Metabolomics (2014) 10:375–385
DOI 10.1007/s11306-013-0616-8

ORIGINAL ARTICLE

Chemometrics models for overcoming high between subject variability: applications in clinical metabolic profiling studies

Yun Xu · Stephen J. Fowler · Ardeshir Bayat · Royston Goodacre

Received: 10 July 2013 / Accepted: 5 December 2013 / Published online: 11 December 2013

Abstract In human metabolic profiling studies, between-subject variability is often the dominant feature and can mask the potential classifications of clinical interest. Conventional models such as principal component analysis (PCA) are usually not effective in such situations and it is therefore highly desirable to find a suitable model which is able to discover the underlying pattern hidden behind the high between-subject variability. In this study we employed two clinical metabolomics data sets as the testing grounds, in which such variability had been observed, and we demonstrate that a proper choice of chemometrics model can help to overcome this issue of high between-subject variability. Two data sets were used to represent two different types of experiment designs. The first data set was obtained from a small-scale study investigating volatile organic compounds (VOCs) collected from chronic wounds using a skin patch device and analysed by thermal desorption-gas chromatography-mass spectrometry. Five patients were recruited and for each patient three sites sampled in triplicate: healthy skin, boundary of the lesion and top of the lesion, the aim was to discriminate these three types of samples based on their VOC profile. The second data set was from a much larger study involving 35 healthy subjects, 47 patients with chronic obstructive pulmonary disease and 33 with asthma. The VOCs in the breath of each subject were collected using a mask device and analysed again by GC–MS with the aim of discriminating the three types of subjects based on breath VOC profiles. Multilevel simultaneous component analysis, multilevel partial least squares for discriminant analysis, ANOVA-PCA, and a novel simplified ANOVA-PCA model—which we have named ANOVA-Mean Centre (ANOVA-MC)—were applied on these two data sets. Significantly improved results were obtained by using these models. We also present a novel validation procedure to verify statistically the results obtained from those models.

Keywords Metabolic profiling · Between-subject variability · Multilevel simultaneous component analysis · Multilevel partial least squares for discriminant analysis · ANOVA PCA · ANOVA simultaneous component analysis · ANOVA mean-centre · Breath analysis

Electronic supplementary material The online version of this article (doi:10.1007/s11306-013-0616-8) contains supplementary material, which is available to authorized users.

Y. Xu (✉) · R. Goodacre
Manchester Institute of Biotechnology & School of Chemistry, University of Manchester, Manchester, UK
e-mail: yun.xu-2@manchester.ac.uk

S. J. Fowler
Manchester Academic Health Science Centre, University of Manchester, Manchester, UK

S. J. Fowler
NIHR Respiratory and Allergy Clinical Research Facility, University Hospital of South Manchester, Manchester, UK

S. J. Fowler
Lancashire Teaching Hospitals NHS Foundation Trust, Preston, UK

A. Bayat
Plastic & Reconstructive Surgery Research, Manchester Institute of Biotechnology, University of Manchester, Manchester, UK

1 Introduction

In recent years untargeted metabolic profiling approaches have been successfully applied to many research areas such as bacterial discrimination, food spoilage detection and plant biology (Cheung et al. 2008; Biais et al. 2009; Xu et al. 2010).

Such methods also have great potential for clinical applications, such as diagnostics and biomarker discovery. However, one of the most common and difficult challenges in such studies is the magnitude of the between-subject variability, which is often the most dominating factor and often unfortunately masks potential differences associated with the factors of interest, typically disease-status or a response to an intervention (Penn et al. 2007; Assfalg et al. 2008). Contributors to the personal metabolic profile are many and include nutrition, health, environmental exposures, age, gender and ethnical background. Conventional data analysis tools such as principal component analysis (PCA) and partial least squares (PLS) (Wold et al. 2001; Brereton 2003) are not effective where there is high between-subject variability because such methods are unable to separate these variation sources. Hence the end results are normally a mixture of all the variation sources and dominated by between-subject variability, except for the unusual situation where the factor of interest had an even more dominating effect.

In univariate statistics analysis, one effective way to overcome such variability is to employ a crossover experimental design in which subjects receive a sequence of different treatments or exposures (e.g. treated by medicine or particular therapy) and then examine the influence of the treatments/exposures applied. For the data obtained from such design, paired statistical tests such as paired *t* test or correlation analysis can be applied and the statistical power is therefore greatly increased. Such methodology has been extended to multivariate analysis under the name of multilevel modelling (Timmerman 2006; van Velzen et al. 2008). Several multilevel models have been proposed and successfully applied to applications which had a similar difficulty such as chemical process monitoring, plant metabolic profiling, and proteomics, most notably multilevel simultaneous component analysis (MSCA) (Jansen et al. 2005a; Ferreira et al. 2009), multilevel partial least squares for regression (ML–PLS–R) (de Noord and Theobald 2005), multilevel partial least squares for discriminant analysis (ML–PLS–DA) (Westerhuis et al. 2010), ANOVA-simultaneous component analysis (ASCA) (Smilde et al. 2005; Jansen et al. 2005b) and ANOVA-principal component analysis (ANOVA-PCA) (Harrington et al. 2005). These models share the same characteristic in their algorithm: to decompose the original data matrix into the sum of a series of sub-matrices, according to the information provided by the experimental design (Smilde et al. 2012). For a crossover design there are generally two sub-matrices: one sub-matrix is the *within-subject* variation matrix while the other is the *between-subject* variation. By doing so, the variation sources were separated and one can apply PCA on the selected sub-matrices of interest (e.g. the *within-subject* matrix) and gain a clearer view of whether the factor of interest does have any significant influence on the data.

However not all experiments are amenable to such approaches. For example, if one wanted to compare the metabolic profiles associated with a chronic disease such as diabetes or rheumatoid arthritis, one would have to have access to patient samples that predated disease onset, which are unlikely to exist in a form suitable for comparative metabolic study. For this type of study, a single factor, one-way ANOVA experiment design has to be adopted and the reported multilevel models may not be an appropriate choice for this type of data. In this study, we propose a new model which was designed to fit the data obtained from such experiment design and can be considered as a supplementary model to those multilevel models as described above to fit the data without an explicit multilevel structure. Finally, we propose a validation method based on a series of permutation tests to ensure that the results given by the multilevel models are genuine and robust.

Two data sets were used in this study and each represented one type of experiment design. The first study analysed VOCs collected from the skin of chronic wounds using a novel skin patch device and analysed by TD–GC–MS (Thomas et al. 2010). A total number of five test subjects were recruited for the study and for each test subject, the skin patches were applied to three different sites: on top of the lesion, the boundary of the lesion and an area of healthy (control) skin. This study, although strictly speaking not a crossover design the resulting data is equivalent to the data obtained from such design and for each subject the samples of all three different classes (i.e. sites) were available. The second data set is pooled from two studies with identical experimental design, involving TD–GC–MS analysis of VOCs in the breath of healthy subjects, as well as patients diagnosed with chronic obstructive pulmonary disease (COPD) or asthma (Ibrahim et al. 2011; Basanta et al. 2012). The aims of these studies were to see whether it was possible to discriminate these three groups of subjects based on VOC profiles in their breath. Unlike the skin VoCs experiment, these studies could not benefit from a crossover design (which is of course impossible to do), instead each subject could only belong to one particular class (healthy, asthma or COPD). In both exemplar data sets, the problem of high between-subject variability appeared to be the main obstacle to achieve clear separation between the expected groups and this obstacle had been overcome by using appropriate chemometrics models.

2 Methods

2.1 Experimental

The full methodologies relating to the collection and analysis of the clinical data sets have been published elsewhere (Thomas et al. 2010; Ibrahim et al. 2011; Basanta et al. 2012). A brief summary of each is presented here.

2.1.1 Skin VOCs data

Five patients were recruited and for each patient VOC samples were collected from three different sites in triplicate: lesion, boundary of the lesion and healthy skin (Thomas et al. 2010). Hence the full data set contained 45 samples. Data pre-processing was the same as described in (Thomas et al. 2010): GC–MS chromatograms were summed into total ion currents (TICs), interpolated and aligned using correlation optimised warping (COW) and subsequently normalised so that the sum of squares (SSQs) of each chromatogram equalled one.

2.1.2 Breath VOCs data

A total number of 115 subjects were recruited for these studies, including 35 healthy subjects, 47 patients with COPD and 33 with asthma. Each subject had four serial breath samples collected and analysed by TD–GC–MS as described (Ibrahim et al. 2011; Basanta et al. 2012). We consider the four serial breath samples as four repeats and do not assume there were significant changes between these repeats. However, some samples had to be removed for various analytical reasons and the final data set contained 400 samples in total. As before the data analysis was performed on the TICs, with the chromatograms baseline corrected, aligned by COW, square root scaled and finally normalised so that the SSQs of each chromatogram equalled one.

2.2 Chemometrics models

For illustration purpose, in this section we assume that there are N samples and J variables in the data under study and thus the full data matrix has a size of $N \times J$. In addition, there are s unique testing subjects and the factor of interest is assumed to have c different levels (i.e. c different classes) in the data set.

2.2.1 Multilevel models

The basic idea of multilevel models is to decompose the original data matrix into the sum of a series of sub-matrices and each sub-matrix carries the variations caused by one particular factor. Then one can model each sub-matrix separately to gain a clearer view of each variation source. We give a brief introduction to a few commonly used multilevel models in this section, namely MSCA, ML-PLS, ANOVA-PCA and ASCA.

MSCA a generalisation of PCA for the data with a multilevel structure, was probably the simplest form of a multilevel model. MSCA separates the *between-level* variations (in this paper it refers to the *between-subject* variations) from the remaining and focus the analysis on what has left after the removal of the *between-level* which is named *within-level* variation. The model of MSCA is given in Eq. (1)

$$\begin{aligned} X &= 1 \cdot m^T + X_{between} + X_{within} \\ &= 1 \cdot m^T + \begin{bmatrix} 1_1 \cdot m_1^T \\ 1_2 \cdot m_2^T \\ \vdots \\ 1_s \cdot m_s^T \end{bmatrix} + X_{within} \\ &= 1 \cdot m^T + T_{between} \cdot P_{between}^T + T_{within} \cdot P_{within}^T + \varepsilon \end{aligned} \tag{1}$$

where $\boldsymbol{1}$ is a vector of the size of $N \times 1$ of 1 s and $\boldsymbol{m^T}$ is a row vector of the size of $1 \times J$ which is the mean vector of $\boldsymbol{X}$; $\boldsymbol{X}_{between}$ is the *between-level* variation matrix and in $\boldsymbol{X}_{between}$ $\boldsymbol{1}_i$ $(i = 1, 2,\ldots s)$ is a column vector of ones and its length equals to the number of samples collected from subject i; $\boldsymbol{m}_i^T$ is the mean vector of all samples collected from subject *i;* $X_{within} = X - 1 \cdot m^T - X_{between}$ which is the *within-level* variation matrix. PCA was then applied to $\boldsymbol{X}_{between}$ and $\boldsymbol{X}_{within}$ separately to obtain the scores and loadings each variation source: $\boldsymbol{T}_{between}$, $\boldsymbol{P}_{between}$, $\boldsymbol{T}_{within}$ and $\boldsymbol{P}_{within}$. Since the *between-subject* variation (i.e. the *between-level* variation) is the one which is not of interest and needed to be removed, the information of interest should be found within the $\boldsymbol{T}_{within}$ and $\boldsymbol{P}_{within}$ pair.

Multilevel partial least squares (ML-PLS) adopted the same methodology MSCA had used which decomposes the data $\boldsymbol{X}$ into the sum of two sub-matrices: the *between-level* matrix $\boldsymbol{X}_{between}$ and the *within-level* matrix $\boldsymbol{X}_{within}$. However, unlike MSCA which uses a PCA model to fit the variance of $\boldsymbol{X}_{within}$ alone, ML-PLS employs a supervised PLS model (Wold et al. 2001) to fit the covariance between the $\boldsymbol{X}_{within}$ and a corresponding response matrix $\boldsymbol{Y}$. The response matrix $\boldsymbol{Y}$ contains the prior knowledge of the sample which could be either continuous such as the concentration levels of one or a few known analytes (the ML–PLS–R model) or categorical such as the class membership of the sample (the ML–PLS–DA model). With the aid of prior information provided by $\boldsymbol{Y}$, the effect caused by the factor of interest would be better reflected in the latent variables in the ML-PLS model than those in the MSCA model which fit the variance of $\boldsymbol{X}_{within}$ only and the interpretability of the model could be further improved.

ANOVA-PCA also takes a similar approach to MSCA but goes further by explicitly modelling all the potentially contributing factors, instead of only isolating the *between-level* variations and placing all else in the *within-level* variation matrix. ANOVA-PCA breaks down the variation sources according to the experimental design by decomposing the original data matrix into the sum of a series of mean matrices. For simplicity reasons, an ANOVA-PCA model for a two factors experimental design, with factors denoted as $f1$ and $f2$ respectively, is given in Eq. (2):

$$
\begin{aligned}
X &= 1\cdot m^T + X_{f1} + X_{f2} + X_{f1\times f2} + \varepsilon \\
&= 1\cdot m^T + \begin{bmatrix} 1_1\cdot a_1^T \\ 1_2\cdot a_2^T \\ \vdots \\ 1_c\cdot a_c^T \end{bmatrix} + \begin{bmatrix} 1_1\cdot m_1^T \\ 1_2\cdot m_2^T \\ \vdots \\ 1_s\cdot m_s^T \end{bmatrix} + \begin{bmatrix} 1_{11}\cdot v_{11}^T \\ 1_{21}\cdot v_{21}^T \\ \vdots \\ 1_{s1}\cdot v_{s1}^T \\ 1_{12}\cdot v_{12}^T \\ \vdots \\ 1_{s2}\cdot v_{s2}^T \\ \vdots \\ 1_{sc}\cdot v_{sc}^T \end{bmatrix} + \varepsilon
\end{aligned}
\tag{2}
$$

where $\boldsymbol{1}\cdot\boldsymbol{m}^T$ is the same as the one in MSCA; $\boldsymbol{X}_{f1}$ and $\boldsymbol{X}_{f2}$ are the factor mean matrices for each factor taken into account by the model respectively. In addition, an interaction matrix $\boldsymbol{X}_{f1\times f2}$ is included. When using ANOVA-PCA as a method to help solving the high *between-subject* variability problem, one could consider *between-subject* variation as one source of variation and assign it to $f1$ while the influence of the factor of interest (e.g. different sampling sites in the skin VoCs study) as another source of variation and assign it to $f2$. Under such assignment, $\boldsymbol{a}_i^T$ $(i = 1,2,\ldots,s)$ is the mean vector of all the samples of subject i; $\boldsymbol{m}_j^T$ $(j = 1,2,\ldots,c)$ is the mean vector of all the samples of class j; $\boldsymbol{v}_{ij}^T$ is the mean vector of all the samples from subject i which belongs to class j. ANOVA-PCA then adds the residue matrix ε where:

$$\varepsilon = X - 1\cdot m^T - X_{f1} - X_{f2} - X_{f1\times f2} \tag{3}$$

back to each of the mean matrices, i.e. $\boldsymbol{X}_{f1} + \boldsymbol{\varepsilon}$, $\boldsymbol{X}_{f2} + \boldsymbol{\varepsilon}$ and $\boldsymbol{X}_{f1\times f2} + \boldsymbol{\varepsilon}$ and performs PCA on each of them separately to obtain the scores and loadings matrices for each variation source. The ANOVA-PCA model can be extended to incorporate more factors and interactions by including more sub-matrices although it could inflate the model rather quickly which is usually unnecessary. Being able to separate the variation sources and focus on a few selected ones is one of the main advantages of the ANOVA-PCA algorithm. Under the context of this study, the sub-matrix of interest would be $\boldsymbol{X}_{f2} + \boldsymbol{\varepsilon}$.

ANOVA-ASCA can be considered as a further development of the ANOVA-PCA model. While ASCA had used the same model as the ANOVA-PCA, the key difference is that the decomposition was performed on the mean matrices directly without adding the residual matrix $\boldsymbol{\varepsilon}$ as shown in Eq. (4) (assuming $f2$ is the factor of interest). After the loadings for each factor were derived from each of the matrix, was then added back to the mean matrix and then projected into the corresponding subspace found by the ASCA using the loadings to obtain the scores as shown in Eq. (5).

$$X_{f2} = T_{f2}\cdot P_{f2}^T \tag{4}$$

$$T_{f2\_ASCA} = (X_{f2} + \varepsilon)\cdot P_{f2} \tag{5}$$

By estimating the loadings from the experiment effect as described in the form of a mean matrix directly, ASCA has better sensitivity in detecting the underlying experiment effect than ANOVA-PCA (Zwanenburg et al. 2011). However such improved sensitivity comes with the drawback that it is no longer easy to interpret the explained variance of each PC as the loadings derived from the mean matrix were unlikely also the ones to maximise the explained variance of the testing matrix (e.g. $\boldsymbol{X}_{f2} + \boldsymbol{\varepsilon}$). Therefore the reconstructed matrix using the loadings and the projected scores could be very different to the testing matrix and resulted in the percentage of the explained variance to be extremely low or even negative (i.e. the fitting error was greater than the total variance of the testing matrix). As we will discuss in later sections that the explained variance of the discriminant PCs could be used to add another layer of confidence of the findings, we have adopted ANOVA-PCA rather than ASCA in this study (although ASCA could be a better choice for the problems when ANOVA-PCA had difficulty finding the deviation caused by the factor of interest).

2.2.2 *ANOVA mean-centring for single factor, one-way ANOVA data*

It can be seen that all the multilevel models described above are essentially performing a localised mean centring to produce a series of sub-matrices, with each sub matrix representing one particular variation source. PCA is then performed on each sub-matrix to generate a model for each variation source. Based on the same methodology, we propose a simplified model to fit the data obtained from a single factor, one-way ANOVA experiment design with the aim to overcome the problem of having high between-subject variability and allow discovering the underlying pattern which relates to the experimental questions. The testing matrix $\boldsymbol{X}_{test}$ which is subject to PCA is given in Eq. (6).

$$
\begin{aligned}
X_{test} &= X_f + \varepsilon_{anova\_mc} \\
&= \begin{bmatrix} 1_1\cdot m_1^T \\ 1_2\cdot m_2^T \\ \vdots \\ 1_c\cdot m_c^T \end{bmatrix} + \varepsilon_{anova\_mc}
\end{aligned}
\tag{6}
$$

In which $\boldsymbol{1}_i$ $(i = 1,2,\ldots,c)$ is a column vector of 1s and the length of each vector equals the number of samples of each class; $\boldsymbol{m}_i^T$ (i = 1,2,…,c) is a row vector which is the mean vector of all the samples of class i. The mean vectors are all calculated from mean centred $\boldsymbol{X}$. The residual matrix $\boldsymbol{\varepsilon}_{anova\_mc}$ is obtained by firstly mean centre original data matrix, then calculate the mean of each subject and subtract them from the corresponding rows (samples) as shown in Eq. (7).

$$\varepsilon_{anova\_mc} = X - 1 \cdot m^T - \begin{bmatrix} 1_1 \cdot a_1^T \\ 1_2 \cdot a_2^T \\ \vdots \\ 1_s \cdot a_s^T \end{bmatrix} \tag{7}$$

In which $\boldsymbol{1}_j$ (j = 1,2,…,s) is a column vector of ones and the length of the vector equals the number of samples of the test subject j; $\boldsymbol{a}_j^T$ is a row vector which is the mean vector of all the samples collected from the test subject j.

If there were no significant dynamic effect between the repeated measurements of the same subject, $\boldsymbol{\varepsilon}_{anova\_mc}$ is essentially the variation caused by experiment itself, e.g. sampling error, instrument measurement error etc. Analysing it together with the $\boldsymbol{X}_f$ is equivalent to superimpose the between-groups difference onto the unavoidable variance introduced by the experiment and assess the significance level of the between groups variance. $\boldsymbol{X}_f$ could either added the residual matrix $\boldsymbol{\varepsilon}_{anova\_mc}$ back and then perform PCA on it (ANOVA-PCA approach), or subject to decomposition directly to obtain the loadings first and then add back $\boldsymbol{\varepsilon}_{anova\_mc}$ and be projected into the subspace via the loadings (ASCA approach). In this study, we employed the ANOVA-MC by using the ANOVA-PCA decomposition approach.

This is a strategy highly similar to the one used by ANOVA tests in statistics and its multivariate extensions such as ANOVA-PCA and ASCA, thus we name this data pre-processing method as ANOVA-mean centring (ANOVA-MC). It is worth noting that it is also possible to apply ANOVA-MC to a data set obtained from a two-way ANOVA design as well by assuming that there is no significant interaction between the factor of interest and each testing subject, i.e. the factor of interest should have a similar effect on every testing subject. If such effect varies significantly from one testing subject to another (i.e. strong interaction exist) then ANOVA-MC could fail to detect it. Under such circumstances one of the multilevel models would be a more appropriate choice, assuming a crossover two-way ANOVA experiment design is applicable.

2.3 Model validation

In ANOVA-PCA, ASCA, and also ANOVA-MC, the labelling information about which samples belong to which group has been used when performing the localised mean centring. Therefore, the mean matrix of interest, itself (e.g. $\boldsymbol{X}_{f2}$ in ANOVA-PCA or $\boldsymbol{X}_f$ in ANOVA-MC) has become a latent factor which would cluster samples into expected groups according to their class labels, thus there is no question about whether the samples from different groups could be separated from each other in the PCA model applied to such testing matrix. The question arises as to whether such separation is statistically significant when comparing it to the background variations, i.e. the residue matrix and, more importantly, the chance of such separation shown in the PCA model had been a false discovery. MSCA only incorporates the subject information into the modelling and thus should have relatively low risk of false discovery.

A validation procedure based on null-hypothesis and permutation test was proposed for ASCA (Vis et al. 2007; Zwanenburg et al. 2011) in which the null-hypothesis assumes the deviation caused by the factor of interest is zero and the alternative hypothesis is that such deviation is significantlly greater than zero. The observed deviation is described as the SSQs of the scores of the PCA performed on the corresponding mean matrix (i.e. the scores of PCA performed on $\boldsymbol{X}_{f2}$). The null distribution was then obtained by permute the labels of the samples a large number of times and calculate the SSQs of each permutation. An empirical p-value assessing the significance level of the observed SSQs can then be found by counting the number of permutations obtained greater SSQs than the observed ones.

In this paper we propose an alternative validation procedure which is also based on random re-sampling and permutation, albeit to be able to provide more useful information in addition to the significance level assessment of the observed deviation caused by the factor of interest. We firstly assume that there were c known classes and the results of ANOVA-PCA, or ANOVA-MC had showed a clear separation between these classes which matched the class labels well and the k PCs are required to separate all the known classes (ideally k should be no greater than c–1). The aim of the validation is to assess the reproducibility of the separation between the groups. This is achieved by randomly re-sample the data sets R times to generate R different subsets of the data. Each subset of the data was then analysed by the method to be validated. The K-means clustering analysis (Hartigan and Wong 1979) was then performed on the final PCA results using a sufficient number of PCs which were able to separate the classes. The number of clusters is set to c and the initial cluster centroid positions were set to be the mean of each class, calculated from the subset of samples using the known group labels. This way the clusters identified by the K-means clustering should have a 1-to-1 correspondence with the expected classes. A pattern with the known classes well separated from each other would be expected to see a high consistency between the known class labels and the labels identified by K-means clustering. Such high consistency should also be reproducible for the models which were built on different subsets of samples. By contrast, if the observed separation was caused by chance or there was no genuine separation the results of the *K*-means clustering would be rather unpredictable. If there was no true underlying difference between the expected classes, the PCA scores obtained would be expected to be a homogeneous mixture and the clusters identified by the *K*-means clustering are merely arbitrary collections of samples, depending on the relative distance between them, and there should be little to no agreement between the expected group labels and those assigned by the *K*-means clustering. Thus for each subset of the data obtained by the random re-sampling, the same analysis as described above was repeated a second time by using the same data but with the class labels randomly permuted, i.e. each sample was randomly assigned a class membership. The consistency between the known class labels and labels identified by *K*-means clustering were calculated both for the model using the original labels, and the one using the permuted labels. If the separation between the known classes were genuine, the label consistency of the models using the original labels (the observed consistency) should be always higher than those using the permuted labels (the null consistency). An empirical *p*-value can be derived by counting the number of cases when the null consistency value had been higher than the observed consistency value and divide it by *R*. In addition, a confusion matrix can be calculated by comparing these two types of labels. In the confusion matrix, each row contains the percentage of the samples in one particular cluster coming from each of the known class while each column contains the percentage of the samples in one particular class allocated into each of the clusters identified by the *K*-means clustering. Compared to the validation procedure reported previously (Vis et al. 2007; Zwanenburg et al. 2011), such a confusion matrix gives a more detailed information of the distribution of the classes, e.g. which class(es) were better separated from others and which classes may have certain amount of overlap between them, similar to the confusion matrix provided by supervised classification models. This is particular beneficial for those studies which contain more than two classes. The procedure of the validation is illustrated as a flowchart in the Supplementary Information Figure S1. It is noteworthy that the percentage of the total explained variance (TEV%) of the first *k* PCs also has an important statistical meaning. It provides a numerical indicator representing how significant the separation between the groups showed by the first *k* PCs is compared to the background noise. However, considering the factor of interest could be a genuine but weak variation source and/or some experiments may have high errors in their measurement, the TEV% alone may not be a suitable measurement for validation purpose, nevertheless having a significant higher TEV% than those of the null distribution could be considered as an extra piece of evidence showing the significance level of the discovery. We also would like to point out that the success of the validation largely rely on the clustering algorithm. Although *K*-means clustering had been used in this particular study for its rapid speed and easy to implement, it is a rather basic clustering algorithm and has some well-known drawbacks such as the end results is only a local minimum and may produce counterintuitive results; it also uses Euclidean distance as the metric which may not always be appropriate to describe a complex cluster shape. For the data with complex structure more sophisticated clustering algorithm such as self-organising mapping (Kohonen 1982), superparamagnetic clustering (Blatt et al. 1996) might be needed for better fit to the data.

In this study, we consider the samples from the same subject as a whole during the re-sampling procedure, i.e. all the samples from the same subject will be either selected or left out together during the random re-sampling; also in the permutation test all the samples of the same subject had the same class membership (albeit randomly assigned). For the skin VOCs study, since the number of subjects available is rather limited (five testing subjects in total), we randomly selected three testing subjects for each re-sampling and all ten possible unique combinations were tested ($R = 10$). The validation of ML–PLS–DA was done by using the same re-sampling of the samples as for ANOVA-PCA and ANOVA-MC except that for the ML–PLS–DA validation, the three randomly chosen subjects were used as the training set and the remaining two subjects were used as the test set. The number of PLS-components of the model were determined by using a threefold cross-validation, performed on the training set only with one subject been left out at one time. The result of ML–PLS–DA was reported as the averaged correct classification rate (CCR) of the test set samples of all the 10 different combinations of the training and test sets. For the breath VOCs study, there were many more test subjects available, thus bootstrap (Efron and Tibshirani 1993) was used for this data set. A total number of 1,000 bootstrap re-sampling procedures with replacement were performed ($R = 1{,}000$) and the distributions of the label consistency between the models using the original labels and those using the permuted labels were compared.

3 Results and discussion

3.1 Skin VOC data set

The PCA scores plots of the skin VOC data are given in Fig. 1a where samples of different classes (i.e. sites) were labelled by different colours and symbols while in Fig. 1b each class has the same colour as in Fig. 1a each sample was labelled by its subject id. In the scores plot, all three classes heavily overlap. The samples collected from healthy skin were somewhat separated from others while there was no separation between the samples from boundary and those from the lesions themselves. By comparing Fig. 1a and b it can be seen that different subjects appeared to have rather different locations (i.e. means) and such variability appears to be the dominant source of variation, masking potential separation caused by other variation sources (e.g. different sampling sites).

The scores plot of MSCA, ANOVA-PCA, and ML–PLS–DA performed on the same data set are provided in Fig. 2a–c. Although ANOVA-MC model was designed to fit a one-way ANOVA data structure, we also applied it to this data set as well to test its applicability as it is reasonable to expect that the VOCs profiles coming from the three different sites would form a similar pattern for all the test subjects. The result of ANOVA-MC is given in Fig. 2d. All four models show a considerably clearer separation between the samples collected from the three different sites and the between-subject variability has been effectively removed. A common trend showed by these models is a gradient from the healthy skin samples to the lesion samples with the healthy skin samples appeared to be better separated from the other two types of samples. Such gradient indicates a gradual change in the VOC profiles from an area of healthy skin to the lesion itself. Between these models, the ANOVA-MC method appeared to have provided the best separation across all three classes, which is rather surprising and indicates that when the influence of the factor of interest is consistent between different subjects (i.e. no strong interactions between the factor of interest and the test subjects), ANOVA-MC model can also be an appropriate choice. By contrast, the result of MSCA shown that the three classes appeared to still have some overlapping, even between the healthy samples and the lesion samples. It is worth noting that the PLS scores plot of the ML–PLS–DA model is provided because it is directly comparable to the results of the other three models. The results of a supervised classification model such as ML–PLS–DA are better presented in a form of properly validated prediction performance such as the prediction accuracies of the blind test samples. Such results are shown in Table 1 and these will be discussed in the validation section (see below).

3.2 Breath VOC data set

The results of PCA on this data set are presented in Fig. 3a, b with class and subject id labelled respectively. The influence of between subject variability is even more dominant than that in the skin VOCs data. In Fig. 3b it is very easy to see that the samples coming from same subject were mostly clustered close to each other and there were no obvious separations between the three expected classes at all. Possible reasons for such strong between-subject variability include the long data collection period (1 year), and also the large number of subjects recruited, when compared with the skin VOCs study. Using the raw data, even supervised classification methods, with appropriate bootstrap validation, such as partial least squares for discriminant analysis (PLS-DA) or support vector machines could not provide a reliable classification model separating the three types of samples (data not shown). This suggests that the reason for the lack of separation in the PCA plot was not that the most discriminant PCs in the PC space had not been identified, but rather because the high between-subject variability had suppressed (overtly affected) the separation caused by the factors of interest (if there were any).

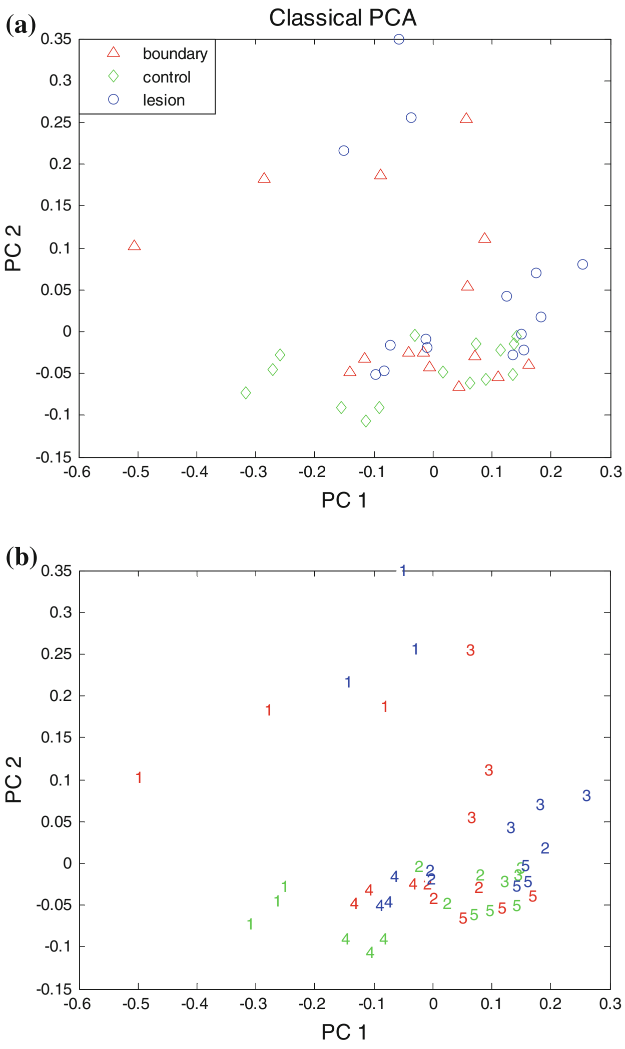

Fig. 1 PCA scores plots of the skin VOCs data set: **a** scores plot with samples labelled according to the sampling site; **b** scores plot with samples labelled according to the subject i.d

Since this experiment does not have a crossover design, the multilevel models such as MSCA or ML–PLS–DA are not applicable to this data set and indeed no sensible pattern could be seen from either $\boldsymbol{X}_{between}$ or $\boldsymbol{X}_{within}$ (data not

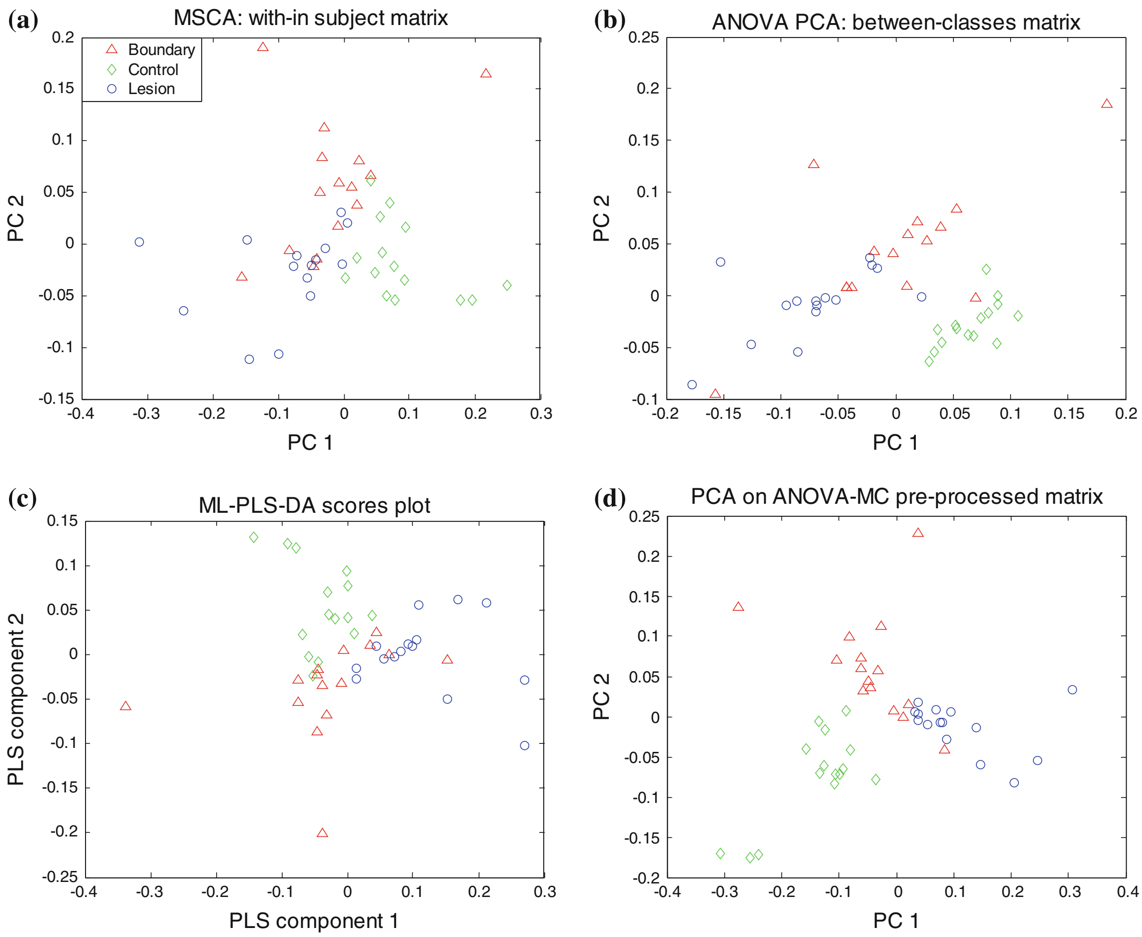

Fig. 2 The results of multilevel models performed on the skin VOCs data set. **a** the scores plot of MSCA; **b** the scores plot of ANOVA-PCA performed on the between-class matrix; **c** the scores plot of ML–PLS–DA model; **d** the scores plot of PCA performed on the ANOVA-MC pre-processed matrix

Table 1 Averaged clustering/classification results of the skin VOCs data

Predicted label	Known label		
	Healthy skin (%)	Boundary (%)	Lesion (%)
ANOVA-MC			
Cluster 1	81.11	3.33	0
Cluster 2	17.78	75.56	6.67
Cluster 3	1.11	21.11	93.33
ANOVA-PCA			
Cluster 1	81.44	10.00	5.56
Cluster 2	16.33	66.67	12.22
Cluster 3	2.23	23.33	82.22
ML–PLS–DA			
Healthy skin[a]	83.33	10.00	0
Boundary[a]	15.00	73.33	13.33
Lesion[a]	1.67	16.67	86.67

[a] The percentage of the predicted labels been assigned to that class

shown). The result of the ANOVA-MC model which is given in Fig. 4 and resulted in a very clear separation between the three classes with the first PC being the most discriminant. The healthy class is in the middle while Asthma class is on one side and COPD class is on the other side. In addition, the separation between the healthy and asthma samples was seemly better than that between COPD and healthy samples. This result is consistent with the nature of the diseases under study. Asthma is clearly a disease characterised by disrupted biophysiological processes and independent of the ageing process. COPD on the other hand is a disease caused by the cumulative deleterious effect of inhaled particles (usually cigarette smoke) on the lung in susceptible individuals; as such its prevalence rises with age and in fact has been said to be a disease of “accelerated lung ageing” (MacNee 2009). It is therefore unsurprising that the data modelled from COPD patients in our study lie at one extreme of “health”, whereas the data from asthmatic subjects were clearly separate. This finding

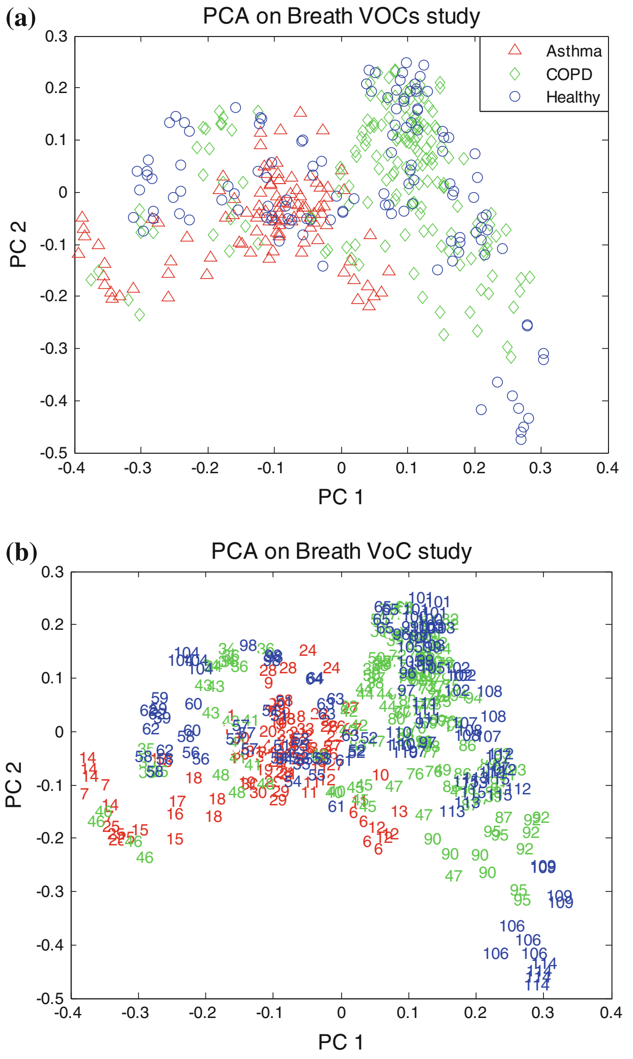

Fig. 3 PCA scores plot of the breath VOCs data set. **a** Scores plot with samples labelled according to the class; **b** scores plot with samples labelled according to the subject i.d

has also been noted in our original modelling of the "COPD *versus* healthy" data, using far simpler statistical methods (Basanta et al. 2012), and also by others in an independent study population and using different methods of sample and data analysis (Fens et al. 2009).

3.3 Validation

For the skin VOCs data set, the averaged consistency of ANOVA-MC using the original labels was 83.33 % (the observed consistency) while that of using the permuted labels (the null consistency) was 43.7 %; ANOVA-PCA had obtained an averaged observed consistency of 76.78 % and the corresponding averaged null consistency was 42.9 %. For both methods, there was no instance that the null consistency had been higher than the observed consistency ($p < 0.1$). The averaged confusion matrix is given in Table 1, where a gradient from the healthy skin samples to the lesion samples is observed. There were very few lesion samples found in the healthy cluster (cluster 1) and vice versa, indicating these two groups of samples were well separated; the boundary samples had been found in all three clusters although it appeared that the "lesion cluster" (cluster 3) had more boundary samples than the healthy skin cluster (cluster 2) had, indicating that the boundary samples were generally more similar to the lesion samples. Comparing the clustering results from ANOVA-MC and those from ANOVA-PCA, it appeared that ANOVA-MC had generated slightly better separations between the three groups, particularly between the healthy and the lesion clusters and also between the boundary and the lesion clusters. The validated CCR of the ML–PLS–DA model is also given in Table 1. The averaged CCR is 81.11 % which is a significant improvement over the results obtained from the classical PLS-DA model as we reported before (73.33 % for healthy skin *vs.* remaining; 56.67 % for boundary *vs.* lesion samples).

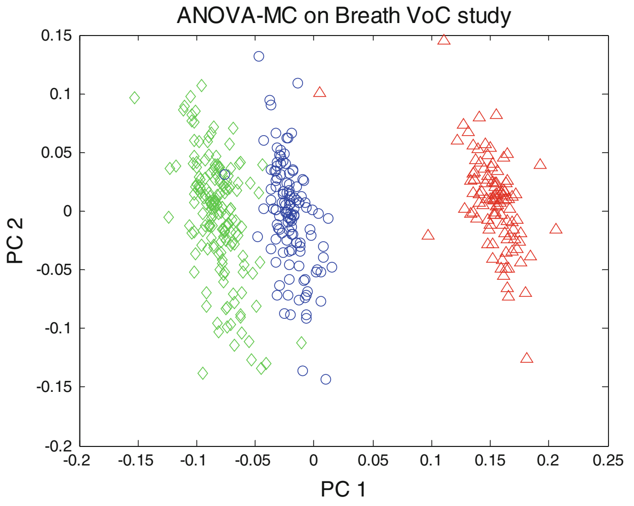

Fig. 4 The scores plot of PCA performed on the ANOVA-MC pre-processed matrix

For the breath VOCs study, ANOVA-MC is the most suitable model for the data and the distributions of the label consistency between the models using the original labels and those using the permuted labels were given in supplementary information, Figure S2. A distinct difference can be seen between the two distributions. The null distribution resembles a normal distribution which is centred at 39.6 % while that of the observed distribution had a mean of 88.1 % and the majority (610 out of 1,000) having a consistency greater than 90 %. For all the 1,000 bootstrapping, there was not a single case that the consistency of the model using the original labels had been lower than that using the permuted labels ($p < 0.001$). In addition, the

Table 2 Averaged clustering results from the breath VoCs data

Clustering label	Known label		
	Asthma (%)	COPD (%)	Healthy (%)
ANOVA-MC			
Cluster 1	98.71	0	0
Cluster 2	0.09	87.67	19.08
Cluster 3	1.21	12.33	80.92

averaged confusion matrix (Table 2) had confirmed the pattern revealed in the initial analysis, the asthma samples were well separated from healthy and COPD samples; there were some overlapping between the healthy and COPD classes although they could still be distinguished as two separated groups. The distributions of the TEV% of the first PC (the most discriminatory PC for this particular data set) using the original labels and the one using permuted labels of the 1,000 bootstrapping procedures on the breath VOCs data is given in supplementary information Figure S3. The TEV% of the first PC of the models using the original labels had, on average, explained three times or more variance than those using the permuted labels, indicating the separation caused by disease status (healthy, asthma or COPD) was in indeed very significant. Similar results were also found in the skin VOCs data (data not shown).

4 Conclusion

Through the analysis of these two clinical metabolic profiling data sets, we have demonstrated that the problem of high between-subject variability can be overcome using a suitable chemometrics model which takes this variability that may not be correlated with disease into consideration. Furthermore, the models generated were successfully validated using a random re-sampling method coupled with a permutation test. As a method which does not need the actual grouping information, MSCA does not require this validation step to authenticate its findings and therefore is particularly suitable for the studies with limited number of samples. Although in this study the pattern revealed by MSCA seemly to be less clear than those found by more sophisticated models such as ANOVA-PCA, ML–PLS–DA and ANOVA-MC. In addition, it is essential for the success of MSCA that a crossover experimental design or equivalent was used, which means all types of samples should be available from each of the testing subject, and this is unlikely in case–control studies. In experiments such as the breath VOCs study, this requirement cannot be met, and ANOVA-MC was the most appropriate choice. The limitation of the ANOVA-MC method presented in this paper is that whilst it is effective tool to confirm or deny hypotheses, it currently has no capability to predict the unknowns. To overcome the high between-subject variability without helps from a crossover experiment design in predictive modelling remains a challenging task and should be an interesting area for future research. Finally a validation procedure has been proposed and the most attractive aspect of this procedure is that it can provide a confusion matrix and it can provide an estimation about the relative positions of each expected classes.

Acknowledgments We thank Dr. Maria Basanta and Dr. Baharudin Ibrahim for providing the breath VOCs data; Dr. Alexi Thomas, Dr. Svetlana Riazanskaia and Dr. William Cheung for providing the skin VOCs data.

References

Assfalg, Michael, Bertini, I., Colangiuli, D., Luchinat, C., Schäfer, H., Schütz, B., et al. (2008). Evidence of different metabolic phenotypes in humans. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 1420–1424.

Basanta, M., Ibrahim, B., Dockry, R., Tal-Singer, R., Douce, D., Woodcock, A., et al. (2012). Exhaled volatile organic compounds as potential biomarkers in chronic obstructive pulmonary disease. *Respiration Research, 13*, 72.

Biais, B., Allwood, J. W., Deborde, C., Xu, Y., Maucourt, M., Beauvoit, B., et al. (2009). 1H NMR, GC–EI-TOFMS, and data set correlation for fruit metabolomics: application to spatial metabolite analysis in melon. *Analytical Chemistry, 81*, 2884–2894.

Blatt, M., Wiseman, S., & Domany, E. (1996). Superparamagnetic clustering of data. *Physical Review Letters, 76*, 3251–3254.

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

Cheung, W., Xu, Y., Thomas, C. L. P., & Goodacre, R. (2008). Discrimination of bacteria using pyrolysis-gas chromatography-differential mobility spectrometry (Py-GC-DMS) and chemometrics. *Analyst, 134*, 557–563.

de Noord, O. E., & Theobald, E. H. (2005). Multilevel component analysis and multilevel PLS of chemical process data. *Journal of Chemometrics, 19*, 301–307.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fens, N., Zwinderman, A. H., van der Schee, M. P. C., de Nijs, S. B., Dijkers, E., Roldaan, A. C., et al. (2009). Exhaled breath profiling enables discrimination of chronic obstructive pulmonary disease and asthma. *American Journal of Respiratory and Critical Care Medicine, 180*, 1076–1082.

Ferreira, D. L. S., Kittiwachana, S., Fido, L. A., Thompson, D. R., Escott, R. E. A., & Brereton, R. G. (2009). Multilevel simultaneous component analysis for fault detection in multicampaign process monitoring: Application to on-line high performance liquid chromatography of a continuous process. *Analyst*, *137*, 1571–1585.

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.

Hartigan, J. A., & Wong, M. A. (1979). A *K*-means Clustering Algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics), 28*, 100–108.

Ibrahim, B., Basanta, M., Cadden, P., Singh, D., Douce, D., Woodcock, A., et al. (2011). Non-invasive phenotyping using exhaled volatile organic compounds in asthma. *Thorax, 66*, 804–809.

Jansen, J. J., Hoefsloot, H. C. J., Greef, J., Timmerman, M. E., & Smilde, A. K. (2005a). Multilevel component analysis of time-resolved metabolomics data. *Analytica Chimica Acta, 530*, 173–183.

Jansen, J. J., Hoefsloot, H. C. J., Greef, J., Timmerman, M. E., Westerhuis, J. A., & Smilde, A. K. (2005b). ASCA: Analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics, 19*, 469–481.

Kohonen, Teuvo. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59–69.

MacNee, W. (2009). Accelerated lung aging: a novel pathogenic mechanism of chronic obstructive pulmonary disease (COPD). *Biochemical Society Transactions, 37*, 819–823.

Penn, D. J., Oberzaucher, E., Grammer, K., Fischer, G., Soini, H. A., Wiesler, D., et al. (2007). Individual and gender fingerprints in human body odour. *Journal of the Royal Society, Interface, 4*, 331–340.

Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., Lamers, R-Jan, 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., Timmerman, M. E., Hendriks, M. M. W. B., Jansen, J. J., & Hoefsloot, H. C. J. (2012). Generic framework for high-dimensional fixed-effects ANOVA. *Briefings in Bioinformatics, 13*, 524–535.

Thomas, A. N., Riazanskaia, S., Cheung, W., Xu, Y., Goodacre, R., Thomas, C. L. P., et al. (2010). Novel noninvasive identification of biomarkers by analytical profiling of chronic wounds using volatile organic compounds. *Wound Repair and Regeneration, 18*, 391–400.

Timmerman, M. E. (2006). Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology, 59*, 301–320.

van Velzen, E. J. J., Westerhuis, J. A., van Duynhoven, J. P. M., et al. (2008). Multilevel data analysis of a crossover designed human nutritional intervention study. *Journal of Proteome Research, 7*, 4483–4491.

Vis, D. J., Westerhuis, J. A., Smilde, A. K., & van der Greef, J. (2007). Statistical validation of megavariate effects in ASCA. *BMC Bioinformatics, 8*, 322–330.

Westerhuis, J. A., van Velzen, E. J. J., Hoefsloot, C. J., & Smilde, A. K. (2010). Multivariate paired data analysis: Multilevel PLSDA versus OPLSDA. *Metabolomics, 6*, 119–128.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*, 109–130.

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.

Zwanenburg, G., Hoefsloot, H. C. J., Westerhuis, J. A., Jansen, J. J., & Smilde, A. K. (2011). ANOVA-principal component analysis and ANOVA-simultaneous component analysis: A comparison. *Journal of Chemometrics, 25*, 561–567.