Journal of Chromatography B, 871 (2008) 227–235

Contents lists available at ScienceDirect

Journal of Chromatography B

journal homepage: www.elsevier.com/locate/chromb

Instrumental and experimental effects in LC–MS-based metabolomics☆

Lyle Burton*, Gordana Ivosev, Stephen Tate, Gary Impey, Julie Wingate, Ron Bonner

Applied Biosystems|MDS Sciex, 71 Four Valley Drive, Concord, Ontario, Canada L4K 4V8

ARTICLE INFO

Article history:
Received 8 February 2008
Accepted 30 April 2008
Available online 10 May 2008

Keywords:
Metabolomics
LC–MS
Multivariate analysis

ABSTRACT

The experimental complexity of a metabolomics study can cause uncontrolled variance that is not related to the biological effect being studied and may distort or obscure the data analysis. While some sources can be controlled with good experimental techniques and careful sample handling, others are inherent in the analytical technique used and cannot easily be avoided. We discuss the sources and appearance of some of these artifacts and show ways in which they can be detected using visualization and statistical tools, allowing appropriate treatment prior to multivariate analysis (MVA).

1. Introduction

Metabolomics requires, and relies on, the ability to make statistically meaningful comparisons of the data for large numbers of variables from diverse samples that may have been acquired and/or analyzed at different times. For this to be feasible the total variance must be carefully controlled and ideally limited to that due to the effect being studied. This prospect is especially daunting given the number of experimental steps involved, each of which can introduce variance that may not always be understood or characterized.

Regardless of the analytical technique used, it is widely understood that particular care must be taken with sample collection, storage and preparation, and that it is essential (but often forgotten) to randomize the sample order prior to data acquisition to minimize instrument effects [1]. For LC–MS using electrospray ionization, one well-known effect is ion suppression where a high-abundance analyte reduces, or eliminates, the response for a weaker analyte. This was studied by Bottcher et al. [2] who concluded, using the nomenclature of Matuszewski [3], that significant absolute matrix effects can occur but relative effects are usually small enough to allow meaningful comparison. Surprisingly, the latter report also described the possibility of ion *enhancement* in real matrices relative to standards made up in pure solvent which, being unexpected, might be hard to distinguish from an up regulated biomarker. Instrument response changes are also well known and a number of compensating experimental strategies have been described. Normalization to one or more internal standards is common in bioanalysis, and Sysi-Aho et al. [4] have described an approach using multiple internal standards including a method to select the optimum standard to be used with any particular analyte. Response changes are particularly important in large-scale, long-term studies where the samples cannot be run in a single batch and may not even be run on the same instrument. A popular strategy in these cases is to use a QC sample obtained by pooling all the analytical samples [5–7] with internal standards and/or a test mixture. The QC sample is injected periodically and used to define intra-batch response curves [6] that can then be used to adjust the response of the analytical samples. The internal standards can also be used for inter-batch response correction and the test mixture to ensure that no gross chromatographic changes have occurred [7] during the batch.

These effects, however, are not the only phenomena that can be introduced by LC–MS; others can be related to the autosampler, such as carryover and sample stability, the LC system (retention time shifts, response changes, contamination build up or wash out) as well as the mass spectrometer (mass shifts, resolution changes), etc. Somewhat surprisingly, however, there seems to have been relatively little mention in the literature of artifacts introduced by the analytical technique, the potential effects that these may have on the data, or how to identify and treat them. Here we show how some of these experimental artifacts can be identified and characterized. We illustrate our approaches using a simple data set but note that they could equally be applied to the QC replicates in a study using pooled QCs. It is clear that there is no standard data processing workflow for LC–MS-based metabolomics [8] or proteomics [9], and that despite similar terminology (feature extraction, peak finding, alignment, pre-processing, pre-treatment, normalization, scaling, etc.) the meaning, usage order and actual algorithms differ

☆ This paper is part of a special volume entitled "Hyphenated techniques for global metabolite profiling", guest edited by Georgios Theodoridis and Ian D. Wilson.

* Corresponding author.
E-mail address: lyle.burton@sciex.com (L. Burton).

1570-0232/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2008.04.044

significantly [10]. Nevertheless, the need for 'clean data' [11–13], i.e. lists of compounds or uniquely identified unknown peaks with quantitative information such as response, is recognized as being essential for effective MVA pre-processing and processing. Data filtering or cleaning is usually performed on a per sample basis, is typically limited to peak finding and background subtraction to detect peaks with offsets, and, for some workers, may include alignment and/or normalization.

As far as we are aware there are no reports of examining the peak lists for cross sample effects, such as instrument artifacts, *prior* to MVA although Fiehn [14] mentions the importance of checking observed clusters *following* analysis and notes that observed cluster patterns correlated with acquisition time may indicate instrumental effects. Artifacts may be removed by variable selection which is often used as part of data analysis to reduce the dimensionality of the data, simplify processing and interpretation [15], and improve the classification/prediction performance of the statistical model [16].

We argue that the presence of artifact peaks should be determined prior to any processing for a number of reasons. First, some processing, such as normalization to biological factors, may obscure the artifact pattern making it hard to detect later. Secondly, the presence of artifacts may distort or overwhelm any sample separation that might otherwise be revealed. Thirdly, the presence of artifacts may indicate that the experiment or study is invalid and should be repeated although, in many cases, artifact peaks can be ignored or 'rescued' by appropriate treatment. Variable rejection in this way can improve the performance of MVA and ease interpretation; for example, Jonsson et al. [17] noted that the presence of instrumental effects caused additional PLS components and made interpretation harder, while Scholz et al. [18] found that the third independent component, obtained from independent component analysis of PCs, was related to acquisition order and represented gradual instrument contamination. We recommend that artifact evaluation include the use of unsupervised tools such as principal component analysis (PCA) [19,20] so that unexpected effects can be detected.

2. Artifact types and manifestations

In general LC–MS artifacts are of two types:

1. Chemical noise, i.e. real peaks that are generated in the instrument, or apparent changes that are instrument related, such as ion suppression.
2. Apparent sample differences introduced by the experimental protocol or the data processing.

One example of the second type has already been reported [21]; in that study the parameters initially chosen for peak picking resulted in peaks being missed in some samples but not others introducing apparent differences. We have observed experimental artifacts [22] in proteomics samples analyzed by "GeLC–MS", i.e. where samples are separated by 1D gel electrophoresis, each strip cut into bands and the bands analyzed by LC–MS, caused by slight differences in the position of the cut bands and hence the proteins selected and analyzed. Similarly, we have found that incorrect alignment can result in the same peak being incorrectly assigned to two variables again resulting in an apparent difference. This has been reported by de Groot et al. [23] who also described a method for detecting and correcting the problem.

The focus of this report is artifacts of the first type that are summarized in Table 1 and that have one of the following characteristics:

1. Constant across all runs (background components in solvents, etc.).
2. Suppression/enhancement that occurs at specific retention times.
3. Trends that are dependent on the acquisition order.

Since the first type do not change between samples they will be ignored by MVA, but it is still important to detect their presence for two reasons: (1) they indicate that the analysis system, particularly the response, is not changing during the analysis, and (2) normalization may change the pattern so that it appears significant. Constant artifacts may be impossible to distinguish from real components that do not change, but neither will have any effect on the MVA results. Their removal is similar to variable selection [15,16] and may improve later MVA performance or ease interpretation.

Suppression can be detected by determining variables that are anti-correlated with variables which increase; if the retention times are comparable, ion suppression must be suspected and the data carefully examined. Similarly, ion enhancement should be suspected in variables that increase when more intense peaks from a different compound at the same retention time also increase. This possibility underscores the need to carefully validate suspected discriminating variables or biomarkers, ideally using a different chromatographic system as recommended for quantitative analyses [3].

The third class of artifacts all show changes that are dependent on acquisition order and hence can be detected by examining the response profile when the data points are arranged appropriately. Manual examination is time consuming and error prone so automated, unsupervised (from the biological viewpoint) techniques are desirable. Here we describe and illustrate several approaches that we routinely use to detect artifacts so they can be removed from the data set or treated in some other way.

3. Approaches to artifact detection

As noted, it is important that artifacts introduced by the analysis system be detected prior to normalization to biological or sample related parameters, otherwise they will be obscured. The presence of invariant artifacts indicates constant system response so that time related artifacts can be reliably detected from their profiles. In our experience, normalization is generally not necessary for relatively small batches of samples that can be acquired on the same instrument at the same time, but is critical if these conditions are not true.

The following summarizes a number of techniques that we have found useful for detecting and characterizing artifacts; all are illustrated in the results section using a simple LC–MS data set.

3.1. Univariate analysis

Since we are looking for effects that depend on acquisition order, we need a technique that can determine differences between samples acquired at the start of the data set and those at the end. If the acquisition order has been randomized, the only systematic difference between these samples will be due to time related artifacts and, if all real classes are roughly equally represented, class effects should cancel out.

Although not statistically rigorous, we have found it useful to create artificial sample classes for the first few (n) and the last few samples in the set and to use a t-test to detect different or similar variables that we then examine manually. A t-test is normally used to determine if a given variable distinguishes two known groups while here we assume that the groups are the same so that any

Table 1
Summary of some significant artifacts, origins and manifestations in acquired data

Artifact	Source	Manifestation
1. Carry over	Autosampler	Incorrect measurement of a variable in sample(s) following a sample containing a large amount of that variable Might be exacerbated by randomization if low-level samples follow intense ones
2. Sample decomposition	Autosampler	Variables that change intensity during the course of the analysis; can be mistaken for a change in response Unstable compounds (variables) will decrease in time while the decomposition products will increase
3. Background components	LC system	Compounds present in the solvent, column, etc. that cause distinct peaks in the data In the absence of instrument changes these will have a constant value across all samples
4. Contamination build up or wash out	Entire system	Variables that change in intensity during the course of the analysis, often in a non-linear way such as sigmoidal These compounds may collect in the autosampler, column or ion source
5. Sensitivity changes	MS	Variables that decrease in intensity during the run If the cause is instrument failure the ions observed may also change
6. Ion suppression or enhancement	MS	Variables that are decreased or increased in some samples because of the presence of intense co-eluting peaks
7. Saturation	MS	Incorrect response for large concentrations because the dynamic range of the detection system has been exceeded May, in severe cases, appear as an intensity decrease

change must be due to the variable. The choice of n depends on the need to represent the sample classes, ideally while restricting the analysis to a few variables to maximize the difference between our artificial classes. Further enhancement can be achieved if the artificial classes contain samples that are known to be the same or similar, for example, they are pooled controls analyzed periodically during the study, from the same group or replicates.

3.2. Variable grouping

We have recently described [24] an unsupervised technique we call principal component variable grouping (PCVG) that uses PCA and processes the loadings data to find variables that are correlated across the samples, i.e. we are using the samples to characterize and group the variables as well as the reverse. Briefly, following PCA, PCVG starts with the most significant variable (the one farthest from the origin for Pareto scaled [11] data) and finds other variables that are within a user specified angle of it in the n-dimensional space defined by the selected PCs. These variables are then assigned to the same group, and the process repeated using the next most significant unassigned variable until all variables (typically above a user specified threshold) have been assigned.

Group members have similar behaviour so those due to artifacts can be detected by examining their profiles to determine any acquisition order or class dependent effects. We are able to treat all group variables together so the behaviour of weak variables, if hard to determine directly, can be inferred from the more intense group members, and all members can be excluded from, or selected for, further analysis in one step.

3.3. Visualization

Despite the power of MVA algorithms, visualization is often the best and simplest way to detect trends in the data and recognize unexpected events; two aspects are especially important in our approach:

First, data points in graphs are depicted with symbols and labels that can be edited by the user in a meaningful way, for example, we often use filled shapes for post-dose samples and empty shapes for pre-dose. This allows us to make changes to emphasize particular events such as acquisition time dependent phenomena that are often more easily revealed by labelling the samples with their acquisition index.

Secondly, we make extensive use of profile plots that show the behaviour of a specific variable across all samples, and the ability to switch the display order between acquisition index and sample class. Profile plots can be used to directly examine the behaviour of a few variables prior to MVA or to review specific variables of interest following analysis.

4. Experimental

4.1. Data set

The concepts described here are illustrated using a simple data set that was generated by taking two urine samples (referred to as P and SB) from the same individual on the same day, and dividing each into two aliquots, one of which was spiked with a standard mixture comprised of minoxidil, safranin orange, buspirone, verapamil, reserpine, and sex pheromone peptide at a concentration of 100 pg/μL. Thus there were two spiked (P+ and SB+) and two unspiked (P− and SB−) samples that were analyzed six times in a randomized order by LC–MS on a QSTAR® Elite QqTOF (Applied Biosystems|MDS Sciex, Toronto, Ontario, Canada). Samples were diluted 1:5 (v/v) with buffer A comprised of 5% (v/v) acetonitrile and 0.1% (v/v) formic acid in water. 5 μL of sample was injected onto a Phemonenex Luna 5 μm C18 100 Å 150 mm × 3 mm column equilibrated with buffer A and a linear gradient of 5% buffer A to 60% buffer B (95%, v/v acetonitrile; 0.1%, v/v formic acid in water) over 25 min used to elute compounds from the column at a flow rate of 250 μL/min. The column was regenerated by washing for 5 min with 100% buffer B prior to re-equilibration with 5% buffer A. The mass spectrometer was scanned from 100 to 1000 amu in positive electrospray mode with an accumulation time of 0.5 s.

To investigate the ability to detect and evaluate significant changes in instrument performance (failure), a further 12 injections were performed with the declustering potential (DP) changed as follows: 60 V (2), 80 V (2), 120 V (2), 180 V (2) and 240 V (4) where the value in parentheses indicates the number of samples acquired at each value. Altering DP to higher values induces fragmentation prior to the mass analyzer so that the intensity of some variables decrease while others increase; at the highest values all ions in the observed mass range have been fragmented and no signal is observed.

4.2. Data processing

Data was analyzed using the MarkerView™ software [21,24] (Applied Biosystems|MDS Sciex, Toronto, Ontario, Canada) which performs feature extraction by peak finding for each sample, alignment using mass and retention time windows for the peaks, scaling, MVA (PCA, PCA-DA and t-tests) and results visualization. The

program also links back to the raw data so that differences can be directly visualized in spectra or chromatograms.

With appropriate peak finding and alignment parameters a total of 1503 peaks were generated from the complete data set (including simulated failures) which was reduced to 1253 by requiring that variables be present in at least two samples. These numbers include all detected peaks, i.e. without de-isotoping, adduct removal, etc.

PCVG was performed using a small specially developed program [24] that interacts with the MarkerView software to extract the data and return the assigned groups. For this data we used three PCs, an angle of 35°, and required that a group have at least three peaks. The only peaks considered were those where the length of the vector joining the variable to the origin in the loading space was at least 0.02.

5. Results

PCA was performed using Pareto scaling and the resulting scores and loadings plots are shown in Fig. 1. It is clear that PC2 separates the spiked samples (P+ and SB+, filled circles) from the unspiked (P− and SB−), and that there is some separation of the P (blue) and SB (red) samples. However, PC1, which explains the bulk of the variance, has some other cause arising from the variables that most affect PC1, such as those in the red square.

Fig. 2 shows the same scores plot but with the samples labelled according to acquisition index; the inset shows the small group of peaks with most negative PC1 loadings. This simple change immediately shows that PC1 is related to the acquisition order, implying that there is some time dependent change, and that the later samples have the most negative PC1 values. The lower part of the figure is the profile plot for the variables in the red rectangle in Fig. 1, also in order of acquisition index, which shows that the effect is due to a decrease in these variables (and an increase in variables with large negative PC1 loadings—not shown). Note that the sample symbols and labels are also used in the profile plot which shows that the effect is probably instrument related since all classes are affected similarly. Drawing this conclusion is facilitated by proper randomization and the use of technical replicates. In a real experiment it would be necessary at this point to determine the implications of this observation: is the study or experimental design still valid? Can the data be used, perhaps by ignoring the affected samples or variables, or by correcting their intensities in some other way,

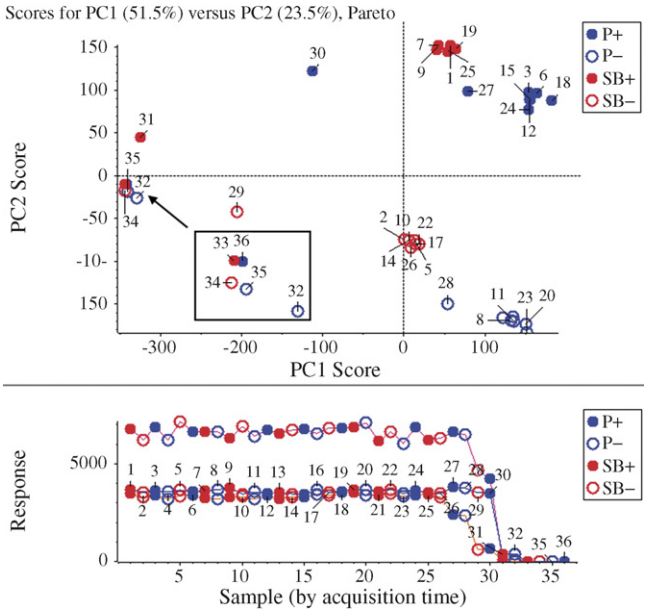

Fig. 2. Scores plot from Fig. 1 (top) with the samples labelled according to acquisition order. The inset shows the detail for the groups of samples with the most negative PC1 scores. The lower part of the plot shows the acquisition time order profiles for the variables with highest positive PC1 loadings and PC2 loadings close to zero (marked with a red rectangle in Fig. 1).

and will this retain sufficient statistical power? The profile plot also helps here since it is apparent that each class is well represented by unaffected variables, so simply ignoring the affected samples may be adequate. In this particular case, since the source of the variation is known to be the simulated instrument failure, ignoring the last 12 samples is acceptable although this may not always be true.

The results of PCA on the first 24 samples, containing 1157 peaks in at least 2 samples, are shown in Fig. 3 with sample labels indicating acquisition index. The PC1/2 scores plot (top left) shows that PC1 separates the spiked samples (filled circles) from the unspiked, and PC2 separates the P (blue) and SB (red) samples, although the 'clusters' are not as tight as might be expected from replicate analyses. PC2/3 (lower left) shows that PC3 results from an effect that is largely time dependent since the samples acquired later have the

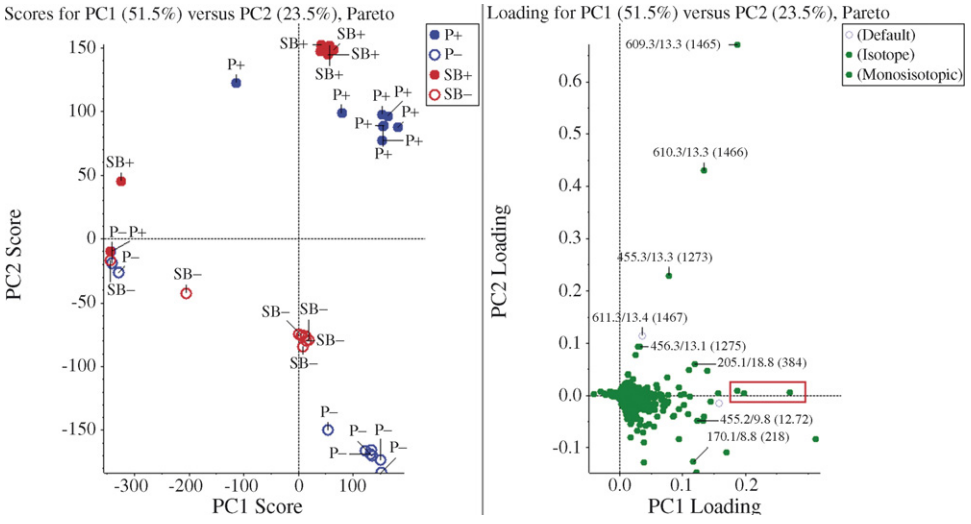

Fig. 1. Scores (left) and loadings plots (right) obtained from PCA of the entire data set (1253 variables). Symbols in the scores plot indicate the sample class (P+, SB+, P− and SB−), those in the loadings plot indicate the isotope status of the variable. For variables assigned to the 'default' class isotope information could not be found; at these masses it is likely that these are very weak monoisotopic peaks. The red rectangle indicates the variables most effecting PC1 only.

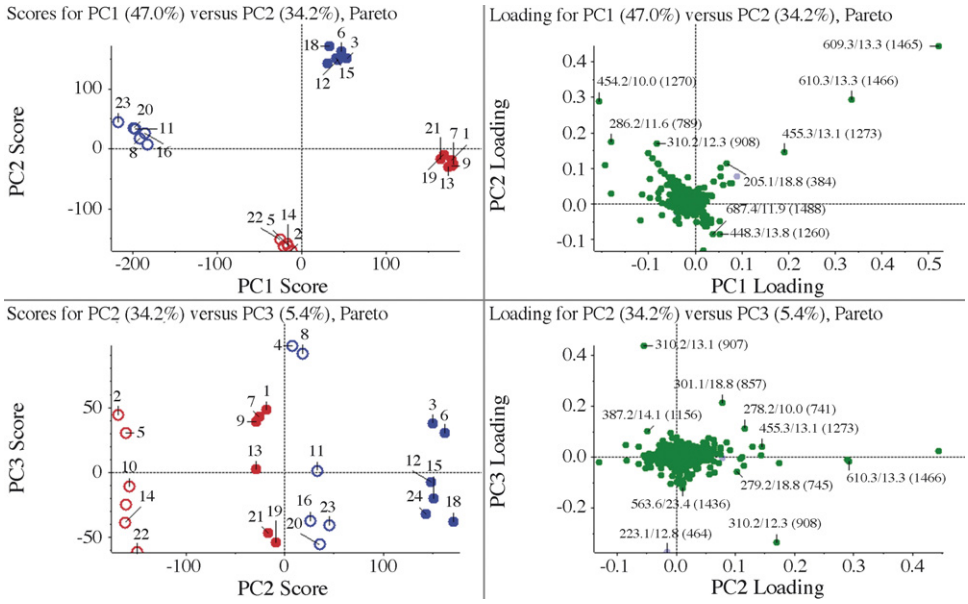

Fig. 3. Scores and loadings plots for the samples excluding those where instrument failure was simulated. The two upper panels show PC1/2 and the lower panels are for PC2/3. Sample symbols are as in Figs. 1 and 2 and the labels indicate acquisition index.

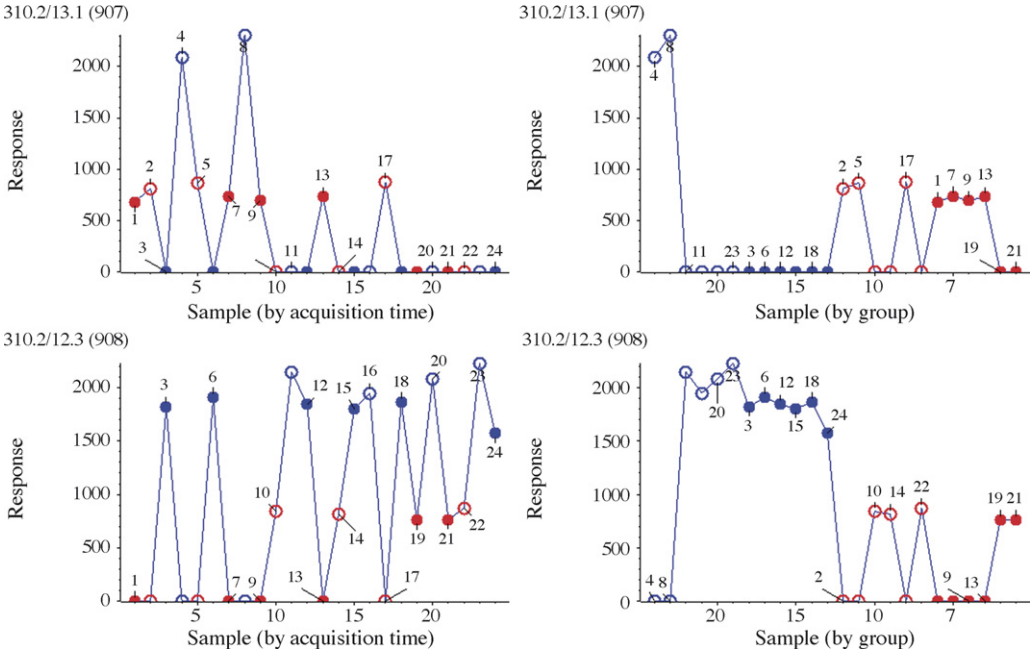

Fig. 4. Profile plots for "310.2/12.3" (lower panels) and "310.2/13.1" (upper) plotted in acquisition index order (left) and group order (right). Symbols are as in Figs. 1 and 2.

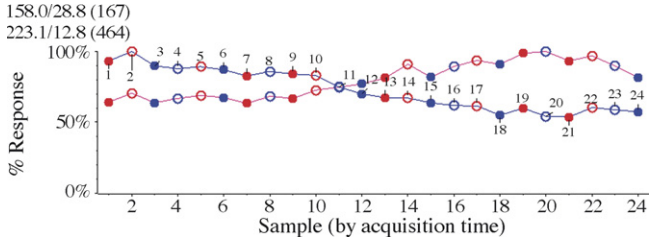

Fig. 5. Time ordered profiles for two variables in Fig. 4 with PC3 loadings of opposite sign. Symbols are as in Figs. 1 and 2.

highest negative PC3 loadings, however the two P− samples with acquisition indices 4 and 8 appear to be different. This underscores the importance of technical replicates, since there is no biological reason why these samples should be different to other members of the P− group. The high PC3 scores for these samples is caused primarily by the variable with the largest positive PC3 loading, i.e. "310.2/13.1 (907)" (upper part of lower right panel) and might arise from carryover from samples 3 and 7 which can quickly be assessed from the time ordered profile plot. As shown in Fig. 4 (top left) the intensity of this variable in samples 3 and 7 is less than in samples 4 and 8 indicating that carryover is not the cause. This figure also shows the time ordered profile for a second variable labelled "310.2/12.3 (908)" which is diametrically opposite in

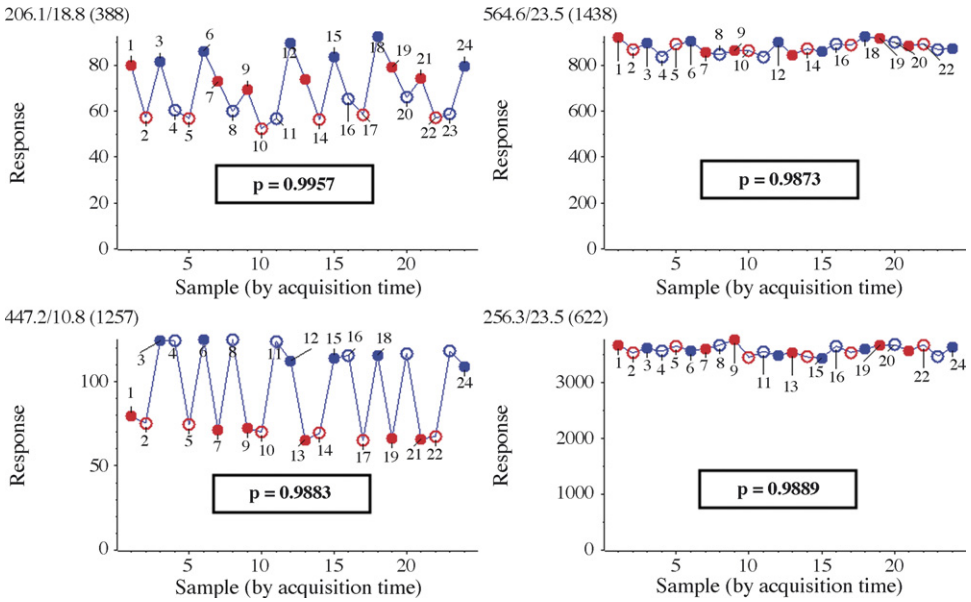

Fig. 6. Acquisition order profile plots for the four variables with the largest *p*-values (most constant) between the first five and last five samples acquired. Symbols are as in Figs. 1 and 2.

Fig. 3, although the mass is the same. We note from the left pair of plots in Fig. 4, that the behaviour of these variables is exactly anti-correlated, a phenomenon that is more obvious from the group ordered plots on the right side of the figure. We have observed that this behaviour is almost always due to peak alignment resulting from a poor choice of parameters or, as in this case, complex, noisy, convoluted peaks (data not shown) that have been assigned to different retention times. Using a larger retention time merge window (1 min cf. 0.4 min) in this case results in a single variable that is present in all samples but greater in P, and samples 4 and 8 are in line with the others.

Fig. 5 shows the time ordered profile plots for two other variables, one with a positive PC3 loading and the other with a negative loading. Clearly these variables are changing over time and contribute to the time dependence of PC3. It is important that they have opposite trends since this shows that instrument response is likely not the cause; given the length of the data acquisition (24 h) it is possible that decreases arise from unstable compounds and increases from the decomposition products, although compound identification would be required to confirm this. Note that normalization would have changed these profiles, potentially making the behaviour unrecognizable. In this example neither variable shows any class related trend so they can be safely ignored, but this may not always be the case and in reality the profile for every variable should be examined in order to determine whether the variable should be kept, removed or treated in some other way. In situations with hundreds or thousands of variables this would be slow and inaccurate so, as mentioned previously, we rely on *t*-tests and PCVG to help.

The *t*-test is used to determine if a given variable separates two classes, in this case ones that we have created to represent the first few, A, and last few, B, samples acquired. The test returns a *p*-value

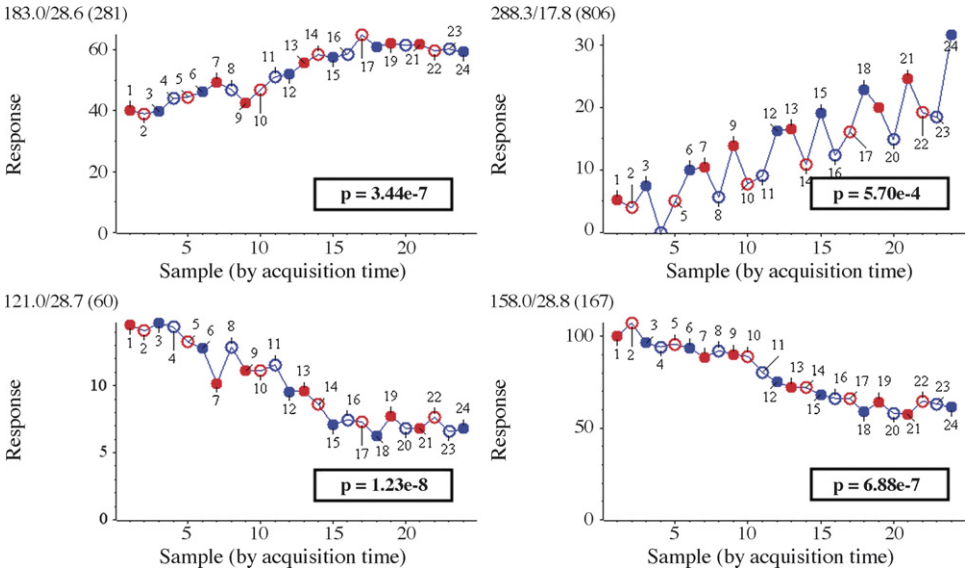

Fig. 7. Acquisition order profile plots for the three variables with the smallest *p*-values between the first five and last five samples acquired. The 30th variable is also shown (top right). Symbols are as in Figs. 1 and 2.

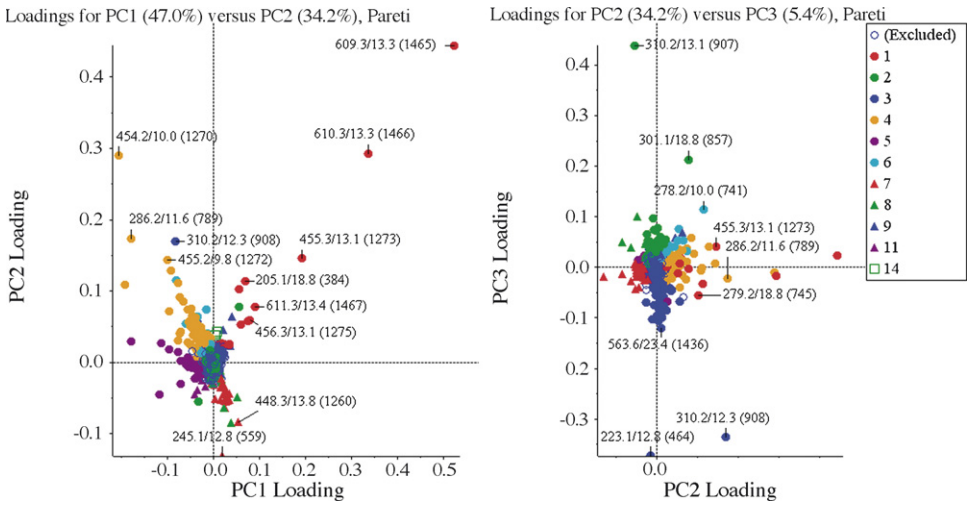

Fig. 8. Loadings plots for PC1/2 (left) and 2/3 (right) as in Fig. 3. The symbols indicate the automatically assigned groups using three PCs, an angle of 35°, and additionally requiring that a group have three variables that are 0.02 from the origin.

which corresponds to the probability that the observed change could be random. If, for a given variable, the *p*-value for the difference between A and B is very small the observed difference is most likely real, i.e. the variable has changed, whereas a value close to one indicates that the variable has not changed and therefore does not separate the groups. Thus we perform the *t*-test using A and B and examine variables that have *p*-values close to one as well as those that are very small. Since we are interested in variables that are present in the majority or all samples, we restricted the test to variables detected in at least 23 (of 24) samples.

Fig. 6 shows the acquisition time ordered profile plots for the four variables with the largest *p*-values, two of which (left panels) combine a constant offset with class differentiation. Fig. 7 shows the profile plots for the three variables with the smallest *p*-value (most different) as well as the 30th variable (top right) which is the first to show a general trend combined with class differentiation. Altogether there are 32 variables with *p*-values less than 10^{-3}.

Clearly using the *t*-test in this way can reveal artifacts that are constant across the analysis or show a systematic change, but it does not distinguish variables that combine this behaviour with class differentiation; this must be determined manually. Artifact variables without differentiation can often be rejected without damaging the data, but the others should be retained; differentiating variables with constant offsets (e.g. left panels in Fig. 6) will be correctly handled by mean-centering, but those on a changing background (e.g. top right, Fig. 7) will not be. Thus, it is not sufficient to define *p*-value thresholds and reject variables without visual examination, although this can be reasonably fast and the *p*-values do limit the number of variables that must be manually examined. Further, the *t*-test is supervised, i.e. we must define the 'classes' to be compared, and we pre-selected variables appearing in 23 or more samples so there are doubtless others with similar trends but low intensity such that they are only observed in 22 samples (or fewer).

In contrast, PCVG finds all variables that share the same pattern and, like PCA itself, reduces a large number of variables to a few groups that can be treated together—rather than looking at thousands of variables we look at just those groups. The technique uses PCA, and so is unsupervised, but the groups returned contain variables in their original context so interpretation is much simpler. For example, applying the technique to the data of Fig. 3 generated just 14 groups that represented 579 variables (the other 578 in the

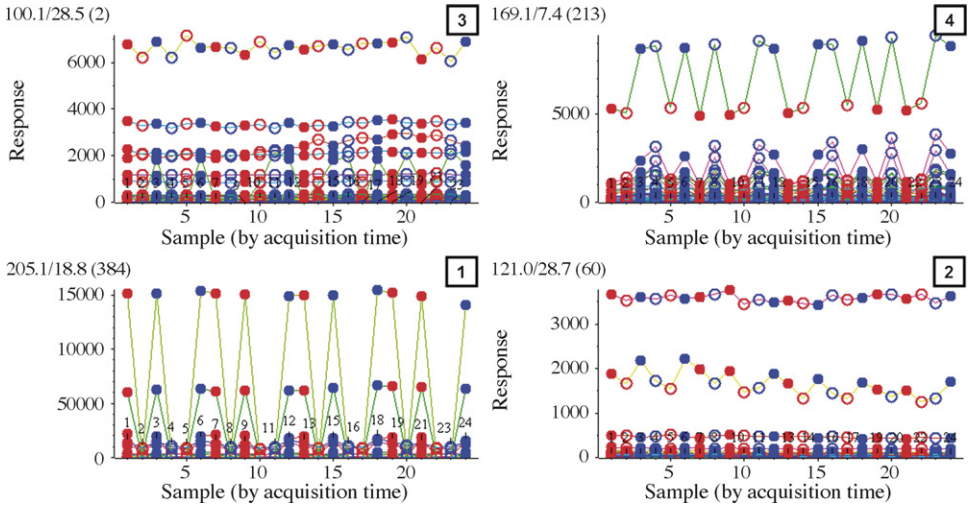

Fig. 9. Acquisition time order profile plots for the first four groups of Fig. 8. Symbols are as in Figs. 1 and 2.

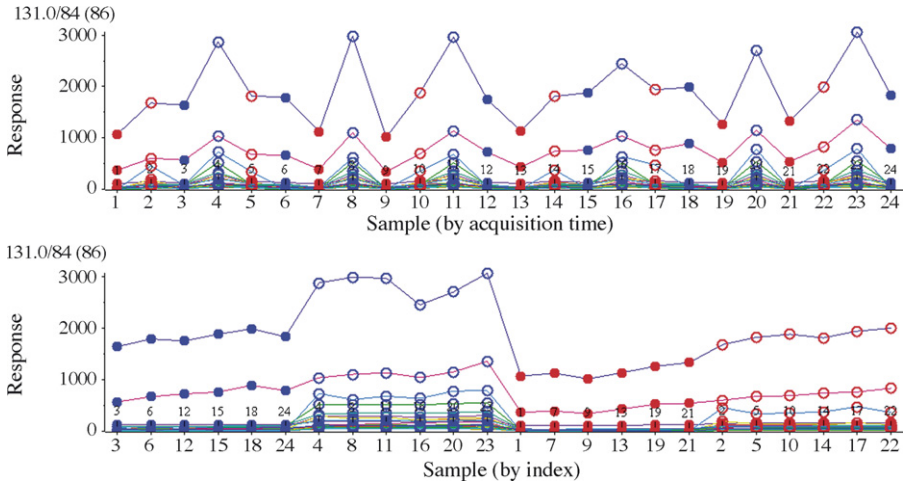

Fig. 10. Profiles for group 5, Fig. 8, ordered by acquisition time (upper) and by group, with acquisition time as a sub-order (lower). Symbols are as in Figs. 1 and 2.

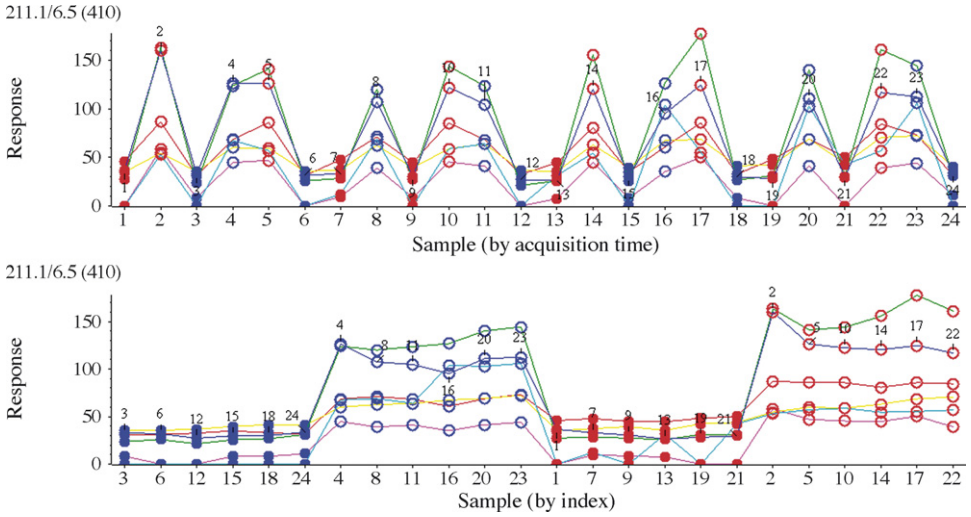

Fig. 11. Profiles for group 11, Fig. 8, ordered by acquisition time (upper) and by group, with acquisition time as a sub-order (lower). Symbols are as in Figs. 1 and 2.

original data have low intensity or low variance, are close to the origin of the loadings plot and are rejected by our minimum distance criterion). Fig. 8 shows the loadings plots from Fig. 3 (right panels), but with symbols used to indicate these 14 groups. Detailed examination of all of these groups is beyond the scope of this report, but relevant examples are given below.

Examination of Fig. 8 and the corresponding scores plots from Fig. 3, suggests that group 1 is related to the spike and that group 4 is related to the difference between the P and SB samples, regardless of spike (left panel), while groups 2 and 3 (right panel) are apparently related to the time-based change that PC3 depicts. Other behaviour can also become apparent, for example, the left panel shows that group 11 (purple triangles) appears to be anti-correlated with group 1, i.e. it distinguishes spiked and unspiked samples but with the opposite sense (higher in the unspiked). Fig. 9 shows the time order profiles for the first four groups assigned in Fig. 8 from which it is apparent that, as predicted, group 1 (bottom left) reflects differences between the spiked and unspiked samples, group 4 (top right) distinguishes the P samples from the SB, and groups 2 and 3 are gradual changes in opposite directions corresponding to decompositions that have little or no ability to distinguish the classes. Fig. 10 shows profiles for group 5 which combines class differentiation with a gradual change; in addition to acquisition time order (upper), the data is plotted in group order with each group sorted by acquisition time (lower) which in this case better reveals the characteristics. As a final example, Fig. 11 shows similar profiles for group 11 which is clearly higher in the unspiked samples than in the spiked. In general, variables that distinguish samples because of reduced intensity should be carefully examined to ensure that this is not caused by ion suppression, i.e. that there are no large peaks that co-elute with the species that appear to be reduced. If this is not the case, the raw data should be checked to verify that the peaks are present in the samples, and were correctly detected and aligned. Here, there is no obvious explanation for the observed behaviour although some of the six peaks in the group do show altered mass profiles, possibly due to saturation effects.

6. Conclusion

For LC–MS-based metabolomics to reveal underlying biological effects, other potential sources of variance must either be controlled or detected and treated. Some can be avoided or minimized by careful experimental technique, especially in sample collection, processing and storage, but artifacts introduced by the analytical system may be unavoidable. Under some circumstances, for example in the analysis of large sample batches requiring hours

or days of analysis time, the combined effects of these untreated artifacts may dominate sample separation. Artifacts such as background peaks introduced in the solvent that are constant across all samples have little variance and hence will not influence separation unless the intensity pattern is changed by normalization to a biological factor. Data treatment may also alter the profile of artifacts that change during the analysis so that they are no longer easily recognized. Hence we recommend that the presence of artifacts be determined prior to other processing so that they can be assessed, characterized and treated.

A variety of potential artifact sources are summarized in Table 1 and we have presented several tools and techniques that can be used to identify the resulting variables. Those that remain constant (background peaks) or change (decomposition, contamination wash out or build up, sensitivity changes) can often be identified using a *t*-test on artificial classes containing the first and last few samples analyzed. The power of this approach can be enhanced if the samples are known to be similar, for example because they are controls or from the same class. Instrument sensitivity changes can be ruled out if peaks with constant response are present, while carryover is best detected using acquisition ordered profile plots and carefully verifying the results from samples following those containing large peaks. Randomization and inconsistency in technical replicates can also help reveal carry over. Intense variables should be checked for other changes at the same retention time; anti-correlated peaks may be the result of ion suppression while correlated peaks may be due to ion enhancement. Suppression is often easier to detect and enhancement may be difficult to distinguish from a real compound increase. This may not be critical during biomarker discovery, but should be carefully validated with different chromatographic conditions before usage in other applications. Saturation effects may also be hard to detect, but instruments tend to have known ranges within which the intensity response is linear and values outside that range should be carefully examined.

We have found PCVG to be a powerful tool for a number of reasons, particularly the dimensionality reduction achieved by representing many variables by a few related groups (in our example 14 groups were obtained from nearly 600 variables). All members of a group share the same behaviour and can thus be characterized together by assigning a common symbol in loadings plots and by visualizing their profiles which, in addition to identifying artifacts, may indicate other effects (such as diurnal variation, xenobiotic metabolites, etc.) that the investigator can choose to pursue or ignore [24]. Groups also assist interpretation since related species (isotopes, adducts, unexpected fragments, etc.) will be present in the same group and will additionally have the same retention time.

Visualization of profile plots arranged in acquisition order and with flexible label and symbol assignment is often a straightforward way of identifying unexpected effects and can be used to examine variables directly or to review variables identified by the other techniques.

We recommend that approaches similar to those described here be applied to LC–MS data *before* response [6] or biological normalization so that artifacts, which would otherwise change and affect the analysis, can be detected. In many cases it is sufficient to reject the variables without damaging the data but others should be treated further before use. Discussion of tools appropriate to these latter cases will be the subject of further reports.

References

[1] M. Otto, Chemometrics, in: Statistics and Computer Application in Analytical Chemistry, Wiley-VCH, Weinheim, Germany, 1999.
[2] C. Bottcher, E.v. Roepenack-Lahaye, E. Willscher, D. Scheel, S. Clemens, Anal. Chem. 79 (2007) 1507.
[3] B.K. Matuszewski, M.L. Constanzer, C.M. Chavez-Eng, Anal. Chem. 75 (2003) 3019.
[4] M. Sysi-Aho, M. Katajamaa, L. Yetukuri, M. Orešic, BMC Bioinform. 8 (2007) e93.
[5] S. Bijlsma, I. Bobeldijk, E.R. Verheij, R. Ramaker, S. Kochhar, I.A. Macdonald, B. van Ommen, A.K. Smilde, Anal. Chem. 78 (2006) 567.
[6] J. van der Greef, S. Martin, P. Juhasz, A. Adourian, T. Plasterer, E.R. Verheij, R.N. McBurney, J. Proteom. Res. 6 (2007) 1540.
[7] H.G. Gika, G.A. Theodoridis, J.E. Wingate, I.D. Wilson, J. Proteom. Res. 6 (2007) 3291.
[8] M. Katajamaa, M. Orešic, J. Chromatogr. A 1158 (2007) 318.
[9] J. Listgarten, A. Emili, Mol. Cell. Proteom. 4 (2005) 419.
[10] A. Craig, O. Cloarec, E. Holmes, J.K. Nicholson, J.C. Lindon, Anal. Chem. 78 (2006) 2262.
[11] R.A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, BMC Genom. 7 (2006) 142.
[12] M.J. van der Werf, R.H. Jellema, T. Hankemeier, J. Ind. Microbiol. Biotechnol. 32 (2005) 234.
[13] Y.I. Shurubor, U. Paolucci, B.F. Krasnikov, W.R. Matson, B.S. Kristal, Metabolomics 1 (2005) 75.
[14] O. Fiehn, Plant Mol. Biol. 48 (2002) 155.
[15] N.M. Al-Kandari, I.T. Jolliffe, Environmetrics 16 (2005) 659.
[16] M. Daszykowski, W. Wu, A.W. Nicholls, R.J. Ball, T. Czekaj, B. Walczak, J. Chemometr. 21 (2007) 292.
[17] P. Jonsson, S.J. Bruce, T. Moritz, J. Trygg, M. Sjöström, R. Plumb, J. Granger, E. Maibaum, J.K. Nicholson, E. Holmes, H. Antti, Analyst 130 (2005) 701.
[18] M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, J. Selbig, Bioinformatics 20 (2004) 2447.
[19] I.T. Jolliffe, Principal Component Analysis, 2nd ed., Springer, New York, 2004.
[20] J.E. Jackson, A User's Guide to Principal Components, Wiley and Sons, New Jersey, 2003.
[21] T.P. Sangster, J.E. Wingate, L.L. Burton, F. Teichert, I.D. Wilson, Rapid Commun. Mass Spectrom. 21 (2007) 2965.
[22] L. Burton, R. Bonner, unpublished results.
[23] J.C.W. de Groot, M.W.E.J. Fiers, R.C.H.J. van Ham, A.H.P. America, Proteomics 8 (2008) 32.
[24] G. Ivosev, L. Burton, R. Bonner, Anal. Chem. 80 (2008) 4933.