of samples are extremes. For the training set (*I* = 80) the expected number of extremes is 20 and for the test set (*I*= 20) it equals 5. The ordinate axis shows the actual number of extremes detected by the PCA model for this α value. For example, for *A* = 4 this number is 18 for the training set (you can see that the corresponding point, marked with a red circle, is slightly below the theoretically expected value on the top left plot) and it is 5 for the test set (top right plot, also marked with a red circle). The light blue ellipse shows the 95% tolerance intervals for the observed number of extremes. So, if points are located within the ellipse, the PCA model performance is satisfactory for a given complexity.

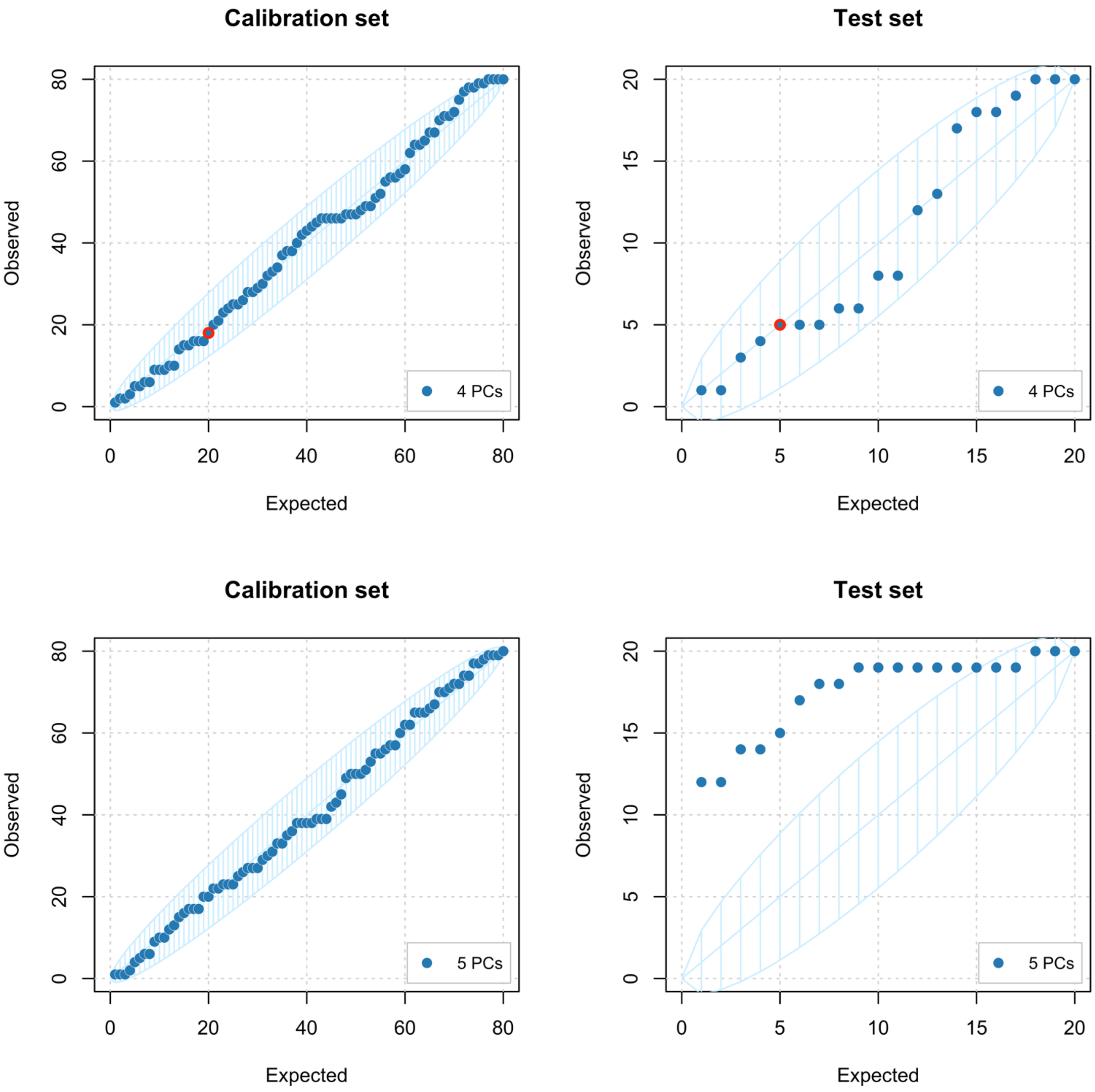

Figure 5. Extreme plots for simulated dataset with effective rank A = 4 (σ = 0.005). The left plots show observations from the calibration set for A = 4 (top) and A = 5 (bottom). The right plots show observations from the test set.

The Extreme plot is sensitive to overfitting and thus can be efficient for estimation of the model complexity regarding the test set. This is illustrated in Figure 5, in which four extreme plots are developed using the PCA model for the simulated data with the noise level σ = 0.005. The left plots represent the training set, and the right plots show the test set. Plots on the top correspond to the PCA model with $A = 4$ PCs, the bottom two plots are for $A = 5$.

We know that the optimal complexity of the model is $A = 4$. The plots that represent the training set (left plots) are similar regardless the number of components. All blue points are located within the tolerance intervals.

The plots for the test set (right) demonstrate a different behavior. In the top-right plot, for $A = 4$, all points are also located within the tolerance intervals indicating that the test set is well described by the model. However, on the bottom-right plot, for $A = 5$, the most of the points are located outside which is a sign of the lack of fit. In other words, if we use 5 PCs, the PCA model starts explaining noise in the training set, which makes the test set modeling worse.

2.4 Procrustes cross-validation

Some of the proposed tools, e.g. the Extreme plots require a test set, which is not always available. This limits the applicability of the proposed tools to a certain degree. Recently, the authors proposed a method that generates a new data set (called *pseudo-validation set*), which can successfully replace an independent test set during the optimization phase, including estimation of the model complexity. The method is called Procrustes Cross-Validation (PCV) [19]. This approach also avoids splitting the data into the training and test, or application of a non-representative test set.

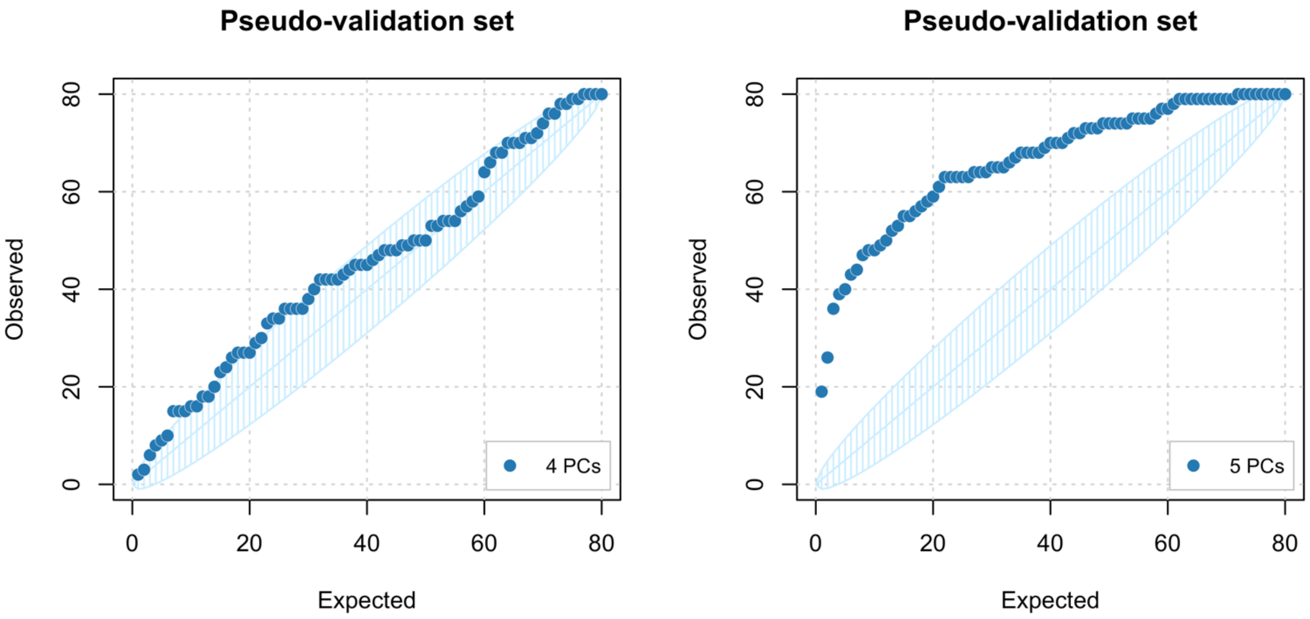

Figure 6. Extreme plots for the simulated dataset with effective rank A = 4 (σ = 0.005). Both plots show results obtained for the pseudo-validation set using PCA model with A = 4 (left) and A = 5 (right).

To demonstrate the applicability of the PCV set, we take the training set from the previous section (left plots in Figure 5), generate the corresponding pseudo-validation set, and use it as a test set. Figure 6 depicts the Extreme plots for the PCV set. It can be seen that the plots demonstrate the same pattern as the plots created for the independent test set — the decent fit for $A = 4$ and the clear sign of overfitting for $A = 5$.

3. Real case studies

In this section we demonstrate the application of the tools for several real-life cases, in which the objectives of the exploratory analysis are different. The choice of a tool is made depending on the objective of the study.

3.1 Case 1. NIR data acquired using two handheld instruments

This case study demonstrates the application of the Distance and Extreme plots for the analysis of the instrument reproducibility.

3.1.1 Dataset description

The dataset consists of the NIR spectra of an anti-inflammatory medicine obtained for 50 tablets from 5 different batches. The spectra are acquired using handheld NIR instruments in the diffuse reflectance mode in the range of 908–1670 nm. Figure 7 (upper plot) depicts the spectra collected by two spectrometers during two days. Subset *S1D1* presents spectra of 50 tablets acquired by Spectrometer 1 during the first day. Subset *S1D2* contains spectra, acquired by Spectrometer 1 during the second day, and subset *S2D1* contains spectra collected by Spectrometer 2 during the first day.

The aim of the study is to compare these three data sets of 50 spectra each. The spectra are corrected using the standard normal variate (SNV) method.

3.1.2 Results

Subset *S1D1* is used for the model training; *S1D2* and *S2D1* are utilized as the test sets. The idea behind these experiments is absolutely clear – we have to be sure that a classification model established on one day is valid for other days. Moreover, the model should be valid for the spectra acquired by another instrument of the same brand. The Distance plots constructed for the *S1D1* model with 2 PCs (Figure 7, left) and 3 PCs (Figure 7, right) show that the *S1D2* samples (the open triangles) are attributed as regular objects in both plots. The samples measured by the second instrument, *S2D1* (open green dots), are located irregularly. Despite the fact that most of the samples for the model with 2 PCs are located in the regular region, all of them have the higher OD values. This abnormality can be observed in the corresponding Extreme plot (Figure 8, right).

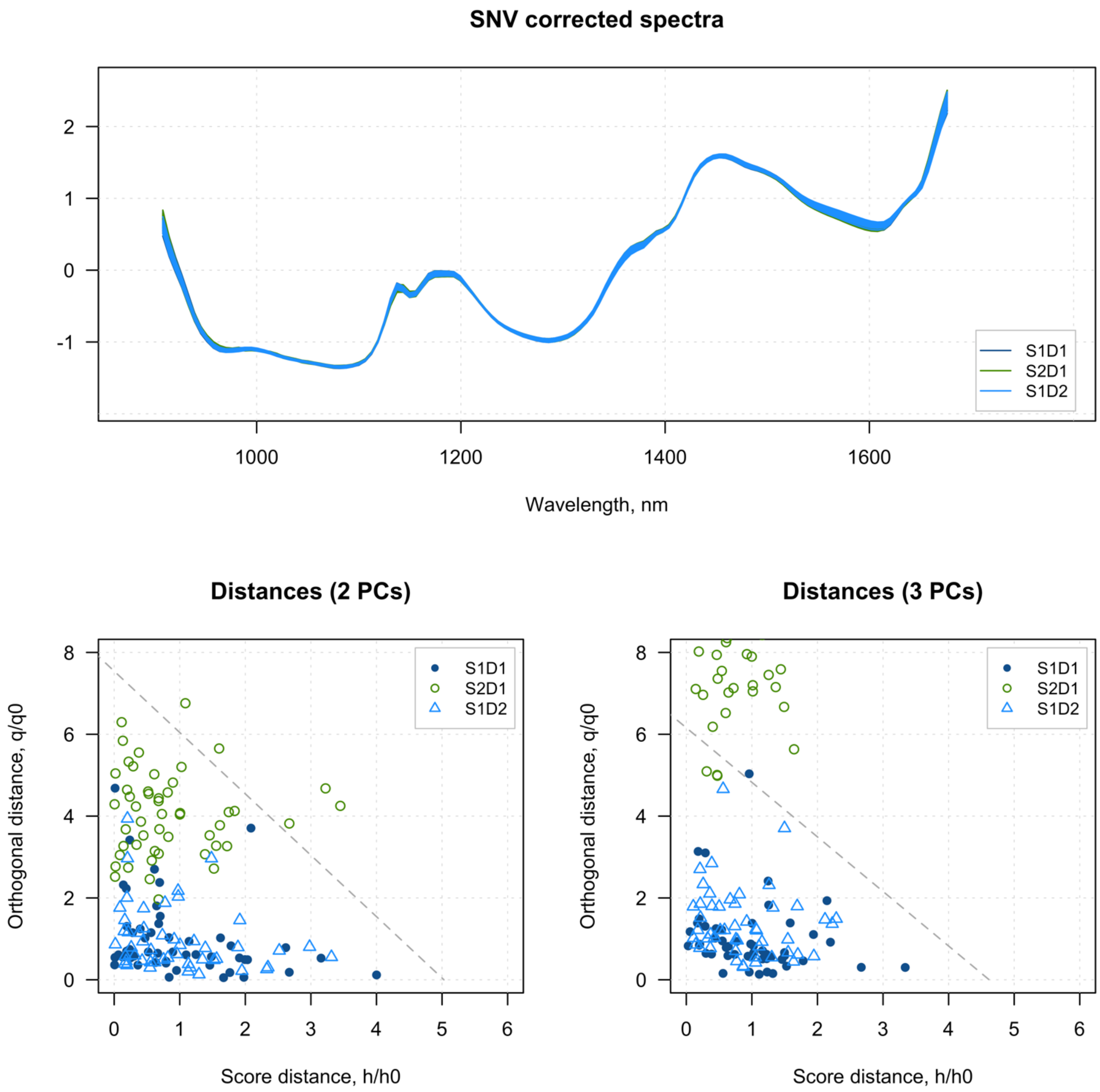

Figure 7. Case1: The SNV-corrected spectra of 50 tablets collected by two NIR instruments during two days (top); the Distance plots built for 2 PCs (bottom left) and 3 PCs (bottom right). Dashed line is the border of regular objects (α=0.01).

For the model with three PCs, the abnormal behavior is evident for both Distance and Extreme plots.

Analyzing the Extreme plot for the subset *S1D2* (Figure 8, left plot) we can conclude that the PCA model is reliable for the routine day-by-day application of Spectrometer 1, when the model complexity is not greater than three. The extreme plot for four PCs goes beyond the tolerance corridor, meaning that a model with $A > 3$ should be applied with caution.

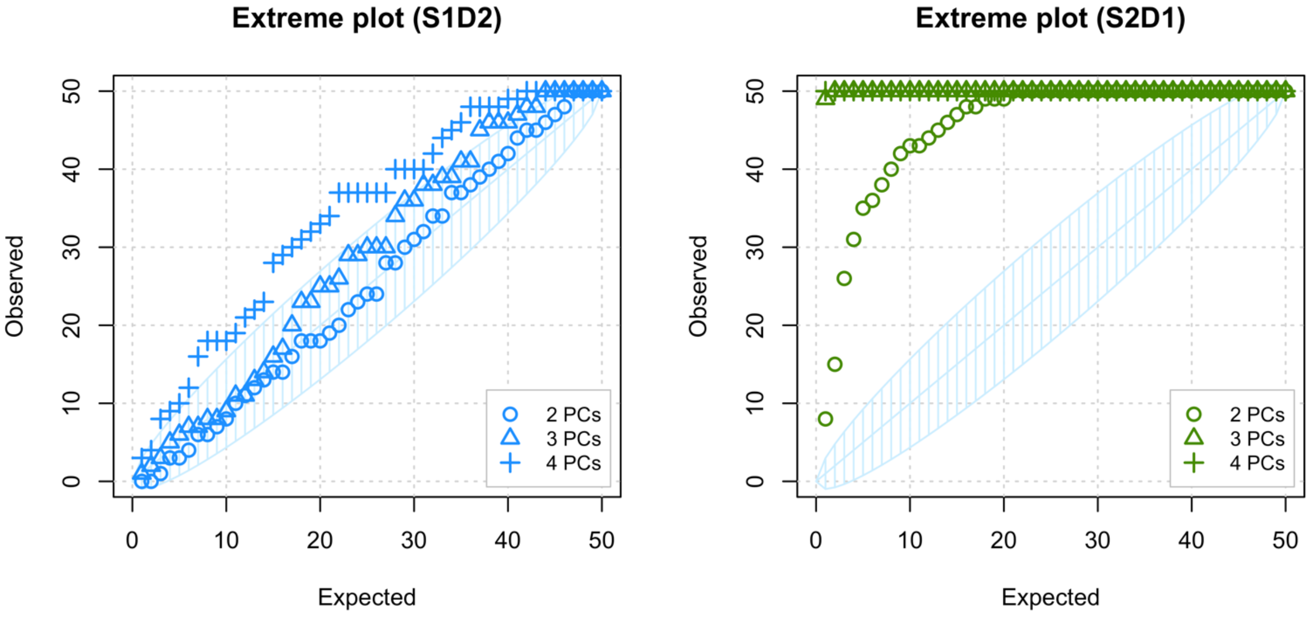

Figure 8. Extreme plot for test set S1D2 and test set S2D1 (for PC2-PC4)

In contrast, the Scores plots for PC1 vs.PC2 and PC1 vs. PC3 (Figure S3 in supplementary materials) do not reveal any visible deviation of subsets *S1D2* and *S2D1* from the training samples. This case study also demonstrates that the Extreme plot is a more sensitive tool than the Distance plot when it comes to assessing the hypothesis that the training and test sets belong to the same population.

It can also be concluded that the complexity of PCA depends on the objectives of the study. In this example, compression of the training set is not the primary concern. The goal is to understand to what extent the complexity can be increased in face of the day-to-day and instrument-to-instrument comparability. From a practical point of view, it can be concluded that the day-to-day reproducibility is acceptable, but a calibration transfer between two instruments is necessary.

3.2 Case 2. Olives data

This example is used to illustrate sensitivity of DoF to the optimal complexity in a PCA model. It also shows how the Procrustes Cross-Validation procedure works.

3.2.1 Dataset description

The data consists of 75 spectra of one selected species of green olives. The spectra are taken by FT-NIR Thermo Scientific spectrometer (Antaris IITM FT-NIR Analyser) and cover the range between 9000 and 4150 cm^{-1} at spectral resolution ≈4 cm^{-1} (1258 spectral values). More details about the data and classification models can be found elsewhere [20,21]. The spectra are SNV corrected prior to the analysis.

3.2.2 Results

From the previous investigations of this dataset, we know that the optimal number of PCs is 4 or 5. Using $A = 6$ or higher leads to overfitting.

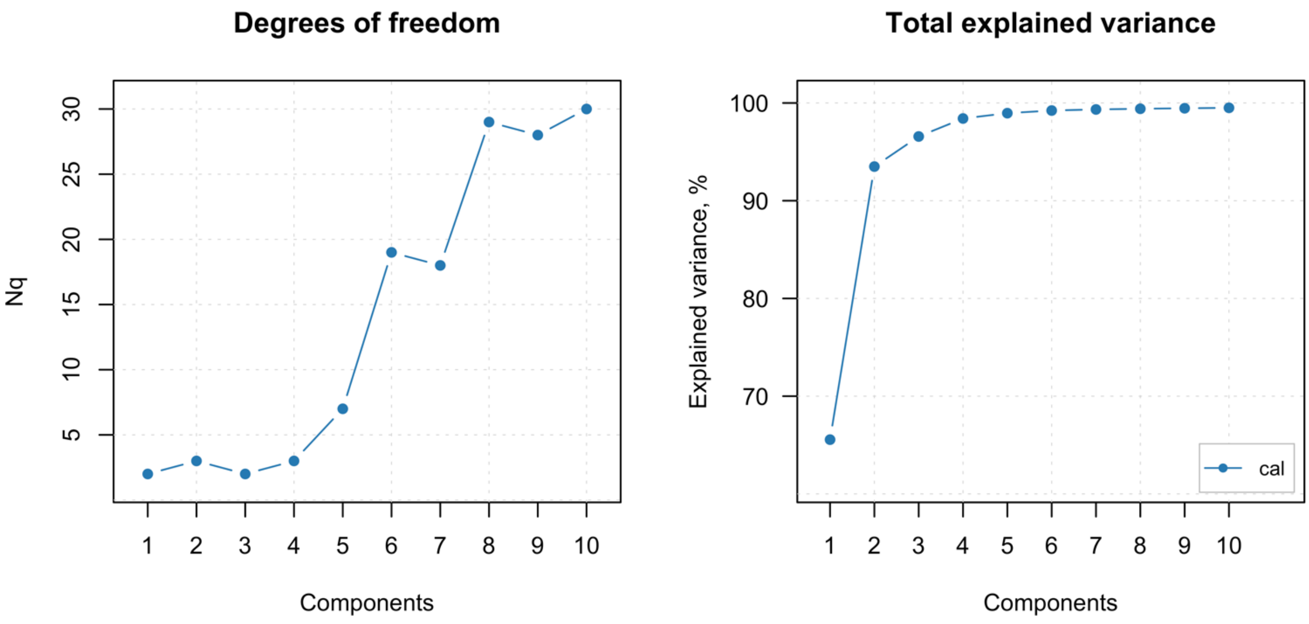

Figure 9. Nq vs PCs plot (left) and Total explained variance plot (right) for the Olives data.

This can also be demonstrated by the DoF plot for the orthogonal distance, as shown in Figure 9 (left plot). Apparently, the DoF plot clearly shows a first small increase between $A = 4$ and $A = 5$. A big jump for $A = 6$, indicates an overfitted model.

In contrast, the total explained variance plot (shown on the right) does not clearly reveal the optimal number of components. The first two PCs explain about 93% of the data variation and contribution of other PCs is very small. This leads us to a wrong decision that $A = 2$.

Instead of an independent test set, we can employ the Procrustes Cross-Validation to create a PCV set for estimation the optimal number of PCs. Figure 10 shows the Extreme plots for the training (left) and PCV (right) sets, for *A* = 4, 5, 6.

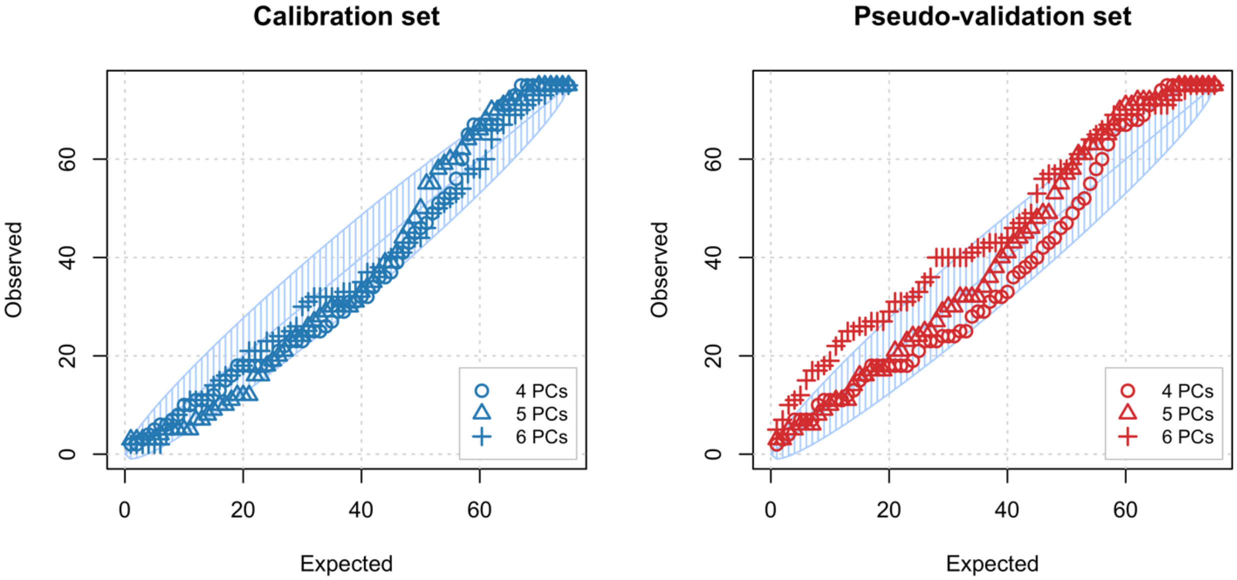

Figure 10. Olives data. The Extreme plots for the training (left) and pseudo-validation sets (right).

For *A*=6 the PCV-set clearly shows the signs of overfitting. This is in accordance with the method discussed above and the previous analysis [21].

3.3 Case 3. Wine data

The data is utilized to illustrate the importance of the robust approach for the data contaminated with outliers.

3.3.1 Dataset description

The dataset [22] has been widely used in various publications mostly related to the classification methods, for example [23,24]. The data consists of 27 variables (chemical and physical characteristics) of 178 wine samples from three different origins: Barolo (59 samples), Grignolino (71 samples), and Barbera (48 samples). Only Grignolino subset is used in this paper. The values are autoscaled prior to the PCA decomposition.

3.3.2 Results

From the original research [22] it is known that each data subset has several outliers. This can be revealed by the comparison the classic and robust estimators of N_h and N_q. In case of a regular data both estimators give quite similar values. On the contrary, the difference can be large when outliers are present. The column "Original data" in Table 1 shows N_q values vs. the number of PCs computed for the initial Wine data (Grignolino subset) using classic (subcolumn 1) and robust (subcolumn 2) methods as well as their difference (subcolumn 3).

The "Outliers free data" column shows the DoF values for the OD obtained after removing the outliers. There is a clear improvement over the original data.

Table 1. Comparison of the Nq values computed using classic and robust approaches for the initial Wine data and after outliers removing

PCs	**Original data, Nq**			**Outliers free data, Nq**		
	Classic	**Robust**	**Difference**	**Classic**	**Robust**	**Difference**
1	8	13	**-5**	9	12	-3
2	7	10	-3	11	14	-3
3	8	9	-1	11	14	-3
4	9	15	**-6**	15	17	-2
5	10	13	-3	14	15	-1
6	12	14	-2	13	11	3
7	13	13	0	13	13	0

Figure 11 shows two Distance plots for $A = 4$. The left plot is built using the classic estimators. The right one is developed using the robust approach. The cut-off levels are calculated for $\alpha= 0.05$ and $\gamma = 0.05$ Obviously, the DoF values estimated using the classic approach ($N_\mathrm{q} = 9$, $N_\mathrm{h} = 3$) and the corresponding thresholds do not help to reveal outliers — all samples are located below the outlier limit shown by the dotted line. The thresholds calculated using the robust DoF values ($N_\mathrm{q} = 15$, $N_\mathrm{h} = 4$) identify three samples as outliers (see in the right plot).

The classic method is not sensitive to outliers, so it leads to biased estimates of DoFs, and, as a result, to expanded outlier limit. The robust approach gets over this issue and detects the outliers correctly.

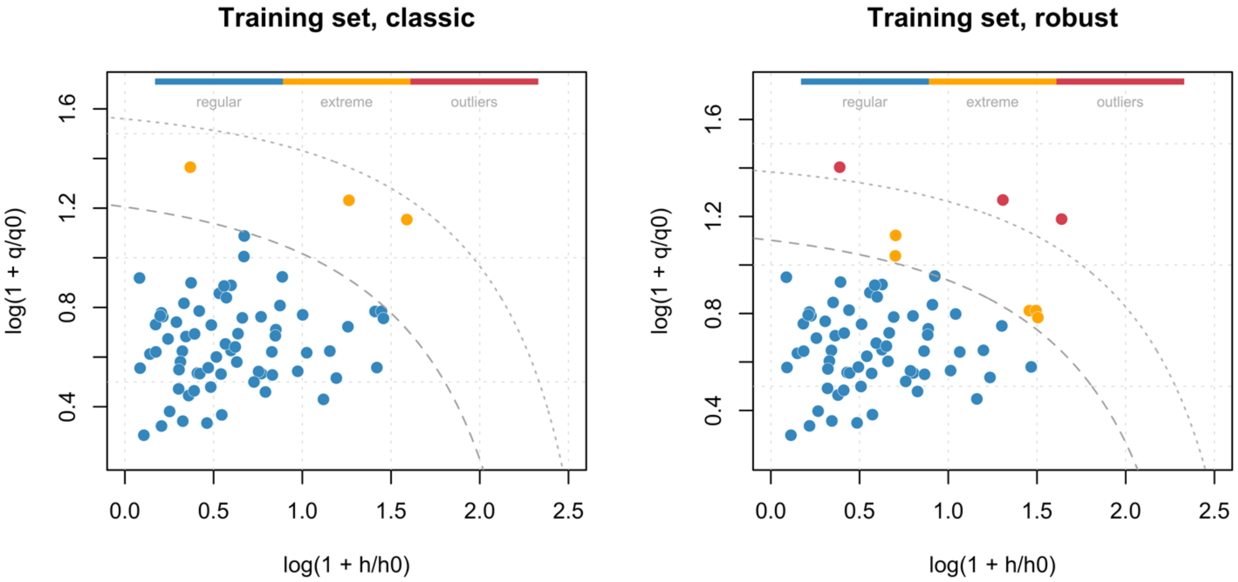

Figure 11. Distance plots for Wine data set, α=0.05 and γ=0.05. The left is for the classic method; the right is for the robust one.

When starting data exploration, it is recommended to build the N_q vs. PCs plots for classical and robust estimates as it is shown in section 2.2

4. Conclusions

In this paper we proposed the new tools that could be useful for the analysis of multivariate data by PCA, especially when it comes to the evaluation of the PCA complexity. Summarizing the presented information, we can conclude the following.

1. The PCA complexity is not just a choice of the number of components, but a broader characteristic, which tells us how well PCA model describes both the training set and the samples from other (test or new) sets.
2. In addition to the conventional tools, the PCA complexity can be investigated and evaluated using the statistics of the score and orthogonal distances. Parameters of their distributions should be estimated using the training set in the same way that is used to estimate other parameters, such as the PLS regression coefficients. This can be referred to as the *data driven* approach.

3. The model complexity depends on the data peculiarities. The analysis of regular data (the data without outliers, one population) and irregular data (the data with outliers and/or a mixture of several populations) requires specific methods. The proposed robust approach can be helpful in latter case.
4. The goodness of fit can be assessed using Extreme plots, which are proven to be a more sensitive tool comparing to e.g. Distance plots.
5. The proposed statistical tools are useful when solving various practical problems of the multivariate data analysis regardless of the analytical instruments used for the data collection and the ultimate goal of a specific study.

Acknowledgement

Oxana Rodionova and Alexey Pomerantsev acknowledge a support within the Russian state assignment No AAAA-A18-118020690203-8

References

[1] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.

[2] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. https://doi.org/10.1039/c3ay41907j.

[3] Y. Kalegowda, S.L. Harmer, Classification of time-of-flight secondary ion mass spectrometry spectra from complex Cu-Fe sulphides by principal component analysis and artificial neural networks, Anal. Chim. Acta. 759 (2013) 21–27. https://doi.org/10.1016/j.aca.2012.11.001.

[4] B. Danylec, C. Kulsing, J.C. Topete, M.T. Matyska, J.J. Pesek, R.I. Boysen, M.T.W. Hearn, Application of linear solvation energy relationships and principal component analysis methods for the prediction of the retention behaviour of E-resveratrol analogues with substituted silica hydride stationary phases, Anal. Chim. Acta. 1090 (2019) 159–171. https://doi.org/10.1016/j.aca.2019.08.072.

[5] O.Y. Rodionova, K.S. Balyklova, A. V. Titova, A.L. Pomerantsev, Quantitative risk assessment in classification of drugs with identical API content, J. Pharm. Biomed. Anal. 98 (2014) 186–192. https://doi.org/10.1016/j.jpba.2014.05.033.

[6] R.G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J.M. Roger, B. Walczak, R. Tauler, Chemometrics in analytical chemistry—part II: modeling, validation, and applications, Anal. Bioanal. Chem. 410 (2018) 6691–6704. https://doi.org/10.1007/s00216-018-1283-4.

[7] A.L. Pomerantsev, O.Y. Rodionova, Popular decision rules in SIMCA: Critical review, J. Chemom. 34 (2020) 1–14. https://doi.org/10.1002/cem.3250.

[8] W.R. Zwick, W.F. Velicer, Comparison of Five Rules for Determining the Number of Components to Retain, Psychol. Bull. 99 (1986) 432–442. https://doi.org/10.1037/0033-2909.99.3.432.

[9] R. Todeschini, Data correlation, number of significant principal components and shape of molecules. The K correlation index, Anal. Chim. Acta. 348 (1997) 419–430. https://doi.org/10.1016/S0003-2670(97)00290-0.

[10] R. Cangelosi, A. Goriely, Component retention in principal component analysis with application to cDNA microarray data, Biol. Direct. 2 (2007) 1–21. https://doi.org/10.1186/1745-6150-2-2.

[11] E. Saccenti, J. Camacho, Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods, Chemom. Intell. Lab. Syst. 149 (2015) 99–116. https://doi.org/10.1016/j.chemolab.2015.10.006.

[12] H.F. Kaiser, The Application of Electronic Computers to Factor Analysis, Educ. Psychol. Meas. 20 (1960) 141–151. https://doi.org/10.1177/001316446002000116.

[13] R.B. Cattell, The Scree Test For The Number Of Factors, Multivariate Behav. Res. 1 (1966) 245–276. https://doi.org/10.1207/s15327906mbr0102_10.

[14] A.L. Pomerantsev, O.Y. Rodionova, Concept and role of extreme objects in PCA/SIMCA, J. Chemom. 28 (2014) 429–438. https://doi.org/10.1002/cem.2506.

[15] S. Kucheryavskiy, mdatools – R package for chemometrics, Chemom. Intell. Lab. Syst. 198 (2020) 103937. https://doi.org/10.1016/j.chemolab.2020.103937.

[16] Y.V. Zontov, O.Y. Rodionova, S.V. Kucheryavskiy, A.L. Pomerantsev, DD-SIMCA – A MATLAB GUI tool for data driven SIMCA approach, Chemom. Intell. Lab. Syst. 167 (2017) 23–28. https://doi.org/10.1016/j.chemolab.2017.05.010.

[17] A.L. Pomerantsev, Acceptance areas for multivariate classification derived by projection methods, J. Chemom. 22 (2008) 601–609. https://doi.org/10.1002/cem.1147.

[18] O.Y. Rodionova, A.L. Pomerantsev, Detection of Outliers in Projection-Based Modeling, Anal. Chem. 92 (2020) 2656–2664. https://doi.org/10.1021/acs.analchem.9b04611.

[19] S. Kucheryavskiy, S. Zhilin, O. Rodionova, A. Pomerantsev, Procrustes Cross-Validation—A Bridge between Cross-Validation and Independent Validation Sets, Anal. Chem. (2020). https://doi.org/10.1021/acs.analchem.0c02175.

[20] P. Oliveri, M.I. López, M.C. Casolino, I. Ruisánchez, M.P. Callao, L. Medini, S. Lanteri, Partial least squares density modeling (PLS-DM) - A new class-modeling strategy applied to the authentication of olives in brine by near-infrared spectroscopy, Anal. Chim. Acta. 851 (2014) 30–

36. https://doi.org/10.1016/j.aca.2014.09.013.

[21] O.Y. Rodionova, P. Oliveri, A.L. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemom. Intell. Lab. Syst. 159 (2016) 89–96. https://doi.org/10.1016/j.chemolab.2016.10.002.

[22] M. Forina, C. Armanino, M. Castino, M. Ubigli, Multivariate data analysis as a discriminating method of the origin of wines, Vitis. 25 (1986) 189–201.

[23] T.H.S. Li, N.R. Guo, C.L. Kuo, Design of adaptive fuzzy model for classification problem, Eng. Appl. Artif. Intell. 18 (2005) 297–306. https://doi.org/10.1016/j.engappai.2004.09.011.

[24] S. Aeberhard, D. Coomans, O. De Vel, Improvements to the classification performance of RDA, J. Chemom. 7 (1993) 99–115. https://doi.org/10.1002/cem.1180070204.

S1. Mathematical notation

PCA and main outcomes

Let **X** be the (I×J) matrix which is obtained from the original data matrix $\mathbf{X}_{\mathrm{raw}}$ by some preprocessing – centering, scaling, etc. Matrix **X** has rank $K \leq \min(I,J)$. The (full) PCA decomposition of matrix **X** is given by the following equation:

$$\mathbf{X} = \mathbf{TP}^{\mathrm{t}} \tag{S1}$$

where $\mathbf{T}=\{t_{\mathrm{ik}}\}$ is the (I×K) *score* matrix, **P** is the (J×K) *loading* matrix. Columns of the loading matrix are orthonormalized eigenvectors that determine the direction of the principal components, $\mathbf{P}^{\mathrm{t}}\mathbf{P} = \mathbf{I}$. The component eigenvalues can be computed as the variance of the corresponding scores:

$$\lambda_k = \frac{1}{I}\sum_{i=1}^{I} t_{ik}^2 \tag{S2}$$

It is evident that:

$$L_0 = \mathrm{Sp}(\mathbf{X}^{\mathrm{t}}\mathbf{X}) = \mathrm{Sp}(\mathbf{T}^{\mathrm{t}}\mathbf{T}) = \sum_{k=1}^{K} \lambda_k \tag{S3}$$

Let us consider the first A (A≤K) PCs in the decomposition given by Eq. (S1):

$$\mathbf{X} = \mathbf{T}_A\mathbf{P}_A^{\mathrm{t}} + \mathbf{E}_A \tag{S4}$$

Matrices $\mathbf{T}_A$ and $\mathbf{P}_A$ include the first A columns of matrices **T** and **P** respectively. The (I×J) matrix $\mathbf{E}_A$ $=\{e_{ij}\}$is the *residual* matrix. A value

$$R(A) = \frac{1}{L_0}\sum_{a=1}^{A} \lambda_a \tag{S5}$$

is called the explained data variation. It varies from 0 (at A=0) to 1 (at A=K).

The relationship between the PCA model and the data points can be represented by two statistics: the orthogonal and score distances. The orthogonal distance (OD), q (also denoted as Q), is the squared Euclidian distance between an observation and the score subspace. For given observation i it is calculated in the original **X** space for the given number of PC as a sum:

$$q_i = \sum_{j=1}^{J} e_{ij}^2 \tag{S6}$$

of the squared residuals presented in matrix $\mathbf{E}_A = \{e_{ij}\}$ that is defined in (S4).

The score distance (SD), h (also known as Hotelling's T^2 distance), is calculated for each sample i as the Mahalanobis distance between a point in the score subspace and its origin:

$$h_i = \sum_{a=1}^{A} \frac{t_{ia}^2}{\lambda_a} \tag{S7}$$

where λ_a are eigenvalues that are defined in (S2). The general property of the two distances, is that both are well approximated by the scaled chi-squared distribution

$$N_q \frac{q}{q_0} \propto \chi^2(N_q),\ N_h \frac{h}{h_0} \propto \chi^2(N_h) \quad \text{(S8)}$$

where h_0 and q_0 are the scaling factors, N_h and N_q are the numbers of degrees of freedom (DoF). These four parameters are considered unknown, and they are estimated using the training dataset [14].

To simplify notations, we will use a generalized symbol $u = \{h, q\}$, which means that the corresponding equation is valid both for h and q. For example, the generalized form of (S8) is as follows

$$N_u \frac{u}{u_0} \propto \chi^2(N_u) \quad \text{(S9)}$$

where u=h, or u=q.

The fact that both distances follow the scaled chi-squared distribution provides a possibility to introduce a new statistic that is called the *full distance*, *f*. It is calculated as a weighted sum of the OD statistics, q, and the SD statistics, h,

$$f = N_q \frac{q}{q_0} + N_h \frac{h}{h_0} \propto \chi^2(N_f) \quad \text{(S10)}$$

It is clear that f also follows the chi-squared distribution with DoF equals:

$$N_f = N_q + N_h \quad \text{(S11)}$$

Given a significance value, α, the limit for the regular objects is determined by an inequality:

$$f \le \chi^{-2}\left((1-\alpha), N_f\right) \quad \text{(S12)}$$

where f is the full distance defined in Eq.(10). The outlier decision rule has a similar form:

$$f > \chi^{-2}\left((1-\gamma)^{\frac{1}{I}}, N_f\right) \quad \text{(S13)}$$

where γ is the outlier significance level.

Classic Estimation of the distribution parameters

For the regular data, the unknown parameters can be found using the method of moments by the following formulae:

$$\hat{u}_0 = \bar{u},\ \hat{N}_u = \text{int}\frac{2\hat{u}_0^2}{s_u^2} \quad \text{(S14)}$$

where “int” stands for rounding to the nearest integer greater than 0, and $\bar{u}$ and s_u^2 are the conventional estimates of the mean and variance:

$$\bar{u} = \frac{1}{I}\sum_{i=1}^{I} u_i,\ s_u^2 = \frac{1}{I-1}\sum_{i=1}^{I}(u_i - \bar{u})^2 \quad \text{(S15)}$$

Robust Estimation of the distribution parameters

The robust approach has been proposed in [14,17], in which the mean and variance of the corresponding distance are replaced with their robust analogues, namely median, M_u, and interquartile range, S_u, statistics. This results in the following expression for the estimation of degrees of freedom:

$$\widetilde{N}_u = \operatorname{int} \exp\left[\left(\frac{1}{d_1}\ln\left(d_2\frac{M_u}{S_u}\right)\right)^{\frac{1}{d_3}}\right] \tag{S16}$$

The parameters, d, are approximated as: d_1=0.72414, d_2=2.68631, d_3=0.84332. The scaling factor, u_0, can be estimated as follows:

$$\tilde{u}_0 = 0.5\,\widetilde{N}_u\left(\frac{M_u}{\chi^{-2}(0.5,\widetilde{N}_u)} + \frac{S_u}{\chi^{-2}(0.75,\widetilde{N}_u) - \chi^{-2}(0.25,\widetilde{N}_u)}\right) \qquad \text{(S17).}$$

S2. DoF plot for matrix with random values

A simple example illustrates the estimation of N_q vs. number of PCs (Figure S2) . The left plot is made for the PCA decomposition of 500 × 20 matrix with random values, while the right plot is made for a similar matrix but with dimension of 500 × 100. The gray dashed line on both plots shows the theoretically expected values. As one can see, in both cases the estimated DoF values are very close to the theoretical expectations, which are equal to (20 – A) and (100 – A) correspondingly.

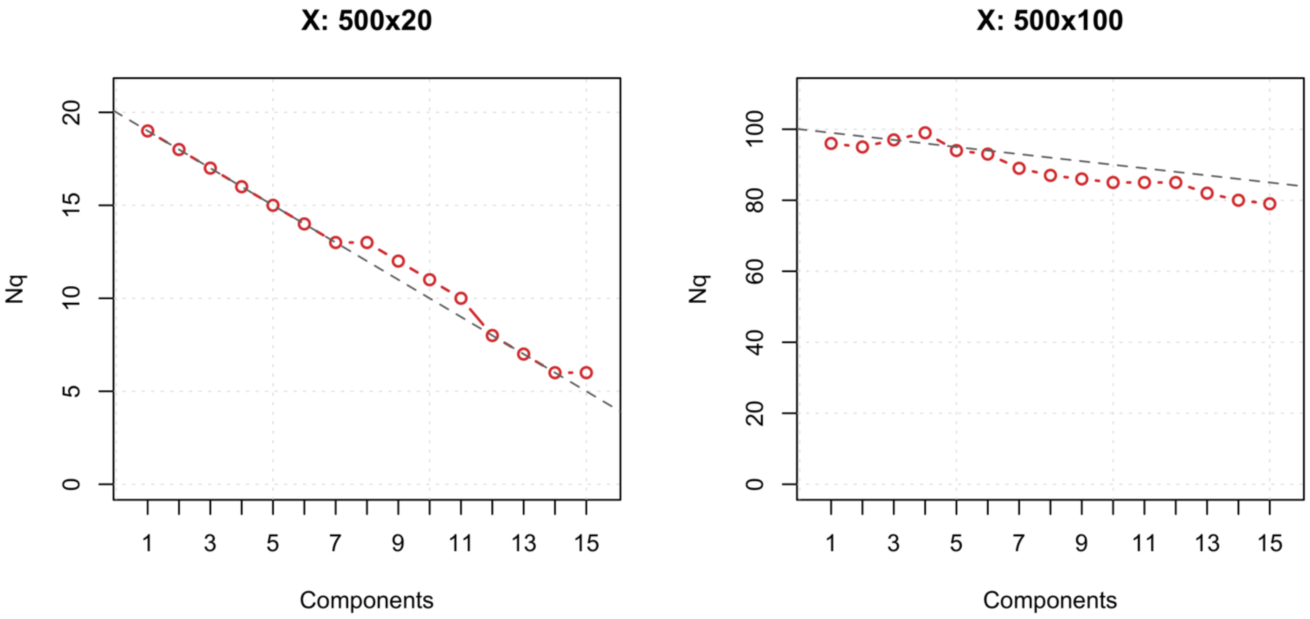

Figure S2. The DoF plot (Nq vs number of PCs) for PCA decomposition of two data sets comprising of normally distributed random values. Left for 500 × 20 and right for 500 × 100.

S3. Scores plots for Case 1

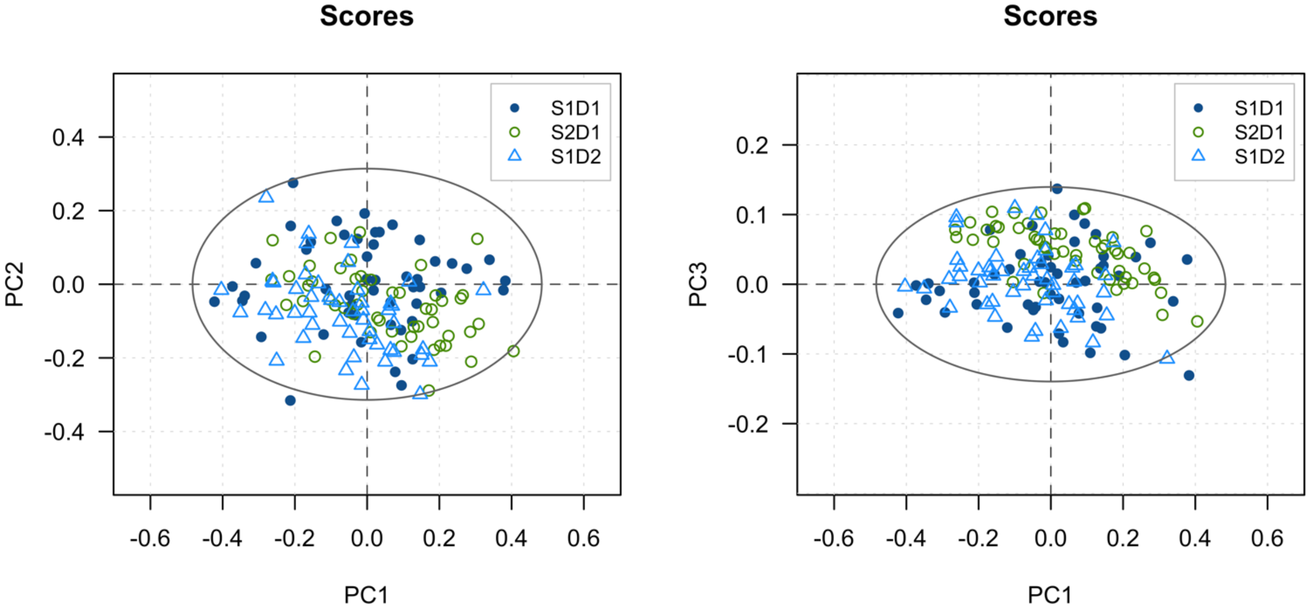

Figure S3. Case 1. Scores plots. Training data S1D1, new subsets S1D2 and S2D1.