The following full text is a publisher’s version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/190831

Please be advised that this information was generated on 2018-10-26 and may be subject to change.
Estimating the number of components and detecting outliers using Angle Distribution of Loading Subspaces (ADLS) in PCA analysis

Y.J. Liu a, b, *, T. Tran a, c, G. Postma a, L.M.C. Buydens a, J. Jansen a

a Radboud University, Institute for Molecules and Materials (IMM), Analytical Chemistry, Heyendaalseweg 135, 6525 AJ, Nijmegen, The Netherlands
b State Key Laboratory of Chemo/Biosensing and Chemometrics, Hunan University, 410082 Changsha, China
c Center for Mathematical Sciences, Merck, Sharp, & Dohme, Oss, The Netherlands

HIGHLIGHTS

• We propose a novel method (ADLS) to quantify the stability of a PCA model.
• ADLS better estimates the number of PCs when compared with other popular methods.
• Different types of simulated and real-life datasets are analyzed.
• We introduce integrated outlier detection in ADLS.

ARTICLE INFO

Article history:
Received 24 November 2017
Received in revised form 20 March 2018
Accepted 23 March 2018
Available online 29 March 2018

Keywords:
PCA
Stability
Bootstrap
Principal components number
Subspace angle
Outlier detection

Abstract

Principal Component Analysis (PCA) is widely used in analytical chemistry, to reduce the dimensionality of a multivariate data set in a few Principal Components (PCs) that summarize the predominant patterns in the data. An accurate estimate of the number of PCs is indispensable to provide meaningful interpretations and extract useful information. We show how existing estimates for the number of PCs may fall short for datasets with considerable coherence, noise or outlier presence. We present here how Angle Distribution of the Loading Subspaces (ADLS) can be used to estimate the number of PCs based on the variability of loading subspace across bootstrap resamples. Based on comprehensive comparisons with other well-known methods applied on simulated dataset, we show that ADLS (1) may quantify the stability of a PCA model with several numbers of PCs simultaneously; (2) better estimate the appropriate number of PCs when compared with the cross-validation and scree plot methods, specifically for coherent data, and (3) facilitate integrated outlier detection, which we introduce in this manuscript. We, in addition, demonstrate how the analysis of different types of real-life spectroscopic datasets may benefit from these advantages of ADLS.

© 2018 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Principal Component Analysis (PCA) is the most widely used multivariate data analysis method, specifically in chemometrics [1,2], which provides an unsupervised interpretation without prior assumptions on grouping or clustering of different samples. PCA
aims to explain the systematic structure of the variability within a multivariate dataset, through the expression of this structure in a relatively low number of new variables (PCs). PCs employ redundancy between the responses of a compound on different measured variables with a certain level of coherence. Coherence, also referred to as collinearity, between spectra of different compounds, may reduce the pseudo-rank of the system due to the converse redundancy of the responses of multiple compounds on the same measured variables. Thereby, coherence may lower the estimated number of PCs when different compounds with coherent spectra are present in the measured mixture.

Selecting an adequate number of PCs for a specific data set is crucial to the interpretation of the results of a PCA analysis. Over- or under-estimation would include noise in the model or leave relevant information in the data unexplained, respectively. Several methods are available to evaluate the most appropriate number of PCs, and the most generic way is the scree plot, which displays the variance explained by each ascending PC [2]. Scree plot is both fast and practical for real datasets of considerable size. However, scree plots often have unclear transition points and their results may be inconclusive with noisy and coherent data: they are more a visual aid based on heuristics. Cross-validation is also commonly used for the same objective. The basic principle of cross-validation is to leave out part of the data, estimate the prediction error and choose the number of components, related to the lowest average prediction error [3–5]. Cross-validation outperforms scree plots with noisy data, but performs equally poor for data with high coherence and is very much dependent on the proportion of the left-out samples [5–7].

Another popular and emerging method to determine the number of components in PCA analysis is bootstrap resampling [6–12]. In this approach, a sufficiently large number of ‘resamples’ are obtained by randomly drawing measured samples from the original data with replacement, which results in resamples of size equal to the originally observed data [13]. The bootstrap resampling can be repeated numerous times, obtaining a bootstrap set with a large number of resamples. The bootstrap set can then be used to estimate the number of components by assessing the stability of PCA models, using the variations of the eigenvectors and eigenvalues, even with a relatively low number of samples and highly non-informative data [6]. Eigenvector-based methods are considered to be more effective as the results based on this type of methods provide more information [6,9]. Most eigenvector-based methods evaluate the variability of each PC, and large confidence intervals correspond to unstable PCA models. However, this type of methods poses some additional challenges, due to the potential ambiguity of component order in PCA [9–11]. Therefore, it is better to evaluate the subspace fitted by all PCs in its entirety, to avoid such challenges. Besse et al. used the Euclidian distance between bootstrap and global sets to represent the stability of the loading subspace, resulting in a poor performance for linear data [14]. Marklewicz et al. obtained the stability of the loading subspace by the variation of the angles between bootstrap and global sets, and the maximum angle between each PC is used as the angle of loading subspace, leading to ambiguous results owing to the problem of reordering of PCs [15,16]. To deal with all the problems, we propose a method named Angle Distribution of Loading Subspaces (ADLS) which is based on the distribution of the angles between the loading subspaces of the bootstrap resamples and the global data set. We have also proposed ADLS to select the number of components in the three-way analysis [17]. In the three-way analysis, components may have physical/chemical relevance, e.g., spectra of particular chemical components in the spectral data set. ADLS estimates this ‘chemical rank’ through the stability of three-way models with different numbers of components. ADLS outperformed a popular method for estimating the chemical rank (Core consistency diagnosis) of noisy and coherent datasets. ADLS may be analogously applied in PCA to estimate the number of components, the ‘pseudo-rank’, by assessing the variation of loading subspaces.

We, therefore, simulate datasets using different types of distributions with different levels of coherence and noise to systematically evaluate ADLS and compared its performance with cross-validation and the scree plot, in line with earlier studies on methods for the selection of an adequate number of PCs [4,7]. These papers provided a benchmark for the component selection in PCA on which we have built our method comparison.

In addition to coherence and noise, outliers are another common disturbing pattern in real-life datasets that influence a PCA model. In this paper, we extend the functionality of ADLS to detect outliers through the variability of the angles between loading subspaces. The enrichment or depletion of specific samples in bootstrap resamples may be associated with being outliers. This allows an integrated estimation of the number of components to match the pseudo-rank of the data, with the detection of outliers that do not match the other samples. We use ADLS to simultaneously detect pseudo-rank and outliers with simulated and real-life datasets. The compare its performance with that of outlier maps with the score and orthogonal distances based on PCA and ROBPCA. ROBPCA is one of the robust versions of PCA expecting to obtain a subspace that is less influenced by the outliers, and therefore has a better performance for outlier detection [18]. The results show that ADLS can detect the outliers for both simulated and real-life datasets. The results are in agreement with those of PCA and ROBPCA.

2. Methods

2.1. Principal component analysis (PCA) and the scree plot

The PCA method reduces the dimensionality of data with a large number of measured variables by transforming these to a new, considerably smaller set of variables called Principal Components (PCs). These variables are fitted to retain as much as possible of the variation present among the specific samples in the dataset. PCA decomposes a matrix $X_{n \times p}$ of which $n$ is the number of samples and $p$ is the number of variables according to Eq (1).

$$X_{n \times p} = T_{n \times p} P_{p \times p}^T + E_{n \times p}$$

The loadings $P$, which describe a new set of orthogonal axes, provide the weights of the original variables in the new PCs. The scores $T$ represent the coordinates of the samples in the space of the loadings. $E_{n \times p}$ is the residue matrix. In PCA model, the first PC describes as much variation as possible by a single dimension; the second and subsequent PCs describe the maximum variability orthogonal to the already fitted PCs. This method is widely used in the analysis of spectroscopic data, as the phenomena that underlie such data may be of considerable lower dimensionality than the measured data itself. Mean-centering is generally used in PCA modeling to have the resulting scores expressing the deviation of each sample from the average spectrum, i.e. to remove information consistent across all samples from the model. The loadings indicate those spectroscopic features most relevant for this variability. Fig.SI-1 is the result of an application of PCA on a spectroscopic dataset. We have utilized three new orthogonal variables (PCs) for the interpretation of the original data, to know the relation between samples and original variables based on the subspace built by a set of PCs. In this example, three PCs are chosen because the first three PCs explain most of the variation, which is commonly
shown by a scree plot. A scree plot displays the explained variation for each PC in descending order versus the number of the components, which is generally used in PCA analysis to choose the number of PCs. However, the scree plot method may provide misleading results when the data contains noise, outliers or coherence. The variability within the noise and these outliers may lead to overestimation of the number of components, while spectral overlap in coherent data may cause underestimation [5]. In a PCA model the subspace defined by a set of loading vectors is the new coordinate basis of the data: the model stability can be represented by the variability in these loadings. Components that do not add systematic variability carried by the majority of samples, i.e. ‘noise’, will lead to large variability upon resampling, which is conducted by the bootstrap method in this paper.

2.2. ROBPCA

ROBPCA reduces the influence of outliers on the estimated scores and loadings, by combining the ideas of both projection pursuit and robust covariance estimation, which yields more accurate estimates than the raw projection pursuit algorithm and PCA [18]. In the ROBPCA approach, the robust covariance estimation is applied to the lower-dimensional data space that is reduced by projection pursuit on the initial dimensions. An important parameter in ROBPCA is the percentage of outliers, and 25% is generally used (which is used in this paper). Currently, there is no effective way to quantitatively estimate this parameter. In this paper, we showed the results for outlier detection by outlier map based on PCA and ROBPCA [18]. The outlier map shows the residuals (orthogonal distance) and score distance (the Mahalanobis distance of samples) based on PCA and ROBPCA. The outlier plot from ROBPCA is robust, which is expected not to be influenced by the outliers.

2.3. Cross-validation

The underlying principle of cross-validation is to leave out part of the data, build a new model, predict the left-out samples, and this procedure is repeated for each part of the data, with an estimation of the Predicted Residual Sum of Squares (PRESS). The number of components with the lowest PRESS should be chosen for the PCA model. There are a number of cross-validation methods applied in PCA, of which the performance of cross-validation by Eigenvector is one of the best, especially for spectroscopic datasets [4,19]. The Matlab code for this method is from PLS-toolbox, Eigenvector Research, Inc.

2.4. Bootstrap

Bootstrap resampling uses random resampling with replacement to construct resamples of equal size with the original dataset [13]. Bootstrap resampling is applicable with small sample sizes, and more importantly, it is not strongly based on distributional assumptions that may not hold up in many real datasets.

In this paper, we conduct bootstrap resampling in the sample dimension (related to the score Y); the loadings from the bootstrap resamples represent the direction (coordinate) of the bootstrap resamples. In this way, we can compare the realization of the loadings for each bootstrap resample using the full complement of measured variables, e.g. wavelength in the case of the spectroscopic dataset. For each model, 1000 bootstrap resamples are constructed as a bootstrap set from which the loadings can be compared to obtain statistical significance.

2.5. The angle between subspaces

In this paper, we utilize the angle between subspaces to show the similarity between loading subspaces, i.e. we do not calculate the angle(s) between the individual loading vectors. The subspace angle between two orthogonal matrices X and Y, in our case two loading subspaces, is calculated in the following way:

$$S = Y - X^T Y$$

$$\theta (X, Y) = \arcsin (||S||)$$

The ADLS [17] approach uses the variation in the angle between loading subspaces calculated in Eqs. (2) and (3) to quantify the model stability. Our earlier studies investigated on the chemical rank estimation in three-way models. In this paper, we extend the method to a stability analysis of PCA models, specifically for estimation of the number of Principal Components. We compare the PCA model calculated on the bootstrap resamples with the global model based on the full dataset [14–17], using the measurement of the angle between the two loading subspaces. For every number of principal components, we obtain an angle distribution. A small variation of angles means a stable loading subspace, indicating a stable PCA model that does not overfit the data. There are four steps for ADLS to select the number of PCs.

The procedure of ADLS model is schematically depicted in Fig. 1. Firstly, the bootstrap resampling is conducted in the sample dimension 1000 times. Then, PCA is applied to both the full data set and the resamples, obtaining global loadings \(P\) and bootstrap loadings \(P_i\). The loading plot in Fig. 1 shows that the first three loadings contain spectroscopic information, as their shape resembles the original spectra, and higher loadings explain the noise and are therefore inherently unstable.

The next step is to compare the spaces defined by the global and bootstrap sub loadings \(P\) and \(P_i\), \(i = 1, ..., 1000\), using the first \(k\) components, by calculating the angle between their subspaces:

$$S = P^* - P_i$$

$$\theta (P, P_i) = \arcsin (||S||)$$

The resulting distribution of the angles is visualized by a boxplot. The upper and lower whiskers (75% and 25%) are used as the range of angles. Fig. 1 shows that the variations of the angles with first two PCs are wide because of the permutation of these PCs having similar variance, or eigenvalue. It is clear that in this case the variance of the loading subspace angles with 3 PCs is small, indicating a stable PCA model with 3 PCs. Besides, the angle variances of the loading subspaces with more than three PCs are wide due to the disturbing pattern structure introduced, indicating that the models are not stable with more than three PCs. Generally, we choose the largest number of components with a small variation of angles, resulting in a set of PCs with stable information. Here, a small variation of angles is defined as the range of angles (upper whisker) smaller than 25°. This is based on the experience of the analysis of ADLS, especially in spectroscopic datasets. However, we
should choose the number of PCs corresponding to the transit point from small angle range to a large angle range in real-life data.

In earlier studies, the other variants of the method based on the variation of loading subspaces are only used to estimate the PCA model stability to select the number of components [14–17]. We extend this in this paper by further investigating the variability of these angles to detect outliers, using the frequency of each sample in the bootstrap set. The variation of the loading subspace will be large in the case of an unstable model, and it may be due to the noise, coherence or outliers. In order to assign the reason for the large angle, we calculate the frequencies of each sample in the bootstrap set, related to the large angle samples. If the widespread of angles are due to noise or coherence, the occurrence of each sample in those bootstrap resamples would be similar. However, in the case of outliers, it is expected that the frequency will be significantly different from the frequency of normal samples. The 99.7% coverage is used to define the ‘significantly different’ samples by adopting the interquartile range (IQR), which is one most common robust measures of scale. IQR is the difference between the 75th percentile and the 25th percentile of the data. It eliminates

Fig. 1. A schematic procedure of ADLS model. The intensities for both spectroscopic and bootstrap resample are in arbitrary units. Variable here means wavelength for this spectroscopic dataset.
the influence of outliers, compared with conventional measures of scale such standard derivation, which is non-robust, greatly influenced by outliers. To assess the stability of a PCA model, there are five sequential steps in ADLS, which we describe below.

I) Create Sub data by bootstrap

The bootstrap resampling is conducted in sample dimension, obtaining 1000 bootstrap subsets $X'$ with the same size as the original data $X$.

II) Obtain the global and sub loadings

The global loading $P$ and 1000 sub loadings $P'$ are obtained from $X$ and $X'$ through PCA analysis.

III) Calculate the angle between the global loadings and bootstrap sub loadings

The similarity between subspaces $P$ and $P'$ is estimated by an angle, represented by $\theta$.

IV) Estimate the range of the angles

The four steps are repeated to calculate the variation of the angles with different numbers of components. The range of the angles is estimated through the upper whisker of the boxplot based on the distribution of $\theta$. Variation of the angles would be small until the extra number of components added to the PCA model. However, the variation of the angles may be large with a smaller number of components than the proper number of components if the reordering of the PCs happens due to similar eigenvalues of the different PCs. In this case, we would obtain a concave shape for the range of different numbers of PCs, with the bottom point corresponding to the appropriate number of PCs. Alternatively, the average angle remains constant for additional components while the range of the angles increases sharply, where the sharp increase indicates the appropriate number of components. Regardless of the situation, we generally choose the largest number of components with a small range of angles (smaller than 25°) to extract all the PCs with useful information.

V) Investigate local enrichment for outlier detection

This step is executed when the smallest variation of the angle distribution is relatively large. The frequency of each sample in the bootstrap set related to the large angles (larger than the angle with median value) is calculated to illustrate the reason for the large angles. Similar frequencies for all samples indicate that outliers are not the reason for the instability; frequencies out of 99.7% coverage are related to outlier samples. The frequency of the outlier samples which is higher or lower than those of the regular samples depends on the influence of outliers which changes the direction of the global loadings or not. If the direction of the global loading is largely changed by the outliers, then the subspace between the global loading and the bootstrap loading with outliers would be small. This is because the global loading subspace, in this case, is biased owing to the outlier(s), that the large angles are contributed by the samples without outliers. If the direction of the global loading subspace is not largely changed by the outliers, then the angle between the subspace of the global loadings and the subspace of the bootstrap loading with outliers would be large. In this case, the frequency of the outlier samples is higher than those of the regular samples in those bootstrap resamples, related the large angle difference with the global model.

3. Data

3.1. Simulated data

3.1.1. Simulated data I

Simulated data I is based on reference [4]. We simulated a spectroscopic dataset with 3 components by $X' = \mathbf{A} \mathbf{C}$, that $\mathbf{A}$ is a spectroscopic (Gaussian peak) matrix and $\mathbf{C}$ is a concentration matrix with random values between 0 and 1. Here the size of $X$ is 100 samples by 100 variables. To investigate more complex datasets, more similar to those encountered in physical experiments, we added disturbing patterns. We introduced 3 levels of coherence ($C_1$, $C_2$ and $C_3$) represented by the correlation coefficient values of matrix $\mathbf{A}$, which are approximate 0, 0.7 and 0.95, shown in Fig.SI-2. Except for the coherence, 4 different levels of noise were introduced here. $N_1$ stands for no noise, $N_2$ stands for 15% heteroscedastic noise ($X_{\text{heter}}$) adding to $X$. The heteroscedastic noise ($X_{\text{heter}}$) is obtained by element-wise multiplication of $X$ and $X_{\text{homo}}$ according to

$$X_{\text{heter}} = X \times X_{\text{homo}}$$

in which $X_{\text{heter}}$ is the heteroscedastic noise matrix, `x` means element-wise multiplication, and $X_{\text{homo}}$ stands for homoscedastic noise with random values at interval [0 1] with uniform distribution. $N_3$ and $N_4$ stand for 1% and 15% homoscedastic noise ($X_{\text{homo}}$) added to $X$. The plot of the spectroscopic datasets with different levels of noise and coherence is shown in Fig.SI-2. Besides, we extended the spectroscopic simulation with datasets having 10 or 100 samples and 15 or 100 variables to explore the performance of ADLS in datasets with different numbers of samples and variables. All the simulations are repeated 100 times.

Furthermore, we showed the performance of ADLS for the datasets with 0, 1 or 5 components in the supporting information. The dataset with 0 component has pure homoscedastic noise. The one-component dataset with Gaussian peak has 1% of homoscedastic noise. The dataset with 5 components has a medium level of coherence (correlation coefficient values of the pure spectra: ~0.7) and 15% homoscedastic noise. All of them have 100 samples and 100 variables.

3.1.2. Simulated data II

In order to investigate the effect of a non-normal distribution, we tested the performance of ADLS on the data generated under a spiked model with skewed distribution from Ref. [7]. This simulation is generated from a Beta distribution with 25 datasets randomly repeated 100 times, with 100 samples and 50 variables. The first dataset has 0 PC, dataset 2–11 have 1 PC, dataset 12–16 have 2 PCs, dataset 17–21 have 3 PCs and dataset 22–25 have 4 PCs.

3.1.3. Simulated data III

Outliers are another common cause of non-linearity in analytical chemical datasets, and the presence of outliers would influence the stability of a PCA model. A PCA model with outlier samples may result in misleading explanations which means that outlier detection is important. Based on the ‘healthy’ spectroscopic dataset shown in Fig.SI-2 (a1), we constructed an outlier dataset by making sample 5 and 9 having different spectroscopic profiles, see Fig. 4 (a). ADLS was applied to analyze this dataset to investigate the effect of outliers on the stability of the PCA model, as well as to detect the outliers.
3.2. Real data

3.2.1. NMR dataset of a ternary mixture design

The first experimental data is an NMR spectroscopic dataset with three simple water-soluble alcohols: propanol, butanol, and pentanol that have spectra with highly overlapping resonances [21]. A tri-axial experimental design, in which the corner of the triangle represents 100% of pure alcohol, is designed for the three alcohol components. Each component has different levels from 0% to 100%. "H NMR spectra were recorded for each of the 231 mixtures on a 400 MHz spectrometer. Prior to Fourier transformation, the corresponding spectra were automatically phased and the baseline was corrected. The original dataset had 65,536 variables; it was reduced to 14,000 variables (3.85–0.65 ppm) in order to remove the water signal and make the investigation more efficient.

3.2.2. Octane dataset based on NIR spectroscopy

The second real-life dataset is the ‘octane’ dataset measured by near-infrared (NIR) absorbance spectroscopy. This dataset was used to show the performance of the ROBPCA method for outlier detection [18]. The dataset concludes 39 gasoline samples and 226 wavelengths, with certain octane numbers. Six of the samples (25, 26, and 36—39) are known as outliers containing added alcohol.

4. Results

4.1. Simulated data

4.1.1. Simulated data I: pseudo-rank estimation

We performed Angle Distribution of Loading Subspaces (ADLS) on the simulated datasets with different levels of coherence and different types of noise to estimate the stability of the PCA model, and subsequently selected the number of components based on the observed angle distributions.

Firstly, ADLS was applied to the ‘healthy’ dataset, and Fig.SI-3 shows the pure spectroscopic profiles (a) and the spectroscopic data with 100 samples (b), indicating the dataset with little coherence and noise. This spectroscopic dataset is also visualized in Fig.SI-2 (a1). Fig.SI-3 (c) shows that the loadings of first 3 components have spectroscopic information and more loadings only introduce noise. The results of the scree plot in Fig.SI-3 (d) shows the accumulated variance explained (left) and the common logarithm of the eigenvalue (right), and both of them have a transit with 3 components, indicating 3 as the proper number of components. Fig.SI-3 (e) illustrates the result of ADLS, clearly indicating a robust PCA model with 3 components. The result in Fig.SI-3 (f) also shows a transit of the value of PRESS with 3 components based on the result of cross-validation. Furthermore, we calculated the angle...
distribution of the loading for each PC separately to illustrate the influence of the reordering and rotation of PCs, as such permutation of PCs is common when two or more PCs have similar eigenvalues. The calculation of angle distribution of the loading for each PC is similar to the calculation of the angle distribution of the loading subspace. In order to show the influence of rotation, we used Procrustes rotation in the bootstrap step to show whether the rotation can solve the problem of the large angle range owing to the similar eigenvalues. Fig. S1-4 (a) shows the loading from one bootstrap resample and Fig. S1-4 (b) shows the corresponding loading after orthogonal Procrustes rotation [22]. Fig. S1-4 (c) and (d) are the results of the angle of distribution of each loading with or without orthogonal Procrustes rotation. The first three components have similar eigenvalue, which causes the permutation of the order of PCs, resulting in a wide spread of angles of the loading vectors for the first 3 PCs, shown in Fig. S1-4 (c). The large angle distribution in Fig. S1-4 (d) with 3 components indicates that rotation cannot solve this problem because reordering of PCs is also the reason for the large range of angles. Permutation of the PC order does not affect the result of ADLS in Fig. S1-3 (e) because ADLS evaluates the entire subspace rather than the individual component.

Coherence is omnipresent in spectroscopy due to spectroscopic overlap, which is also known as collinearity between different compounds, and the estimation of the number of PCs for the dataset with high coherence is difficult. Fig. 2 illustrates the loading plot with the first 4 components, result of the scree plot (d), ADLS (e) and cross-validation (f) for the estimation of the number of PCs for the dataset with high coherence indicated by the overlap of the 3 components shown in Fig. 2(a) and (b). This spectroscopic dataset also corresponds to the dataset in Fig. S1-2 (a8). The accumulated explained variance (left) and the common logarithm of the eigenvectors (right) in Fig. 2 (d) show unclear transit points, while the lowest PRESS with 2 PCs of the cross-validation in Fig. 2 (f) indicates that 2 components should be chosen, owing to the high coherence. The result of ADLS in Fig. 2 (e) shows that the PCA models with one, two and three components are associated to stable subspaces. The variation of angle distribution becomes larger with more than three PCs, due to the introduction of disturbing patterns, such as noise, into the model due to the overfit. Based on the loading plot with 4 PCs in Fig. 2 (c), we can conclude that the first three PCs contain spectroscopically interpretable information. The fourth component introduces considerable instability-inducing noise. The first three loadings in Fig. 2 (d) represent spectroscopic information, as their smooth shape corresponds to that expected from the 3 simulated spectra in Fig. 2 (a). As commonly observed in PCA, the three loadings linearly combine the pure spectroscopic profiles according to maximum variability across the samples and therefore do not correspond to the specific pure spectroscopic profiles of different compounds [1,2]. We may, therefore, select the stable model with three components to preserve all spectroscopic information within the data that is carried by the majority of samples. The ADLS result in Fig. 2 (e) furthermore shows that PCA models with one two and three components are stable across bootstrap resamples, which indicates they may be interpreted as describing the majority of variability in a set of stable latent variables. The results of this dataset with a high level of coherence show that ADLS provides a correct number of components as well as the stability of the PCA model for this dataset with high coherence, while a model with the lower number of principal components recommended by the scree plot and by the cross-validation method does not capture all spectroscopic information.

To better approximate physical reality, we illustrate the result of the dataset with 15% homoscedastic noise \( N_2 \) added to the spectroscopic simulated dataset having median \( C_2 \) level of coherence in Fig. S1-5. This spectroscopic dataset is also corresponding to the dataset in Fig. S1-2 (a8). We then compare the performance of scree plot, cross-validation and ADLS to estimate the number of components. The scree plot in Fig. S1-5 (d) shows that there is no crisp transition point to choose the number of components from both plots with accumulated variance explained (left) and the common logarithm of the eigenvectors (right). However, cross-validation (Fig. S1-5 (f)) shows the lowest PRESS for three components, indicating a satisfactory performance of cross-validation for this level of noise. The result of ADLS in Fig. S1-5 (e) shows that the PCA model with 3 components has a relatively small range of angles, indicating 3 as the number of components. However, this PCA model is unstable because the absolute angle distribution is still considerably large. Also, the loading plot (Fig. S1-5 (c)) shows noisy loadings for the first 3 PCs with spectroscopic information and the fourth PC with only noise, which corroborates this observation. ADLS thereby
provides model stability information that is lost from cross-validation. ADLS provides a model heuristic that does not depend on the measurement units of the data and can, therefore, be directly compared between models of different datasets. As the range of the angles with 3 components is relatively large, we estimate the frequency of each sample related to the large angles (larger than the median angle with 3 components) in the bootstrap set to explore the reason for the unstable PCA model. Fig.SI-6 shows that each sample has similar frequencies, and the upper and lower limits are defined based on 97.3% coverage, indicating that an outlier is not the reason for the instability in this PCA model.

Furthermore, we analyzed the spectroscopic simulated datasets with different levels of noise and coherence as well as different numbers of samples and variables, each dataset consisting of 100 simulations. We only present the result of ADLS and cross-validation and omit of scree plot as there is no clear objective rationale to decide on the number of components for the scree plot method. We show the percentage of the correct results in Fig.SI-7. Fig.SI-7 (a1) and Fig.SI-7 (b1) of ADLS and cross-validation with 10 samples and 15 variables. The results show that both methods have good performances with the datasets with different levels of noise if the level of coherence is low (C1, little coherence). However, both of them provide wrong results for the datasets with the highest level of coherence (C3), with ADLS performing better in the datasets having no noise (N1). Fig.SI-7 (a2) and Fig.SI-7 (b2) visualize the results of ADLS and cross-validation with 100 samples and 15 variables. The results show that ADLS is more sensitive to heteroscedastic noise (N2) than cross-validation. Fig.SI-7 (a3) and Fig.SI-7 (b3) visualize the results of ADLS and cross-validation with 10 samples and 100 variables. They show that ADLS is relatively more sensitive to hetero/homo-scedastic noise (N2 and N4) compared to cross-validation. Fig.SI-7 (a4) and Fig.SI-7 (b4) visualize the results of ADLS and cross-validation with 100 samples and 100 variables. These results show that ADLS performs better for the highest level of coherence (C3) with no or low level of noise. Based on the results of the datasets with different levels of noise and coherence, as well as different numbers of samples and variables, we conclude that ADLS provides better results for the datasets having a high level of coherence. However, ADLS is more sensitive to noise than cross-validation if the number of samples or variables is small.

As a control experiment, we investigated a dataset with only pure noise, which is expected to have 0 PC in PCA analysis. Fig.SI-8 (a) represents the pure noise dataset, and Fig.SI-8 (b) is the loading plot showing that none of the loadings has any pattern. Fig.SI-8 (c) and (d) show the result of the scree plot with accumulated variance explained and the common logarithm of the eigenvalue. Both of them show no transit point with variance evenly distributed among the different PCs, leading to a soft conclusion of 0 PC. Fig.SI-8 (e) shows the result of ADLS, clearly indicating that the loading subspaces are unstable with any number of components. Fig.SI-8 (f) is
the result of cross-validation showing that the PRESS value ascent with the increase of the number of PCs. Here, the result of ADLS gives the information of stability of the PCA model which is not provided by scree plot or cross-validation.

Additionally, we simulated a spectroscopic dataset with 1 component. Fig.SI-9 shows the data (a), the loading plot (b), the scree plot with accumulated variance explained (c), the common logarithm of eigenvalue (d), the result of ADLS (e) and the result of cross-validation (f). The scree plot shows that 1 component explained over 99% variance with a clear transit point in the plot of common logarithm of eigenvalue in Fig.SI-9 (d). The result of ADLS clearly shows that 1 component indicates a stable PCA model while cross-validation PRESS value goes up with the increase of the number of PCs. Here, we can see that the result of ADLS is clear with 0 or 1 component while the cross-validation method is less conclusive: the first PRESS value is the smallest for the datasets with both 0 and 1 component.

In order to investigate the dataset with a large number of components, we show the results of the simulated spectral dataset with 5 components in Fig.SI-10. This 5-component dataset has a medium level of coherence and 15% homoscedastic noise indicated by the pure spectra of 5 components in Fig.SI-10 (a) and the spectra of the 100 samples in Fig.SI-10 (b). The loading plot in Fig.SI-10 (c) shows that 5 components have spectral information. Fig.SI-10 (d), (e), and (f) show the results of the component number estimated by the scree plot (the scree plot with accumulated variance explained, left, the common logarithm of eigenvalue, right), ADLS and cross-validation. Cross-validation indicates that only 4 components are necessary for PCA modelling, which is insufficient. ADLS and the scree plot using the common logarithm of eigenvalue indicate 5 components as the number of components in PCA modelling. ADLS shows an advantage of estimating the number of components in this spectral dataset with a relatively large number of components compared with cross-validation.

4.1.2. Simulated data II: dataset with skewed distribution

The boxplot in Fig. 3 shows the predicted numbers of components based on ADLS (a) and cross-validation (b) for the 25 simulated skewed distribution datasets, repeated 100 times. The numbers on top of the boxplots are the true number of the components. We can see from Fig. 3 (a) that ADLS provides correct results for dataset 1 (0 component), dataset 5–10 (1 component), dataset 11–20 (2 components), dataset 21–30 (3 components), dataset 31–40 (4 components), dataset 41–50 (5 components), dataset 51–60 (6 components), dataset 61–70 (7 components), dataset 71–80 (8 components), and dataset 81–90 (9 components). The frequency of each sample related to the angle larger than the median angle (c) for simulated dataset III. The upper and lower limits are defined based on 97.3% coverage.

Fig. 5. The score plot (a) and outlier map (b) based on ROBPCA, and the frequency of each sample related to the angle larger than the median angle (c) for simulated dataset III. The upper and lower limits are defined based on 97.3% coverage.
dataset 14–15 (2 components), dataset 19–20 (3 components) and dataset 24–25 (4 components). The results in Fig. 3 (b) show that cross-validation only provides correct results for the data with 1 component as the first component is always corresponding to the smallest PRESS value for 25 datasets. The conclusion is further confirmed by Fig.SI-11 showing the results provided by ADLS (a) and cross-validation (b) of the 24 datasets (dataset 2 to dataset 25).

We didn’t show the result of the first dataset as there is 0 component involved in the first dataset which is the situation discussed before (Fig.SI-9). From the ADLS graphs one can conclude that it is sometimes more difficult to estimate the number of components for these simulated skewed datasets. The cross-validation method mostly provides wrong results [7]. ADLS provided better results for the datasets with a skewed distribution compared with cross-validation, as ADLS estimates the robustness of the loading space by bootstrap resampling, which is not based on the assumption of a normal distribution.

4.1.3. Simulated data III: outlier analysis

Fig. 4 (a) shows the spectroscopic profile of the dataset with sample 5 and 9 as outliers having different spectroscopic profiles. The results of estimating the number PCs for the dataset with outlying samples 5 and 9 are plotted in Fig. 4(e and f), illustrating that the number of PCs should be 3 having the smallest PRESS value with the cross-validation method and the smallest variation of angles using ADLS. Besides, the sum of the variation for the first 3 PCs is 99%, according to the score plot in Fig. 4(b), which means 3 PCs are enough for a PCA model (similar function as a scree plot).

Fig. 4 (b) shows that sample 5 and 9 are outliers. Based on the cutoff lines and the position of the samples in the score space and/or orthogonal distances space, the outliers are then classified into three types: orthogonal outlier, bad leverage outlier and good leverage outlier, shown in Fig. 4 (d). The outlier map indicates sample 5 and 9 as bad leverage outliers. For a comparison, we also show the result of ROBPCA to detect outliers, using 3 components and 25% as the percentage of outliers (default value). Fig. 5 shows that the outliers can be detected by the ROBPCA score plot (a) and the outlier map (b). Also the angle distribution of loading subspace with 3 components is considerably large (Fig. 4 (e)), which means that this is an unstable PCA model resulting from noise, coherence and/or outliers. We used ADLS to investigate the reason for the unstable PCA model, whether the instability was caused by either outliers or other reasons. We calculated the frequency of each sample in bootstrap set with angles larger than the median angle related to 3 components, which are shown in Fig. 5 (c).

If the large angles were due to noise or coherence, each sample would have similar frequencies in the large angle bootstrap sets. However, large angles due to outliers would lead to a significantly different frequency of the outlier samples. Fig. 5 (c) shows that the frequencies of samples 5 and 9 are clearly out of the lower limit line, indicating them as the outliers.

According to the results for this simulated dataset, we show that ADLS can detect outliers as well as estimate the number of PCs, based on the information of the variation of the loading subspace.
4.2. Experimental data

4.2.1. NMR dataset from a ternary experimental design

Fig. 6 (a) shows the result of ADLS for the NMR dataset of a ternary experimental design. It shows that a PCA model with two components is most stable because the variation in angles is smallest, with a similar result of cross-validation shown in Fig. 6 (b). However, here it took hours of the cross-validation method to estimate the number of the components owing to the size of the dataset (14,000 variables), while it took less than 1 min of ADLS to finish the calculation. A robust PCA model with 2 components due to the ternary experimental design is verified by the small variation of the angle distribution of the loading subspace with 2 components [21]. The score plot in Fig. 6 (c) shows that the shape of the score plot with 2 PCs is similar to the ternary experimental design for the concentrations of propanol, butanol, and pentanol [21], and the variances for the first two PCs are over 97%, indicating that 2 components are enough for the PCA model.

4.2.2. Octane dataset based on NIR spectroscopy

Fig. 7 shows the results based on the methods of ADLS (a) and cross-validation (b) for real data II. We selected two PCs because the ranges of the angles are relatively larger with more than two PCs based on ADLS, see Fig. 7 (a). The result of cross-validation is unclear with a continuous decreasing PRESS which may be owing to the spectroscopic information shared by all the PCs, shown in the loading plot in Fig. 7 (d). The score plot in Fig. 7 (c) shows that the variance for the first 2 PCs is over 97% and the six known outlier samples can clearly be observed with 2 PCs. The PCA-based outlier map (Fig. 7 (e)) indicates unclear information for the outlier detection. Furthermore, this PCA model with 2 PCs is not stable indicated by the result of ADLS with an angle range larger than the median angle with 2 PCs (Fig. 7 (a)). The reason for the unstable model might be resulting from the outliers indicated in the score plot in Fig. 7 (c), and this is verified by the frequency of each sample in the bootstrap set, related to the large angles based on ADLS in Fig. 8 (c). The frequencies of the six samples (25, 26 and 36–39) are zero, which are clearly significantly different from those of the

Fig. 7. The results of ADLS (a), cross-validation (b), the score plot with 2 PCs (c), loading plot for the first 4 PCs (d), and the outlier map based on PCA (e), for real dataset II.
other samples, indicating them as outliers. For a comparison, Fig. 8 (a) shows the result of ROBPCA with 2 components, with a similar score plot based on PCA. Furthermore, the outlier map in Fig. 8 (b) shows that both the orthogonal and score distances of the outliers are large, indicating them as bad leverage outliers. The results indicate that ADLS, similar to ROBPCA, can detect outliers.

5. Conclusion

We propose to estimate the stability of the PCA model by the variation of loading subspace, using Angle Distribution of Loading Subspace (ADLS) coupled with bootstrap resampling. Compared with other methods based on the variation of the loading subspace, ADLS has better performance for the linear dataset and resolves the problem of reordering of PCs. With the information on stability, ADLS can be used to select the number of PCs with a stable PCA model, and its performance was compared with the methods of cross-validation and scree plots. The results show that ADLS outperforms the methods of cross-validation and scree plots. ADLS provides the information for the stability of the PCA model and the correct number of components, especially for datasets with a high level of coherence while the other two methods give biased results for coherent datasets without the information of the stability. In addition, the spread of the angles is for the first time investigated to detect outliers for the unstable PCA model, by estimating the frequency of each sample in bootstrap set related to large angles. With the analysis of simulated and real datasets, the results of ADLS show that it can detect outliers, and that the results are in agreement with those based on the outlier maps of ROBPCA. ADLS is proven to be a method to estimate the number of components as well as detect outliers with the stability information of a PCA model.

Acknowledgement

The China Scholarship Council is gratefully acknowledged (grant to Y.J. Liu). We would also like to show our gratitude to Prof. Hai-Long Wu for his advice and support.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.aca.2018.03.044.

References

[2] H. Abdi, L.J. Williams, Principal component analysis, Wiley interdisciplinary


