Metabolomics (2012) 8:S37–S51
DOI 10.1007/s11306-011-0361-9

ORIGINAL ARTICLE

Multiblock principal component analysis: an efficient tool for analyzing metabolomics data which contain two influential factors

Yun Xu · Royston Goodacre

Received: 1 March 2011 / Accepted: 26 August 2011 / Published online: 11 September 2011
© Springer Science+Business Media, LLC 2011

Abstract Principal component analysis (PCA) is probably one of the most used methods for exploratory data analysis. However, it may not be always effective when there are multiple influential factors. In this paper, the use of multiblock PCA for analysing such types of data is demonstrated through a real metabolomics study combined with a series of data simulating two underlying influential factors with different types of interactions based on 2×2 experiment designs. The performance of multiblock PCA is compared with those of PCA and also ANOVA-PCA which is another PCA extension developed to solve similar problems. The results demonstrate that multiblock PCA is highly efficient at analysing such types of data which contain multiple influential factors. These models give the most comprehensive view of data compared to the other two methods. The combination of super scores and block scores shows not only the general trends of changing caused by each of the influential factors but also the subtle changes within each combination of the factors and their levels. It is also highly resistant to the addition of 'irrelevant' competing information and the first PC remains the most discriminant one which neither of the other two methods was able to do. The reason of such property was demonstrated by employing a 2×3 experiment designs. Finally, the validity of the results shown by the multiblock PCA was tested using permutation tests and the results suggested that the inherit risk of over-fitting of this type of approach is low.

Keywords Multiblock PCA · Consensus PCA · ANOVA-PCA · Metabolomics · Experiment design · Simulation

1 Introduction

After suitable data pre-processing principal component analysis (PCA) (Jolliffe 2002; Brereton 2003) is normally the first step in metabolomics data analysis in order to gain an initial view of the data in terms of exploratory analysis. One of the main reason is that PCA does not require labelling information (i.e. it is an unsupervised approach) and generally does not need a validation step to verify the clustering trend shown by the analysis. Therefore it can work on a relatively small sample sized data where additional validation samples may not be readily available. PCA summarise all the major variance in the data set into a subset of mutually orthogonal latent factors (principal components, PCs) and each PC represent one major trend of independent variance. In metabolomics studies, there can be many different variations and, in most cases, only a small subset of these are of interest. The hope with PCA is that the variance of interest can be separated from other unwanted ones and be summarised into one or a few PCs which can be visualised in a lower dimensional space. However, due to its unsupervised nature, PCA lets the data "tell" the story all by itself and it is possible that the major

Electronic supplementary material The online version of this article (doi:10.1007/s11306-011-0361-9) contains supplementary material, which is available to authorized users.

Y. Xu (✉) · R. Goodacre
School of Chemistry, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK
e-mail: yun.xu-2@manchester.ac.uk

R. Goodacre
Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

variance (i.e., in the first few PCs) revealed by the PCA are not the ones the study aimed to discover. In addition, when more than one influential factor exists in the data and there are interactions between these influential factors, PCA is not always able to separate adequately the variations caused by each factor. These will be rarely presented in different individual PCs, and more often the variations caused by these factors are spread across a number of different PCs which makes the results rather difficult to interpret. Even when multiple influential factors are independent to each other, if the variation caused by these factors are similar the resulting subspace constructed by PCA can be a rotated space of the original one (i.e., each PC is a linear combination of several underlying latent factors) instead of having one PC represented one individual influential factor and thus the results will still be difficult to understand. There are some methods which aim to rotate PC axis found by PCA to make the results easier to interpret and this can be done both in an unsupervised way (e.g., varimax rotation (Kaiser 1958)) or a supervised way (e.g., PCA followed by discriminant function analysis, PC-DFA (Manly 2005)). The drawback of unsupervised rotation is that its objective still has no direct link to the aim of the study. For example, the objective of varimax is to find a set of PCs so that for each PC high loadings are generated from a few variables while the rest are close to 0, there is no guarantee that these "a few variables" with the highest loadings in the first a few PCs are the ones of interest. Hence its success is rather situation dependent. Supervised rotation, on the other hand, such as PC-DFA (where the rotation is to minimise within class variance and maximise variance between different classes) has the inherit risk of over-fitting, the axes are no longer orthogonal (even if they are treated as such) and like all supervised methods proper validation is generally required. Another important aspect is that when there is more than one influential factor that exists in the data and there are interactions between the factors, the difference between classes (i.e., a unique combination of the studying factors and their levels) can be obscured by the interactions. When this happens, even supervised rotations cannot show the differences because, as a result of such interaction, there may be no real difference at all between certain classes in the data.

Multiblock PCA (Westerhuis et al. 1998; Smilde et al. 2003) is an extension of PCA which is designed to find the underlying relationships between several sets of possibly related data with the emphasis to reveal the "common trend" between these data. These methods have been extensively used to integrate the data generated by multiple analytical platforms such as combining the data from denaturing gradient gel elecrophoresis (DGGE; a DNA-based community typing approach) and gas chromatography-mass spectrometry (GC-MS) (Zomer et al. 2009); or combining the readings from multiple chemical sensors in process control (Qin et al. 2001; Ferreira et al. 2010). There is another potential use of such methods. Multiblock PCA can be used to analyse the data generated by a balance-designed experiment with multiple and possibly interacting factors. For example, assuming there are 2 factors, denoted as $F1$ and $F2$, in the data and each factor has a and b different levels respectively, one can re-arrange the data in 2 different ways to make it fit the multiblock PCA model. One is to build an a blocks data matrix with each block contains b different levels of factor $F2$ at one particular level of $F1$. In this data matrix, the influence of $F1$ becomes a baseline while that of $F2$ becomes a common trend between different blocks and thus stands better chance to be discovered by multiblock PCA. The second way is to construct another b blocks data matrix which can be made in the same way so that the change caused by $F1$ can be revealed. The hope is that by re-arranging the data into multiblocks, the influence of potentially interaction factors are explicitly separated and the existence of multiple influential factors is no longer a hurdle and instead it increases the chance of discovering the influence caused by the underlying factors. Even the changes caused by the influential factors which have been obscured by the interactions between the factors can still be reveal after the multiblocking. Previously we have reported several metabolomics studies using multiblock PCA for analysing the data with multiple factors with satisfactory results. In one study (Kassama et al. 2010) we were able to reveal the difference in the metabolic profiles between different strains of bacteria and also the change caused by different extraction techniques; in another study (Xu et al. 2010) multiblock PCA was able to reveal the changing in the volatile organic compounds (VOCs) in spoiled meat over time and also the difference between the natural spoilage and the spoilage caused by an inoculated pathogen. In this current paper we firstly use the data from one real metabolomics study which has been published before (Kassama et al. 2010) to highlight the advantage of multiblock PCA over classical PCA. Then we analyse the use of multiblock PCA to analyse this type of data in detail through a series carefully designed simulations. For comparison purpose, ANOVA-PCA (Harrington et al. 2005) which is another PCA extension designed to model multiple influential factors in the data was also applied to the simulated data sets. The results of these two PCA extensions are compared with that of the classical PCA. It is worth noting that there is another extension of PCA under the name of ANOVA simultaneous component analysis (ASCA) (Smilde et al. 2005) which is closely related to ANOVA-PCA. However since the first 2 modes in ASCA are the results of PCA performed on the means of the different levels of the

factors, the number of the data points of these modes is the same as the number of the different levels of one factor and the number of the possible combinations of the factors and their levels respectively. This makes the interpretation rather trivial when the number of levels is low. For example, when it is applied to a 2 × 2 experiment design, the scores of the first mode would only have 2 data points and 4 points for the 2nd mode. Thus this method is not included in this paper. Nevertheless we expect these two methods would reach very similar results as they are both based on a very similar methodology. In addition to comparing the performance of the 3 models, the robustness of using the multiblock PCA for analysing such data is also assessed using a permutation test procedure.

2 Experimental and data analysis

2.1 Real metabolomics data

For brevity reason we only give an brief description of the study in this article, the details can be found in (Kassama et al. 2010). The study was to investigate the metabolomics profiles of three strains of *Streptomyces lividans* TK24: (1) wild type (W); (2) empty pIJ486 plasmid (P) and (3) pIJ486 expressing urine mTNF-α (T) using GC-MS. Two different extraction methods were used to extract intracellular metabolites: (1) conventional manual extraction through vortex and (2) using a ultrasonic adaptive focus acoustics (AFA) treatment in a Covaris™ S1 single tube system (Covaris Inc., Woburn, MA, USA). The aim was to investigate the differences in the metabolic profiles of the three different bacterial strains and also assess whether there is any significant difference in the metabolic profiling by using the two different extraction methods. A 3 × 2 experiment was conducted: for each strain, seven cultures were cultivated and for each culture two equal volume samples were taken, one for manual vortex extraction and another for AFA treatment. This gives a total number of 42 samples. The GC-MS analysis detected 120 unique metabolite peaks and this resulted in a data matrix of 42 × 120. In the original paper, PCA and consensus PCA were performed on the data matrix and in this work we also employed ANOVA-PCA for comparison purpose.

2.2 Simulation

Component analysis methods such as PCA are based on the concept of latent factors (Borsboom et al. 2003). A latent factor is a mathematically inferred variable from the observed variable which captures the variance caused by a set of observed variables which follow a similar trend in the data. When this applies to real data analysis, the variables of interest to be discovered are the ones under the influence of the factors which the experiment aimed to study and they change in a specific way when the levels of the influential factors changes. It is common in real applications that multiple observed variables follow a similar (or the opposite) trend of changing under the influence of the influential factors and thus can be effectively described by a single latent variable. This is the main reason that PCA and other component analysis methods can be used for dimensionality reduction. Based on this concept, two types of latent factors can be found in the data, one is the trend the experiment aimed to discover and these are the latent factors of interest. There will also be other trends which are irrelevant to the aim of the study and these will also form a number of other 'irrelevant' latent factors that compete with those of interest during the component analysis. The relative ratio of the variance caused by these two types of latent factors determines how easy or difficult the pertinent trend of interest can be discovered by the component analysis. The simulation conducted in this paper mirrors this concept with simulated data.

First, a set of 10 variables with the trend of interest were generated. Assuming these variables are under influence of 2 factors, denoted as $F1$ and $F2$, each factor has 2 different levels, denoted as $L1$ and $L2$. To discover the influence of these 2 factors at different levels, a 2 × 2 experiment design is required to record the response of the variables at the 4 different statuses: ($F1L1$, $F2L1$) (means $F1$ at $L1$ *and* $F2$ at $L1$), ($F1L2$, $F2L1$), ($F1L2$, $F2L1$) and ($F1L2$, $F2L2$). If there were no interactions between $F1$ and $F2$, the response of the variables under the influences of these 2 factors will be like Fig. 1a: 4 distinctive responses are seen, one for each class (i.e., a unique combination of the factors and their corresponding levels) can be found and the trajectory of changing from $F2L1$ to $F2L2$ at $F1L1$ is *parallel* with the that at $F1L2$ and vice versa. Thus an effective component analysis will be able to show 4 distinct clusters in one of its component (i.e., a latent factor) and the results should be easy to interpret. However, if the two factors are interacting, that is to say the trend of changing of one factor may be different while the other factor is changing its level simultaneously, the responses of the variables under these influential factors will be more complicated. There could be numerous possible interactions and Fig. 1b–f illustrated 5 different possible interactions which have been simulated in this study:

Interaction 1, Fig. 1b: the trend of $F2$ is still the same when $F1$ changes its level, however the trajectory of changing of $F2$ is no longer parallel with each other between the different levels of $F1$. For this type of interaction happens, 4 distinctive clusters can still be observed with a subtle change in that there is less difference between

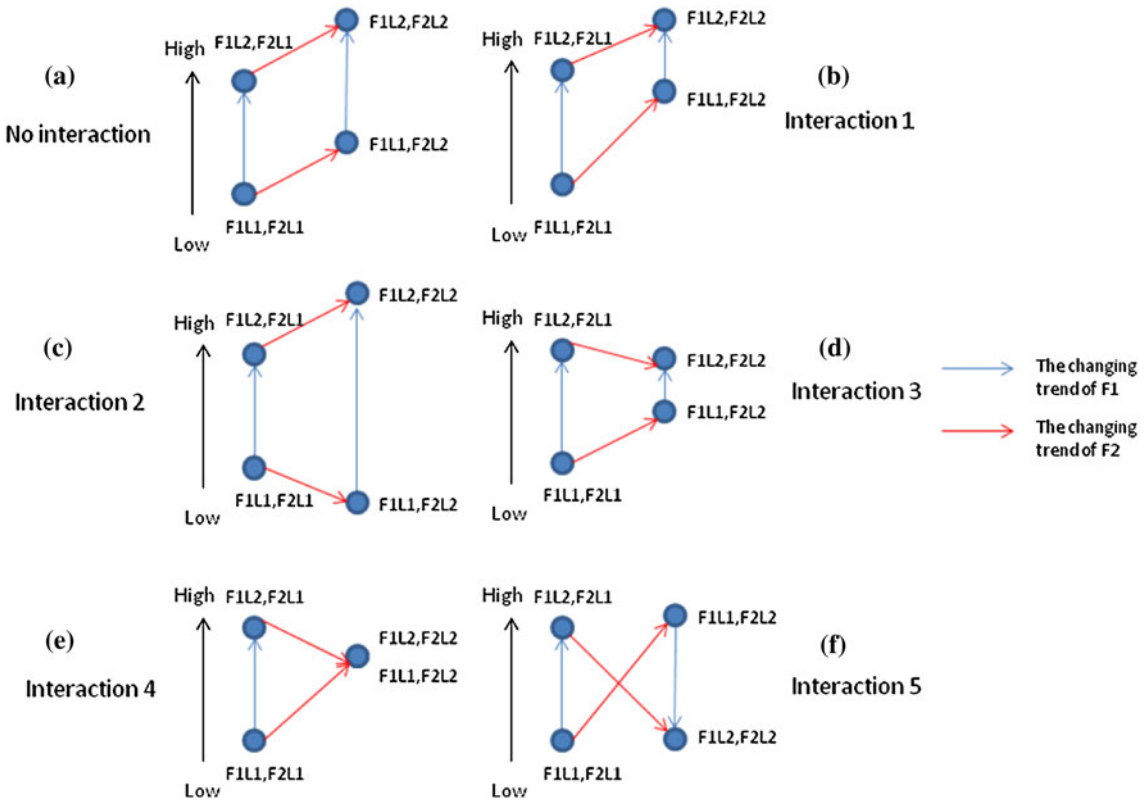

Fig. 1 The different types of interactions simulated in this study

(*F*1*L*1, *F*2*L*2) and (*F*1*L*2, *F*2*L*2) compare to the same pair at *F*2*L*1.

Interaction 2, Fig. 1c: the trend of *F*2 is reversed when *F*1 changes its level and as the result the difference between the two levels of *F*1 is enhanced when *F*2 changes its level, yet the trend remains the same, 4 distinctive clusters exist.

Interaction 3, Fig. 1d: this is similar to Interaction 2, the trend of *F*2 is reversed when *F*1 changes its level except that the difference between the two levels of *F*1 is reduced when *F*2 changes its level, still with 4 distinctive clusters.

Interaction 4, Fig. 1e: This is a special case of Interaction 3, the trend of *F*2 is reversed when *F*1 changes its level and the difference between the 2 levels of *F*1 disappears when *F*2 changes its level to *L*2, in this case only 3 distinctive clusters exist. There is no difference between the 2 levels of *F*1 when *F*2 is at *L*2.

Interaction 5, Fig. 1f: the trend of *F*2 is reversed when *F*1 changes its level and vice versa. This type of interaction can lead to only 2 distinctive responses to be observed, the difference between the two levels of *F*1 cannot be distinguished if *F*2 changes its level accordingly and vice versa.

For each "interesting" variable, it contains the responses of the 4 different classes. Each class is characterised by its mean (μ) and standard deviation (s). The exact number of the means were randomly chosen from 1 to 10 while they follow a certain order for each particular interaction as illustrated in from Fig. 1a–f. For example the order of the means at 4 different classes of the Interaction 1 are $\mu_{(F1L1,\ F2L1)} < \mu_{(F1L1,\ F2L2)} < \mu_{(F1L2,\ F2L1)} < \mu_{(F1L2,\ F2L2)}$ with an additional condition that $\mu_{(F1L2,\ F2L2)} - \mu_{(F1L1,\ F2L2)} < \mu_{(F1L2,\ F2L1)} - \mu_{(F1L1,\ F2L1)}$. Such order of the Interaction 4 becomes $\mu_{(F1L1,\ F2L1)} < \mu_{(F1L1,\ F2L2)} = \mu_{(F1L2,\ F2L2)} < \mu_{(F1L2,\ F2L1)}$. The standard deviation of each class was set to be 15% of the corresponding mean. A total of 100 responses were generated for each class using a normal distribution random number generator with the corresponding means and the standard deviations. A total of 10 "interesting" variables were generated in this manner and they form a latent factor of interest to be discovered by the component analysis.

In addition to the variables of interest, a number of "not-so-interesting" variables were also generated. A random vector $\boldsymbol{x}$ of 400 elements were generated from a normally distributed random number generator with a randomly chosen mean and standard deviation and used as a "seed", an additional 9 variables $\boldsymbol{x}_{\text{ext}}$ were derived from the "seed" using a linear extension:

$$\boldsymbol{x}_{\text{ext}} = b \cdot \boldsymbol{x} + c + \boldsymbol{\varepsilon}$$

where the slope b and the offset c are also randomly chosen from 1 to 100 and $\boldsymbol{\varepsilon}$ is a random noise vector to prevent a

perfect correlation between the derived variable; this represents one type of trend which is not of interest. Sufficient noise was added so that the correlation coefficients between the derived variables as well as the corresponding seed varied from 0.7 to 0.9 which is similar to those between the "interesting" variables. One "seed" with its 9 derived variables form a competing latent factor and another 9 competing latent factors were formed using different randomly generated seeds.

For each type of interaction, a series of matrices were generated by combining the corresponding latent factor of interest with 1–10 competing latent factors. Each of the matrices is thereafter auto-scaled so that each variable has a mean of 0 and a standard deviation of 1. This makes each latent factor, including the one of interest, have the same variance. These matrices represent one problem with different levels of difficulties. It should be very easy to discover the latent factor of interest when the ratio of these two types of latent factors is 1:1 while it should be the most difficult to discover it when such ratio becomes 1:10.

2.3 Multiblock PCA

A graphic illustration of multiblock PCA is given in Fig. 2a. Assuming a data matrix $\boldsymbol{X}$ comprises k blocks, denoted as $\boldsymbol{X}_1$, $\boldsymbol{X}_2$,… to $\boldsymbol{X}_k$, the multiblock PCA is defined as:

$$\boldsymbol{X} = [\boldsymbol{T}_1 \cdot \boldsymbol{P}_1 \boldsymbol{T}_2 \cdot \boldsymbol{P}_2 \ldots \boldsymbol{T}_k \cdot \boldsymbol{P}_k] + \varepsilon \text{ with}$$

$$\boldsymbol{T}_{\text{sup}} = \sum_{i=1}^{k} w_i \cdot \boldsymbol{T}_i$$

and

$$\boldsymbol{X}_i = \boldsymbol{T}_i \cdot \boldsymbol{P}_i + \boldsymbol{\varepsilon}, \ \boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \ldots \boldsymbol{k}$$

The model comprises 4 parts: super scores $\boldsymbol{T}_{\text{sup}}$, block scores $\boldsymbol{T}_i$, block loadings $\boldsymbol{P}_i$ and block weights $\boldsymbol{w}$. The "global" trend of the data is revealed in the super scores $\boldsymbol{T}_{\text{sup}}$ while the trend of each block is shown in the corresponding block scores $\boldsymbol{T}_1$, $\boldsymbol{T}_2$,… to $\boldsymbol{T}_k$ and the variable contributions to the trends are shown in the corresponding block loadings $\boldsymbol{P}_1$, $\boldsymbol{P}_2$,… to $\boldsymbol{P}_k$. The block weights $\boldsymbol{w}$ shows the relative contribution of each block to the "global" trend of the data shown in the $\boldsymbol{T}_{\text{sup}}$.

There are several multiblock PCA algorithms reported in the literature, which differ by whether the blocks scores or block loadings are normalised and how this is done. A comprehensive review of these algorithms can be found in (Smilde et al. 2003). In this paper a modified consensus PCA algorithm, which normalised the block loadings and under the name of CPCA-W (Westerhuis et al. 1998), is used and other multiblock PCA models can also be applied in the same way. A brief description of the algorithm of the CPCA-W is given below:

(1) choose a sensible start $\boldsymbol{t}_{sup}$ with the constraint of $\|\boldsymbol{t}_{sup}\| = 1$, in this study we use the eigenvector with the largest magnitude eigenvalue of $\boldsymbol{X} \cdot \boldsymbol{X}^T$ as the start $\boldsymbol{t}_{\text{sup}}$.
(2) $\boldsymbol{p}_i = \boldsymbol{X}_i^T \cdot \frac{\boldsymbol{t}_{\text{sup}}}{\|\boldsymbol{X}_i^T \cdot \boldsymbol{t}_{\text{sup}}\|}$, for all the $i = 1,2,\ldots k$
(3) $\boldsymbol{t}_i = \boldsymbol{X}_i \cdot \boldsymbol{p}_i$, for all the $i = 1,2,\ldots k$
(4) $\boldsymbol{T} = [\boldsymbol{t}_1 \ldots \boldsymbol{t}_k]$
(5) $\boldsymbol{w} = \boldsymbol{T}^T \cdot \boldsymbol{t}_{\text{sup}}$
(6) $\boldsymbol{t}_{\text{sup}} = \boldsymbol{T}^T \cdot \boldsymbol{w}$
(7) $\boldsymbol{w} = \frac{\boldsymbol{w}}{\|\boldsymbol{t}_{\text{sup}}\|}$
(8) $\boldsymbol{t}_{\text{sup}} = \frac{\boldsymbol{t}_{\text{sup}}}{\|\boldsymbol{t}_{\text{sup}}\|}$
(9) $\boldsymbol{p}_i = \boldsymbol{X}_i^T \cdot \boldsymbol{t}_{\text{sup}}$, $\boldsymbol{E}_i = \boldsymbol{X}_i - \boldsymbol{t}_{\text{sup}} \cdot \boldsymbol{p}_i$
(10) $\boldsymbol{X}_i = \boldsymbol{E}_i$, for all the $i = 1,2,\ldots k$ and $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2 \ldots \boldsymbol{X}_k]$, return to (1) and repeat until the algorithm converges.

It has been proved that the super scores of CPCA-W is in fact equivalent to those applying PCA on $\boldsymbol{X}$ directly which is under the name of Sum-PCA (Smilde et al. 2003), the earliest form of multiblock PCA. The difference between the two methods is that CPCA-W also provides blocks scores and loadings which Sum-PCA lacks. Nevertheless, this suggests that CPCA-W and PCA are essentially the same, the only real difference is that the PCA has been performed on a re-arranged data. The properties of PCA can be applied to CPCA-W directly.

When the multiblock PCA was applied to the data with two factors, the data were re-arranged according to the experimental design. An example of a 2×2 experiment design is given in Fig. 2b. Two test matrices are needed, one matrix for one testing factor while put the other factor as the background factor. There is a concern about using multiblock methods in the way as described above though. Multiblock methods assume that the samples on the same row on each block are directly related and in conventional applications they are normally the same sample analysed by different analytical tools. This assumption does not strictly hold when it applies to multiblocking the data according to the experimental design. The samples on each row are the ones at the same level of the testing factor but different levels of the background factor. The consequence is that the results of multiblock PCA are no longer unique: with different orders of samples in each block, the results will not be exactly the same. However we have previously (Kassama et al. 2010) showed that even multiblock PCA gave different results when the order of the samples were changed, the general patterns (i.e. the separation between classes) were still well preserved should it did exist. A thorough permutation tests were conducted by permuting the order of samples a large number of times. The results

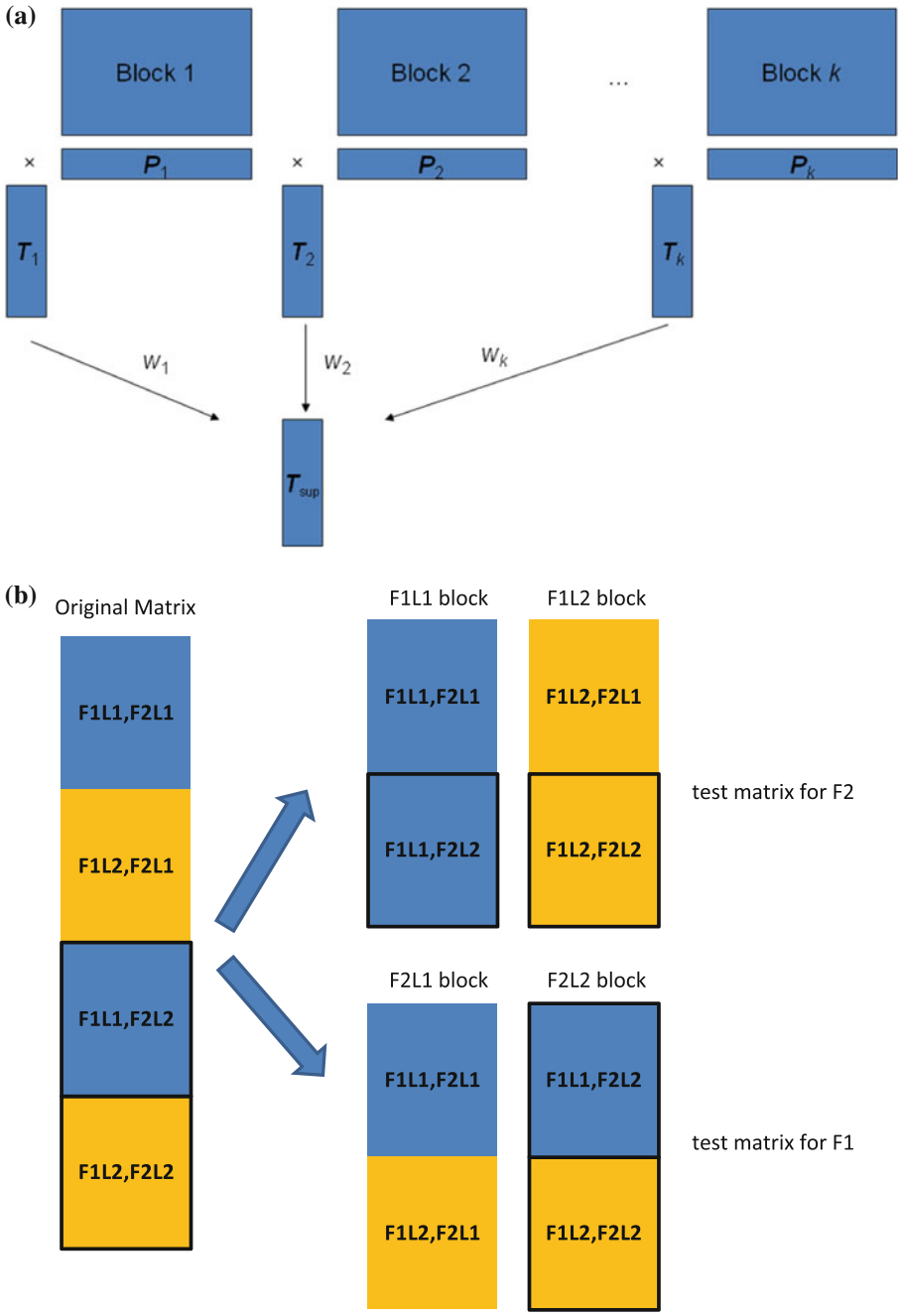

Fig. 2 **a** a graphic illustration of multiblock PCA model; **b** a graphic illustration of re-arranging data according to the experimental design for the multiblock PCA analysis

suggested that when the permutations were conducted with-in each class, the results of CPCA-W were highly similar to each other, both in super scores level and block scores level. We also conducted the same permutation test on the simulated data and have reached the same conclusion (data not shown). The reason is that even the samples on the same row, different blocks are at the different levels of the background factor, the change caused by that factor is effectively a baseline change, regardless how much difference were there between its different levels, a mean centring can effectively remove it and the influence of the testing factor remains as the main source of variance. Considering that in exploratory data analysis, precise modelling is generally not of the main concern, small disparities can be tolerated should the general pattern remains the same. Should it is needed to confirm the pattern obtained from the multiblock PCA is reproducible, one can perform a permutation test as described in (Kassama et al. 2010) and compare the pattern of the scores of different permutations by using a multivariate pattern comparison metric such as Procrustes distance (Gower and Dijksterhuis 2004). One can conclude that the observed

pattern is genuine should the Procrustes distances are consistently low.

2.4 ANOVA-PCA

The main idea of ANOVA-PCA is that it decomposes the data matrix into a sum of series mean matrices which contain the mean of each level of one particular factor. To study the influence of one particular factor, the corresponding mean matrix is superimposed on the residual matrix which is the data matrix after the means of all known factors been subtracted from it and subject it to PCA. For example the model of ANOVA-PCA on the data matrix $\boldsymbol{X}$ obtained from a 2 factors ($F1$ and $F2$) experimental design is defined as:

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\mu}_{F1} + \boldsymbol{\mu}_{F2} + \boldsymbol{\mu}_{F1F2} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\mu}$ is the grand mean matrix which has the same size as $\boldsymbol{X}$ with each row equals to the mean of $\boldsymbol{X}$, $\boldsymbol{\mu}_{F1}$, $\boldsymbol{\mu}_{F2}$ and $\boldsymbol{\mu}_{F1F2}$ is the mean matrices of factor A, B and the interaction.

Based on this model, each sub matrix is obtained in a sequential manner. Firstly the grand mean is subtracted from the $\boldsymbol{X}$.

$$\boldsymbol{X}_1 = \boldsymbol{X} - \boldsymbol{\mu}$$

The $\boldsymbol{\mu}_{F1}$, $\boldsymbol{\mu}_{F2}$ and $\boldsymbol{\mu}_{F1F2}$ are calculated from $\boldsymbol{X}_1$. In $\boldsymbol{\mu}_{F1}$,each row is the mean for the corresponding level of factor $F1$, e.g. given a data matrix obtained from a 2 × 2 experiment, first 200 are the samples of level 1 of the factor $F1$ and the next 200 are the level 2 of the same factor, then in $\boldsymbol{\mu}_{F1}$ the first 200 rows are all the same which is the mean of the first 200 rows and the second 200 rows are the mean of the second 200 rows. Another similar calculation is performed to obtain $\boldsymbol{\mu}_{F2}$ which contains the mean of the corresponding levels of factor $F2$. In $\boldsymbol{\mu}_{F1F2}$ there will be 4 different means, for the rows of the 4 different combinations of the 2 factors and their levels (assuming a 2 × 2 experiment design). More mean matrices can be calculated if there are more influential factors to be examined.

Finally, the residual matrix $\boldsymbol{\varepsilon}$ is obtained by subtracting the 3 matrices $\boldsymbol{\mu}_{F1}$, $\boldsymbol{\mu}_{F2}$, and $\boldsymbol{\mu}_{F1F2}$ from $\boldsymbol{X}_1$

$$\boldsymbol{\varepsilon} = \boldsymbol{X}_1 - \boldsymbol{\mu}_{F1} - \boldsymbol{\mu}_{F2} - \boldsymbol{\mu}_{F1F2}$$

Two test matrices, $\boldsymbol{X}_{F1}$ and $\boldsymbol{X}_{F2}$, were then calculated by adding $\boldsymbol{\varepsilon}$ back to the corresponding mean matrix:

$$\boldsymbol{X}_{F1} = \boldsymbol{\mu}_{F1} + \boldsymbol{\varepsilon} \text{ and } \boldsymbol{X}_{F2} = \boldsymbol{\mu}_{F2} + \boldsymbol{\varepsilon}$$

These two test matrices are then subject to PCA separately to reveal the influence of $F1$ and $F2$. If needed, a interaction matrix $\boldsymbol{X}_{F1F2} = \boldsymbol{\mu}_{F1F2} + \boldsymbol{\varepsilon}$ can also analysed to see if there were any significant interactions between the two factors.

In difficult problems, the variance left in $\boldsymbol{\varepsilon}$ can be equal or higher than that in the mean matrices and still hinder the chance of discovering the significant pattern. An extension to the ANOVA-PCA was made by Climaco-Pinto et al. (2009) to improve the sensitivity of detecting significant patterns. It is achieved by subtracting first few PCs from $\boldsymbol{\varepsilon}$ matrices:

$$\boldsymbol{\varepsilon}_r = \boldsymbol{\varepsilon} - T \cdot P^T$$

Then form a new $\boldsymbol{X}_{\text{new}}$

$$\boldsymbol{X}_{new} = \boldsymbol{\mu} + \boldsymbol{\mu}_{F1} + \boldsymbol{\mu}_{F2}\boldsymbol{\mu}_{F1F2} + \boldsymbol{\varepsilon}_r$$

ANOVA-PCA is then performed on the $\boldsymbol{X}_{\text{new}}$. With sufficient number of PCs been subtracted from $\boldsymbol{\varepsilon}$ the separation caused by the testing factor can eventually be revealed on the first a few PCs. The caveat is that with enough PCs subtracted $\boldsymbol{\varepsilon}$, the separations between the different levels of the testing factor can *always* be observed on the first PC, even when there were no real differences. Take an extreme example, if all possible PCs in $\boldsymbol{\varepsilon}$ were subtracted, $\boldsymbol{\varepsilon}_r$ will become a 0 matrix and the PCA would be applied to the mean matrix itself, the scores of all the samples belonging to the same class will be all exactly the same, equal to the mean of the corresponding class. Since there will always be some minor differences between the means of the classes, separation can always be observed. Thus the number of PCs to be subtracted from $\boldsymbol{\varepsilon}$ need to be determined by using permutation tests. In the permutation test, a series of null mean matrices were obtained using randomly selected groups of samples. Assuming n PCs are needed to be subtracted from $\boldsymbol{\varepsilon}$ to able to enable ANOVA-PCA to show the separations using the original labelling, if similar separation, which is measured using the Mahalanobis distance (Brereton 2003) between the centres of different classes, can also be seen using the null mean matrices with n (or less) PCs subtracted from $\boldsymbol{\varepsilon}$, the observed separation using the original labelling is considered as a false discovery. In this study, if there were no separation on either of the first 3 PCs, we subtract sufficient number of PCs from $\boldsymbol{\varepsilon}$ until the separation can be seen on one or a few of the first 3 PCs, providing more than 95% (i.e. $p < 0.05$) of the corresponding permutation tests requires more PCs to be subtracted from $\boldsymbol{\varepsilon}$.

2.5 Model validation

Although multiblock PCA does not use the labelling information explicitly, the data itself have been re-partitioned based on their labels. Thus it is necessary to assess the risk of over fitting when such partitioning of the data is employed. A cross-validation approach coupled with a permutation test with a simple distance classification is employed for this purpose. For CPCA-W models, the

validation is carried out according to the following procedure:

(1) Remove one sample from each block and build a CPCA-W model on the remaining samples and keep the first 3 PCs.
(2) Calculate the centre and covariance matrix of each class (a unique status of factors and their corresponding levels) within each block.
(3) Project the samples being left out in step (1), i.e. the 'test samples', into the corresponding block via the loadings and the Mahalanobis distance (Brereton 2003) from the block scores of the test sample to each class centre within the block that the test sample was originally assigned to. The predicted class membership of the test sample as the one with the shortest distance.
(4) Repeat steps (1) to (3) and remove a different set of samples each time for testing until all the samples have been left out once. The correctly classified rates (CCR) of the class membership are then calculated.
(5) Compare the CCRs obtained via cross-validation as described above against a null distribution of CCRs obtained from a permutation test (Good 2005). The permutation test was performed by randomly shuffling the order of data, then rearranging the data into k (k is the number of the blocks in the data before the permutation) arbitrary blocks with the same number of samples for each block as was applied to the original model. The operations described from step (1) to (4) were applied to the permuted data, the CCRs of the class membership predictions were then recorded. This test was repeated 1,000 times, each time with different randomly shuffled data. The CCRs obtained from these tests formed the null distributions and were used to assess the significance level of the CCRs obtained from the original data. If the block and/or class structure is authentic, it can be expected that the CCRs from the original data will be much better than the majority of those from the null distribution. In addition, we also tested the classification error of the auto-predictions of the *permuted* data sets. That is to project all the samples of the *training* set into to the multiblock model and assign the class membership of these samples according to step (3). This is to assess the inherit risk of over-fitting of the multiblocking. A classification error close to that of an random classifier classification indicates that the risk of over-fitting is low.

A similar cross-validation was also performed on the results of ANOVA-PCA with Climaco-Pinto's procedure. Each data set was randomly split into 10 parts, one part (40 samples, 10 for each class) was removed from the data and used as the test set; the remaining samples were used as the training set. The number of PCs to be removed from $\boldsymbol{\varepsilon}$ was determined using the training set only and the class membership of the test set was determined using the distance classification approach as above. This procedure was repeated 10 times, until every sample had been included in the test set once. It is important to note that although each test matrix have all the samples from the 4 classes, the test matrix itself is designed to reveal the separation of 2 classes, i.e. the two levels of the testing factor. Therefore in the validation, we consider the 2 levels of the factor which the testing matrix was *not* designed to test as one, e.g. in the $\boldsymbol{X}_{F1}$, the samples of ($F1L1$, F2L1) and ($F1L1$, $F2L2$) are considered as belonging to the same class.

3 Results and Discussion

3.1 Results on the real MS-based metabolomics data

The results of the three PCA models are presented in Fig. 3. All the three models had been able to reveal the difference between three bacterial strains which is the main source of variation in this data set (the results relates to the separation between bacterial strains of CPCA-W and ANOVA-PCA are not shown for brevity reason). However it is difficult to conclude whether there was any significant difference between using the two different extraction methods based on the results of classical PCA. By contrast, such difference becomes very clear when a CPCA-W model was employed. The super scores clearly showed there was a significant difference between the two extraction methods and, by inspecting the blocks scores, albeit such difference was observed on all 3 blocks, it was not consistent across 3 different strains. In the T block, the two classes were less well separated compared to the other 2 blocks. This clearly indicates that different strains have different responses to different extraction methods, i.e. there were interactions between the two factors. Bar charts of a few selected significant metabolites from the loadings plots of the CPCA-W confirmed this (Kassama et al. 2010) Such information is not easily discernable in the classical PCA results. ANOVA-PCA was also able to show the difference between the extraction methods in its corresponding testing matrix. This example has demonstrated the clear advantage of using CPCA-W or ANOVA-PCA over classical PCA on the data with interacting factors.

3.2 Results on the simulations: Modelling performance

Due to the large number of figures generated by the experiment, the scores plots of each type of interactions obtained by different PCA models are presented separately

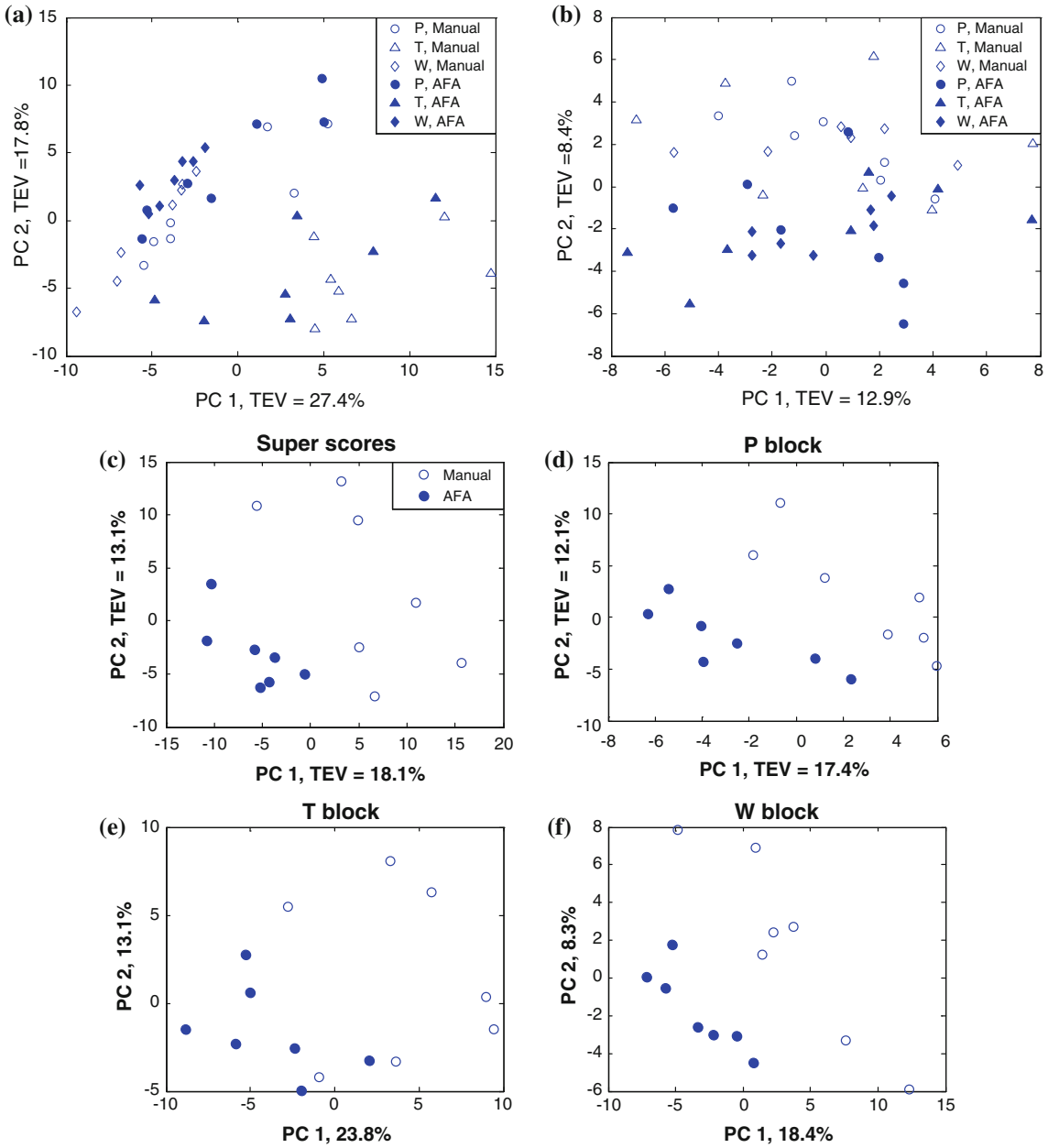

Fig. 3 The results of the real metabolomics data: **a** classical PCA scores plot; **b** ANOVA-PCA scores plot on the extraction methods testing matrix; **c** CPCA-W super scores plot on the 3 blocks data; **d** CPCA-W block scores on the P block; **e** CPCA-W block scores on the T block; **f** CPCA-W block scores on the W block

in the supplementary information (Figures S-1 to S-18). From the illustrations in Fig. 1, it is already clear that when there are interactions between factors, PCA will not always be able to reveal the differences between all the classes and some of the classes overlap each other due to the interactions such as (*F*1*L*1, *F*2*L*2) and (*F*1*L*2, *F*2*L*2) in Interaction 4. Nevertheless, PCA can still reveal some separations when the data are "clean" enough and the full picture has to be inferred from the results in the light of prior knowledge (e.g., the number of expected classes). By partition the data into several matrices according to the experiment design, ANOVA-PCA and multiblock PCA should be able to reveal the difference caused by the influential factors, even when there are interactions between them. With the ratio of the interesting latent factor and the competing latent factor set to 1:1, all three methods have had little difficulty showing the clustering trend. PCA was able to show 4 distinctive clusters in Interactions 1, 2 and 3; while 3 clusters in Interaction 4 and 2 clusters in Interaction 5 were revealed. It is interesting to see that ANOVA-PCA

failed to reveal any difference caused by the underlying influential factors in Interaction 5 in its two factor testing matrices $\boldsymbol{X}_{\boldsymbol{F}1}$ and $\boldsymbol{X}_{\boldsymbol{F}2}$, even with 1:1 latent factor ratio. This is due to the fact that the trend of changing both factors reversed while the other factor changes its level and cause the mean of each factor at each particular level close to the global mean which is 0 after the mean centre step. However, separations can be seen in the interaction matrix $\boldsymbol{X}_{F1F2}$. In fact, the results from $\boldsymbol{X}_{F1F2}$ were almost exactly the same as those from the PCA when the number of competing latent factor was low. The reason is that when the two factor mean matrices are close to *0*, $\boldsymbol{X}_{F1F2} = \boldsymbol{\mu}_{F1F2} + \quad \boldsymbol{\varepsilon} = \boldsymbol{\mu}_{F1F2} + (\boldsymbol{X}_1 - \boldsymbol{\mu}_{F1} - \boldsymbol{\mu}_{F2} - \boldsymbol{\mu}_{F1F2}) \approx \boldsymbol{X}_1$ and $\boldsymbol{X}_1$ was the matrix used by PCA. Nevertheless, the disparity between the results of the factor testing matrices and the interaction matrix clearly suggested the existence of interaction between the two factors and also the type of the interaction. In addition, for Interactions 2, 3 and 4, the separation between *F*2 is generally worse than that of *F*1. This is because that the trend of changing of *F*2 reversed when *F*1 changes its level thus the differences caused by *F*2 at different levels of *F*1 are counter-acting each other when calculating the mean matrix $\boldsymbol{\mu}_{\boldsymbol{B}}$. This suggests that when such interaction happens ANOVA-PCA may lose its sensitivity detecting the significant pattern, although this problem can be circumvented by implementing Climaco-Pinto's procedure (see below). By contrast, CPCA-W has no difficulties in revealing the changes caused by each of the influential factors, both in the super scores and each of the corresponding block scores. Unlike ANOVA-PCA, the problem of reversing trend can be easily solved by reversing the sign of either the corresponding loadings or the scores and makes CPCA-W naturally robust against such problems. A changing of the direction of the block loadings or scores provides an evidence of such interaction existing in the data.

With the addition of more competing latent factors, the most notable change in PCA is that it becomes increasingly harder to find discriminant PC(s). The first few PCs were generally dominated by the competing latent factors while the trend of the latent factor of interest can only be found in one or a few minor PCs (e.g. PC 7 and 8 in interaction 1 with 10 competing latent factors added, Figure S-6). It is also impossible to predict which PC would be the most discriminant one as it depends on which data set was used it can be any PC between 2nd to $k + 1$th PC (k is the number of competing latent factors added). The separation between different classes also becomes weaker with more competing latent factors added in. For example, in Interaction 1 with 10 competing latent factors are added, only the difference between the most different classes (i.e. (*F*1*L*1, *F*2*L*2) and (*F*1*L*2, *F*2*L*2)) can be observed in a lower PC score while other less well separated classes resulted in from most to completely overlapping. In the results of ANOVA-PCA, the "position" of the discriminant PC was also affected by the addition of competing latent factors (data not shown). For example, the most discriminant PC in Interaction 4 data with 10 competing latent factors appeared to be PC 10 for *F*1 and PC 12 + PC 14 for *F*2. However, the separation shown in the ANOVA-PCA is generally much better than those in the PCA (except Interaction 5 which its factor testing matrices could not reveal any separations while the interaction matrix showed very similar results to that of PCA) and the loadings of the discriminant PCs are much less "contaminated" by the variables coming from the competing latent factors. After applying Climaco-Pinto's procedure clean separations can always be observed on the first 2 PC after 1–13 PCs removed from $\boldsymbol{\varepsilon}$, more competing latent factors existed, more PCs were needed to be removed from $\boldsymbol{\varepsilon}$. For the interaction 5, with more competing latent factors added, ANOVA-PCA with Climaco-Pinto's procedure applied could still show the same pattern on the first 2 PCs while PCA could not do so with more than 5 competing latent factors added (Figure S-18). It is worth noting that, the factor with its trend reversed while the other factor changing its level, i.e. *F*2 in interaction 2–4 generally require more PCs, sometimes even more than the number of actual latent factors existed (e.g. a data with 10 competing latent factors added should have the actual number of latent factors of $1 + 10 = 11$) to be removed from $\boldsymbol{\varepsilon}$ to enable ANOVA-PCA show its effect. This confirmed that averaging two contradicting trends had diminished the difference between the different levels of the studying factor as stated above and sometimes such difference after averaging could be smaller than the background variation. Another observation need to be aware of is that in interaction 4, the subtle detail that (*F*1*L*1, *F*2*L*2) and (*F*1*L*2, *F*2*L*2) were the same could not be seen in the *F*1 test matrix (Figures S-13 to S15). This is because the difference shown in the figure is the averaged difference of the two levels of *F*1 across two levels of *F*2 and the information about the details of *F*1 on each level of *F*2 has been lost through averaging. CPCA-W appeared to be the most successful model which is seemingly completely resistant to the addition of competing latent factors. Even with 10 competing factors, the first PC remains the most discriminant one and the only difference was the variance explained by that PC proportional decreased when more competing latent factors were included; this is to be expected. This makes the model of CPCA-W the easiest one to build and interpret.

We conclude that the reason of such seemly flawless success of CPCA-W on these simulated data is that by rearranging the data into blocks, the relative ratio of the variance of the useful latent factor is increased. A graphic

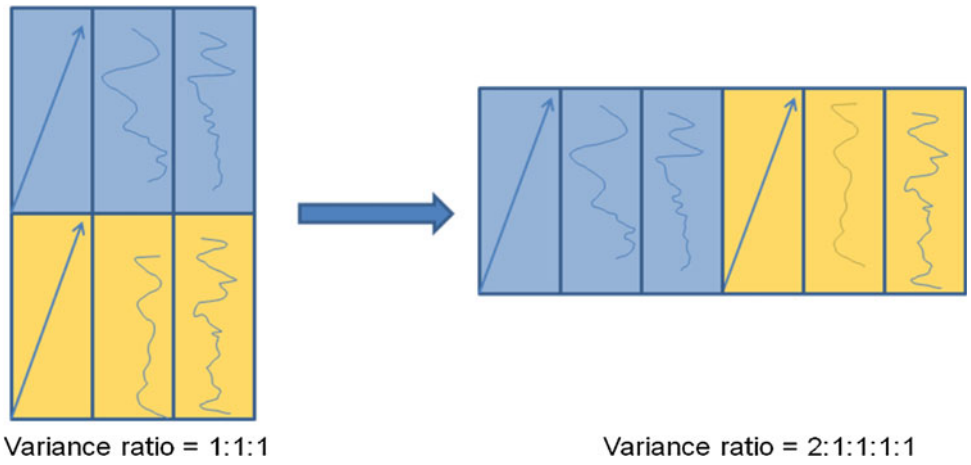

Fig. 4 Multiblocking increases the relative variance ratio of the trend of interest over competing ones. Assuming the monotonic increasing trend is the trend of interest, represented by the *arrows*. If there were no interactions between the two factors, before the multiblocking the variance ratio of the latent factor of interest over the other 2 competing ones is 1:1:1; however, after the multiblocking this becomes 2:1:1:1:1. If the interactions does exist, the relevant ratio may vary albeit one can still expect the increase of such ratio

illustration with a 2 × 2 experimental design which has 2 competing latent factors is given in Fig. 4. In this example, by re-arranging the data into 2 blocks, according to the experiment design, the number of the variables was doubled and the ones which affected by the influential factor will follow the similar (or the opposite) trend after the multiblocking. Thus the variance caused by the influential factor is increased accordingly. Other variables which are not affected by the influential factor are not likely to be able to benefit from the re-arrangement of the data in the same way, this results in increasing the number of the *uncorrelated* competing latent factors with similar variance of each. By increasing the variance caused by the factor of interest whilst keeping the others more or less the same, the relative ratio of the variance of interest over competing ones is hereby increased and the chance of it to be discovered by the component analysis is also increased. This can be demonstrated by employing a 2 × 3 experiment design. The data were generated as described in 2.1 except that each competing latent component possess 20 variables instead of 10, thus the variance of each competing latent factor is twice as much as the one of interest after autoscaling. *F*1 now has 3 levels and *F*2 remains the same with 2 levels. The relative distribution of the means of each status is given in (Fig. 5). The results of the CPCA-W with 10 competing latent factors is presented in (Fig. 6). CPCA-W has clearly shown the difference caused by *F*2 which has been modelled in a 3 blocks model, both in the super scores and the block scores. With 10 competing latent factors, the first PC remains as the most discriminant one and even a subtle detail that the two levels of *F*2 were less well separated in *F1L2* block than the other 2 blocks has been shown (Fig. 6a–c). However, CPCA-W failed to discover the difference caused by *F1* in its two block model

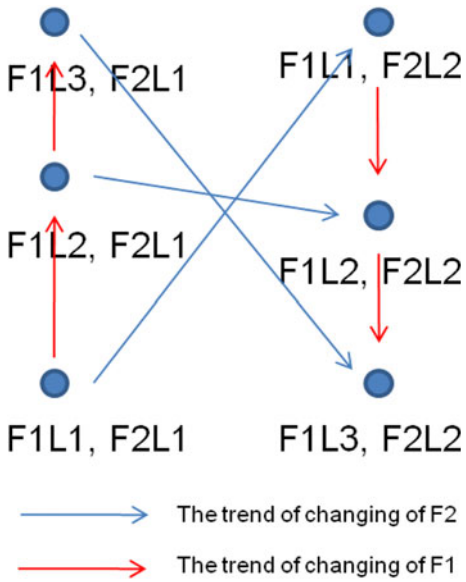

Fig. 5 The 2 × 3 experiment design

completely (Fig. 6d, e). No discriminant PC can be found at all. Furthermore, it is interesting to compare the results with those of PCA, the results of PCA performed on the same experiment design is presented in Fig. 7. In fact, with the number of variables in each latent factor doubled, it is now possible to find a discriminant PC (PC 11) to show 2 well separated classes with an additional one in the between for each factor, just as expected (Fig. 7a). Such discriminant PC cannot be found if the number of variables in each competing latent factor was set to 10 (Fig. 7c). The same observations were found from other types of interactions as well (data not shown). The reason becomes obvious by comparing the loadings plots (Fig. 7b, d) of

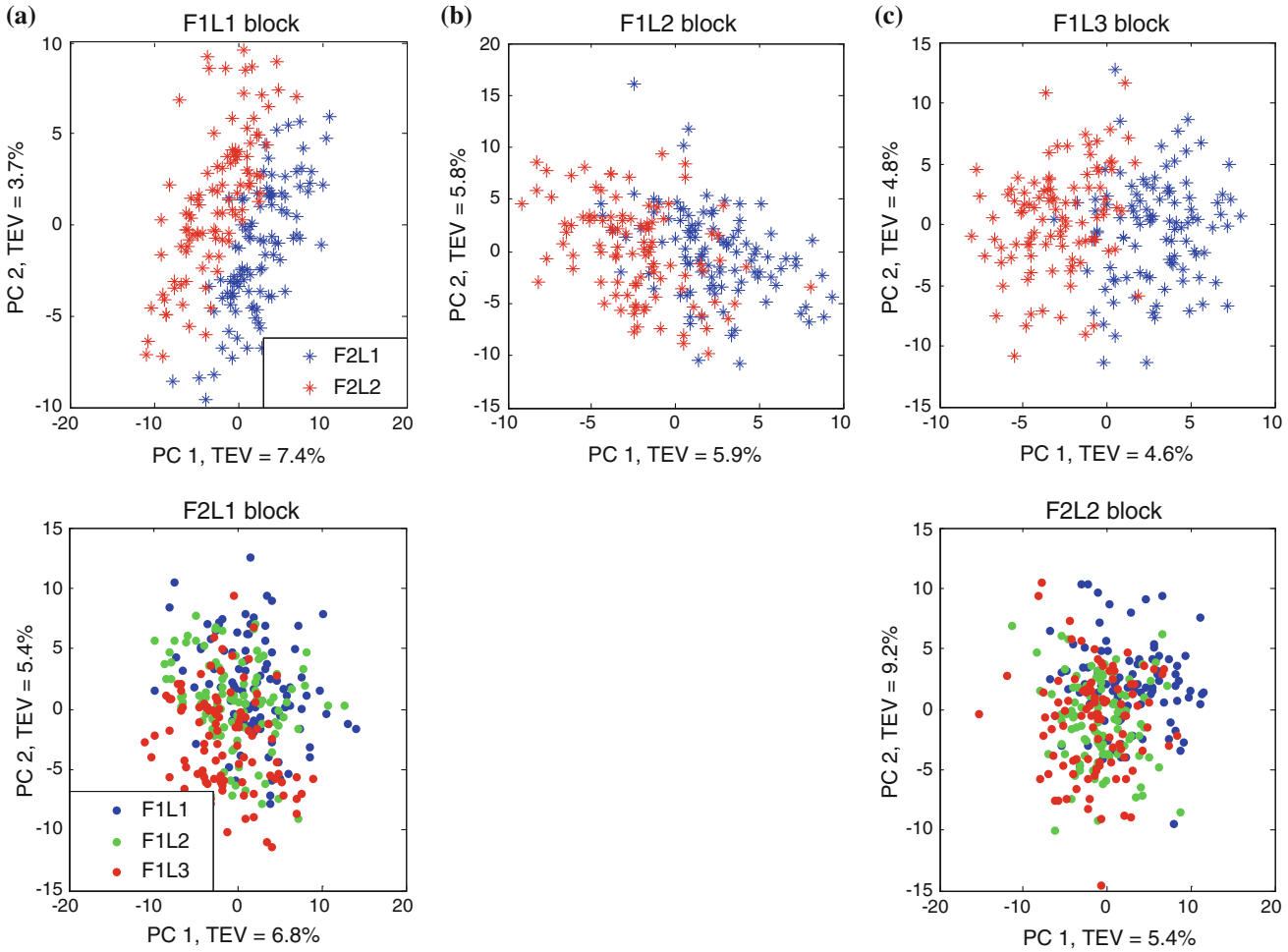

Fig. 6 The block scores plot of the CPCA-W performed on the 2 × 3 experiment design. F1 has 3 levels while F2 has 2 levels

these two types of data. In the data with the competing latent factors containing twice the number of variables, the discriminant PC is clearly dominated by the first 10 variables, i.e. the ones of interest while all other variables have had little contributions. By contrast, when PCA was applied to the data with the same design but each competing latent variable having 10 variables, the loadings of PC 6, the one with the variables of interest having the highest weight, clearly shows that there were no less than 3 competing latent factors having similar or even higher contribution to this PC. This demonstrates the fact that when the underlying latent factors having similar variance, the latent factor inferred by PCA can be a mixture of several of them rather than one unique latent factor for one PC even there were little correlation between them (since the competing latent factors were generated from different random vectors little correlation can be expected between them); i.e., the subspace of PCA is a rotated space of the original one. Since CPCA-W share the same properties as PCA, the advantage of multiblocking turns into a disadvantage when some of the competing latent factors have approximately k times variance (k is the number of levels of the underlying factor) as the latent factor of interest has. Although the chance of this problem happening is probably very small in real metabolomics applications and can be circumvented by increasing the level of the factor of interest to further increase the variance of this influential factor and hopefully it will out weight the variance of the competing latent factors.

3.3 Model validation

The validation results of CPCA-W are listed in Table 1. For brevity reason, only the results of the 1 + 10 data sets (the most difficult ones) of the different types of interactions are listed. The CCRs obtained from the permuted data sets are all close to an expected results from a 2-class purely random classifiers which is 50%, regardless how many competing latent factors were in the data. This was not the case for Interaction 4 which is a special case and will be discussed separately, the CCRs from the data without permutation are much higher. With 1 to 3

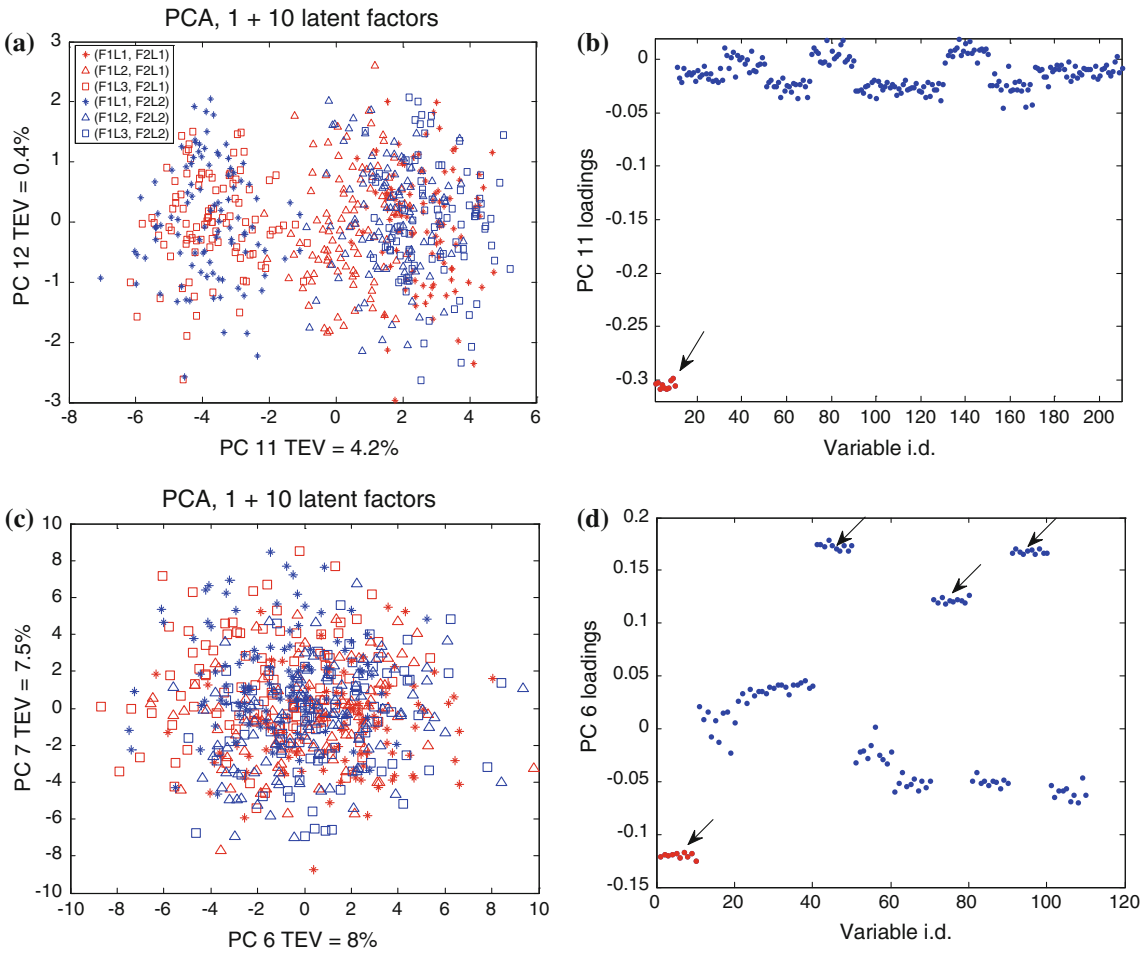

Fig. 7 PCA scores and loadings plot on the 2 × 3 experiment design: **a** and **b** are the scores and loadings plots respectively of PCA performed on the data with each competing latent factor having 20 variables respectively; **c** and **d** are the scores and loadings plot respectively of PCA performed on the data with each competing latent factor having 10 variables

competing latent factors, perfect classifications were achieved and the CCRs were slowly decreasing when more competing latent factors were added into the data. Nevertheless, even with 10 competing latent factors the CCRs were above 90% except that of Interaction 1 block 1 which had a CCR of 80.8% which is still well above 50%. This means that the separation shown in the scores plots are genuine and multiblock PCA is highly resistant to the irrelevant competing latent factors.

Interaction 4 is a special case, there were no real separation between the class of (*F*1*L*1, *F*2*L*2) and (*F*1*L*2, *F*2*L*2) thus the expected CCR of an ideal model for *F*2 blocks is 75% (i.e., a perfect separation result from the *F*2*L*1 block and a random classification result with 50% of accuracy from the *F*2*L*2 block). With 1 to 2 competing latent factors the cross-validated CCRs are both 76% which is very close to the expected accuracy. However, such accuracy decreased rather quickly while more competing latent factors were included and with 10 competing latent factors, the CCR was reduced to 62.5%. This can be explained that only one block out of 2 has had real separation and the multiblock approach is designed to increase the chance of discovering the "common trend" between the blocks. If there were no real "common trend" between the blocks the multiblock approach will no longer be very effective.

The auto-prediction results of the permuted data sets are also presented in Table 1. Similar to those of the cross-validated CCRs from the permuted data, the CCRs of the auto-predictions from the permuted data were appeared not affected by the number of competing latent factors and the CCRs varied from 50.3 to 57.5%, although seemingly marginally higher than the cross-validated results they are still significantly lower than those from the data without permutation. This suggests that the risk of over-fitting introduced by the multiblocking is low and probably can be considered as an unsupervised method.

The validation results of ANOVA-PCA are listed in Table 2. With Climaco-Pinto's procedure applied, ANOVA-PCA had achieved similar validated compared to those of CPCA-W except interaction 5. For interaction 5,

Table 1 Results of CPCA-W validation

	Interaction 1		Interaction 2		Interaction 3		Interaction 4		Interaction 5	
	Block 1	Block 2	Block 1	Block 2	Block 1	Block 2	Block 1	Block 2	Block 1	Block 2
Cross-validated accuracy without permutation	80.8%	98.8%	95.3%	97.8%	95.3%	97.8%	99.0%	62.5%	94.8%	90.5%
Cross-validated accuracy with permutation	50.5%, $p < 0.001$	50.5%, $p < 0.001$	53.8%, $p < 0.001$	50.2%, $p < 0.001$	51.5%, $p < 0.001$	49.5%, $p < 0.001$	54.0%, $p < 0.001$	53.5%, $p = 0.005$	46.8%, $p < 0.001$	45.0%, $p < 0.001$
Auto-prediction accuracy with permutation	54.5%, $p < 0.001$	54.3%, $p < 0.001$	57.5%, $p < 0.001$	51.0%, $p < 0.001$	50.3%, $p < 0.001$	54.5%, $p < 0.001$	52.3%, $p < 0.001$	53.8%, $p = 0.007$	51.2%, $p < 0.001$	53.8%, $p < 0.001$

The p-value is defined as the number of the permutations which obtained better results than the original labelling divided by the total number of the permutations. A $p < 0.001$ indicates that none of the permutations obtained better results than the original labelling

Table 2 Results of ANOVA-PCA with Climato-pinto procedure validation

	Interaction 1		Interaction 2		Interaction 3		Interaction 4		Interaction 5		
	$\boldsymbol{X}_{F1}$	$\boldsymbol{X}_{F2}$	$\boldsymbol{X}_{F1}$	$\boldsymbol{X}_{F2}$	$\boldsymbol{X}_{F1}$	$\boldsymbol{X}_{F2}$	$\boldsymbol{X}_{F1}$	$\boldsymbol{X}_{F2}$	$\boldsymbol{X}_{F1}$	$\boldsymbol{X}_{F2}$	$\boldsymbol{X}_{F1F2}$[a]
Cross-validated accuracy	92.1%	85.5%	91.9%	89.5%	90.4%	85.9%	96.9%	92.1%	50.5%	48.9%	90.3%
p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.499	0.505	<0.001

[a] The class labelling of $\boldsymbol{X}_{F1F2}$ is to consider ($F1L1$, $F2L1$) and ($F1L2$, $F2L2$) as one class while ($F1L1$, $F2L2$) and ($F1L2$, $F2L1$) as another class

the interaction matrix $\boldsymbol{X}_{F1F2}$ can reveal two clearly two separated groups: one group is a mixture of ($F1L1$, $F2L1$) and ($F1L2$, $F2L2$) and the other group is a mixture of ($F1L1$, $F2L2$) and ($F1L2$, $F2L1$). When the class labelling was made according to this, the validation results were satisfactory. The validation results suggested that the separation showed in ANOVA-PCA with Climaco-Pinto's approach were also genuine albeit the interpretations might not always be straightforward.

4 Conclusion

Through a series of carefully designed simulations, this investigation has demonstrated the capability of multiblock PCA for analysing multiple factor/multiple level data with a balanced experiment design. This success was due to the fact that by re-arranging the data into a series of blocks according to the experiment design, the relative ratio of the variance caused by the factor under study (if it does exist) over others is increased and thus increases the chance of being discovered by the variance driven component analysis methods such as PCA. It has also been demonstrated that such methodology could fail under some special circumstances which is when there are 1 or more competing latent factors having approximately k times variance as the latent factor of interest has. Although increasing the level of the factor under study can usually help circumventing such problem. ANOVA-PCA attempts to tackle the same problem through a different route. In ANOVA-PCA, a hypothetical deviation over experimental error approach is employed, the deviation caused by the factor to be studied was presented in a form of a mean matrix. This mean matrix was subsequently superimposed over the residual matrix which represents the variance which was not related to the studying factor. If the deviation caused by the factor was significant, the separation would be revealed in one or a few major PCs. In the case that the influence of the factor is weak or irrelevant variations were too strong, Climaco-Pinto's procedure can be applied to increase its sensitivity. However, should Climaco-Pinto's procedure be used, the risk of over-fitting was also increased and a permutation test for validation must be used. In our simulations, the multiblock approach has better sensitivity at detecting the significant pattern than the ANOVA-PCA without using Climaco-Pinto's procedure. When Climaco-Pinto's approach was applied, ANOVA-PCA can achieve similar sensitivity. It worth noting that Climaco-Pinto's approach must be done with an appropriate permutation tests while the multiblock approach does not need it as we have showed that the inherit risk of over-fitting of the multiblock approach is very low.

Metabolomics studies are almost always involved with investigating multiple influential factors and more often than not such knowledge is known to the researchers. Thus the experiment design are normally planned accordingly to enhance the chances of revealing such influence factors. The multiblock approach discussed in this article has demonstrated that such information can also be actively incorporated into the data analysis step and greatly improve the interpretability of the model. Any metabolomics study investigating the changing caused by known or hypothetical influential factors with a balanced experiment design suits this type of approach well.

The main limitation of this multiblock approach in its current form is that it can only investigate 2 influential factors at the moment. How this methodology could be extended to cope with 3 or even more factors is a very interesting area for further research. A possible approach is to perform a hierarchical type of multiblock re-arranging and analyse two influential factors at one time while all other factors being baseline. Another drawback is that multiblock PCA made no direct attempts to estimate the size of interactions between factors directly, the interaction effect is shown in the form of subtle differences in the patterns shown in different blocks, both in the block scores and the corresponding loadings. While by using ANOVA-PCA or ASCA such information can be obtained more easily from the interaction matrix, in a single unified frame.

Acknowledgments We want to thank Dr. Yankuba Kassama for providing his metabolomics data. We also want to thank the anonymous reviewers for many constructive suggestions. YX and RG acknowledge the Symbiosis-EU (www.symbiosis-eu.net) project (No. 211638) financed by the European Commission under the 7th Framework programme for RTD. The information in this document reflects only the authors' views and the Community is not liable for any use that may be made of the information contained therein.

References

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203–219.

Brereton, R. G. (2003). *Chemometrics: Data analysis for the laboratory and chemical plant*. Chichester: Wiley.

Climaco-Pinto, R., Barros, A. S., Locquet, N., Schmidtke, L., & Rutledge, D. N. (2009). Improving the detection of significant factors using ANOVA-PCA by selective reduction of residual variability. *Analytica Chimica Acta, 653*, 131–142.

Ferreira, D. L. S., Kittiwachana, S., Fido, L. A., Thompson, D. R., Escott, R. E. A., & Brereton, R. G. (2010). Windows consensus PCA for multiblock statistical process control: Adaption to small and time dependent normal operating condition regions, illustrated by on-line high performance liquid chromatography of a three stage continuous process. *Journal of Chemometrics, 24*, 596–609.

Good, P. I. (2005). *Permutation, parametric and bootstrap tests of hypotheses* (3rd ed.). New York: Springer.

Gower, J. C., & Dijksterhuis, G. B. (2004). *Procrustes problems*. Oxford: Oxford University Press.

Harrington, P. B., Vieira, N. E., Espinoza, J., Nien, J. K., Romero, R., & Yergey, A. L. (2005). Analysis of variance-principal component analysis: A soft tool for proteomic discovery. *Analytica Chimica Acta, 544*, 118–127.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*, 187–200.

Kassama, Y., Xu, Y., Dunn, W. B., Geukens, N., Anné, J., & Goodacre, R. (2010). Assessment of adaptive focused acoustics versus manual vortex/freeze-thaw for intracellular metabolite extraction from Streptomyces lividans producing recombinant proteins using GC-MS and multiblock principal component analysis. *Analyst, 135*, 934–942.

Manly, B. F. (2005). *Multivariate statistical methods: A primer*. London: Chapman & Hall.

Qin, S. J., Valle, S., & Piovoso, M. J. (2001). On unifying multiblock analysis with application to decentralized process monitoring. *Journal of Chemometrics, 15*, 715–742.

Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., Lamers, R.-J. A. N., van der Greef, J., & Timmerman, M. E. (2005). ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics, 21*, 3043–3048.

Smilde, A. K., Westerhuis, J. A., & de Jong, S. (2003). A framework for sequential multiblock component methods. *Journal of Chemometrics, 17*, 323–337.

Westerhuis, J. A., Kourti, T., & MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics, 12*, 301–321.

Xu, Y., Cheung, W., Winder, C. L., & Goodacre, R. (2010). VOC-based metabolic profiling for food spoilage detection with the application to detecting *Salmonella typhimurium* contaminated pork. *Analytical and Bioanalytical Chemistry, 397*, 2439–2449.

Zomer, S., Dixon, S. J., Xu, Y., Jensen, S. P., Wang, H., Lanyon, C. V., et al. (2009). Consensus multivariate methods in gas chromatographic mass spectrometry and denaturing gradient gel electrophoresis: MHC-congenic and other strains of mice can be classified according to the profiles of volatiles and microflora in their scent-marks. *Analyst, 134*, 114–123.