Peak Alignment and Robust Principal Component Analysis of Gas Chromatograms of Fatty Acid Methyl Esters and Volatiles

Stina Frosch Møller and **Bo M. Jørgensen***

Danish Institute for Fisheries Research, DTU Build. 221, DK-2800 Kgs. Lyngby, Denmark

Abstract

Gas chromatograms of fatty acid methyl esters and of volatile lipid oxidation products from fish lipid extracts are analyzed by multivariate data analysis [principal component analysis (PCA)]. Peak alignment is necessary in order to include all sampled points of the chromatograms in the data set. The ability of robust algorithms to deal with outlier problems, including both sample-wise and element-wise outliers, and the advantages and drawbacks of two robust PCA methods, robust PCA (ROBPCA) and robust singular value decomposition when analysing these GC data were investigated. The results show that the usage of ROPCA is advantageous, compared with traditional PCA, when analysing the entire profile of chromatographic data in cases of sub-optimally aligned data. It also demonstrates how choosing the most robust PCA (sample or element-wise) depends on the type of outliers present in the data set.

Introduction

Chemometric tools, such as principal component analysis (PCA) for visualisation and data mining, are frequently used to analyse chromatographic data. In most cases, chromatographic data are transformed to peak areas, which are then used for further analysis. The method relies on subjective peak selection and peak identification and on integration parameters, which if not properly set, may cause great errors in the calculated peak areas. Implications of the data extraction method, thus, are incorporated in the PCA analysis. The disadvantages concerned with peak area analysis, such as loss of information due to the selection of a subset of peaks and to erroneous peak areas, can be avoided by using the entire chromatographic profile *per se* when analysing the data. In addition, peak shapes and information about the absence or presence of peaks are automatically included in the data analysis.

Unavoidable retention time shifts from one run to another obscure differences due to chemical variations between samples. Because multivariate data analysis requires uniform presentation of data [i.e., all data vectors have to be of the same length with corresponding elements (variables) representing similar phenomena in all samples], an appropriate pre-processing technique to align the chromatograms is needed. Variations, thus, are not dominated by shifts between variables but by different levels of the variables as they should.

Several retention time alignment algorithms have been reported in the literature (1–3). In the present study, the correlation optimization warping (COW) algorithm (2), originally developed as a data pre-processing step in multivariate modelling of chromatographic data. The COW algorithm has been successfully employed to align chromatograms from gas chromatography (GC)–flame ionization detection (FID) (3,4) and GC–mass spectrometry (5) measurements. According to Tomasi et al. (4), COW is less flexible than other warping methods, thus giving fewer artefacts and improving the quality of the alignment when applied to complex chromatographic data. COW allows aligning complex chromatograms with different number of peaks, peak intensities, and peak widths. Furthermore, it corrects peak shifts in both directions and aligns many chromatograms simultaneously, without any knowledge or identification of peaks.

PCA, like most other common chemometric methods, is based on the less robust least squares estimation. This means that the presence of even one single outlier in the data set can hamper the analysis and lead to incorrect conclusions. Outliers are measurements that do not fit into the pattern or grouping shown by the majority of measurements in a properly designed experiment. The most common outlier types are complete sample measurements (data vectors), but also individual "strange" data elements in the chromatogram may be considered as outliers.

The outlier problem can be solved in two ways: (*i*) by diagnostics or (*ii*) by robust estimators (6). In the first approach, outliers are identified and expelled from the data set prior to making the chemometric model. A complication is that it may be difficult to identify outliers, even when multivariate data are available, and

[*] Author to whom correspondence should be addressed: email boj@difres.dk.

the task gets harder and more time-consuming when the amount of data is huge. In the second approach, which is used in this paper, robust estimators are used instead of the ordinary non-robust least squares estimator. Robust methods reduce or remove the effect of outlying data points, allowing the remainder to predominantly determine the model.

In this study, the advantage of using all collected data points from the GC in the chemometric analysis combined with COW pre-processing is illustraited. Because of the outlier problem, concerning both sample-wise and element-wise outliers, the advantages and drawbacks of two robust PCA (ROBPCA) methods, ROBPCA (7) and robust singular value decomposition (RSVD) (8), are also investigated for the analysis of GC data. Opposite to the methods that rely on subjective peak selection and peak areas, the PCA analysis is able to identify relevant peaks and use all information contained in the chromatograms.

The analyses are performed on two data sets differing in quality. The first is GC–FID data from fatty acid methyl esters (FAME), which are "well behaved" in the sense that outliers are expected to be due to insufficient peak alignment only. The second data set consists of GC–FID data of volatile lipid oxidation productions (ATD), which have a relatively higher risk of artefacts and with larger sample differences and peak shifts.

Materials and Methods

Data sets

Gas chromatograms of FAMEs and of ATDs collected by dynamic head-space were kindly provided by the lipid group of the authors' institute. An FID was used for both types of chromatograms. The data from gas chromatograms of FAMEs show the fatty acid composition of triglycerides or phospholipids. In the present case, samples of fish oil from farmed rainbow trout fed two different diets were included. The data from gas chromatograms of ATDs show volatile lipid oxidation products (mostly aldehydes, ketones, and short-chain fatty acids). The samples included were from farmed rainbow trout kept frozen at –20°C, –30°C, or –80°C for 0–24 months. Detailed results concerning the experiments and the chemical findings are under preparation for publication.

The chromatograms were imported from the instrumental result files (ASCII text format) into MatLab 7.0.4 (The MathWorks) where the pre-processing (normalization, baseline correction, and alignment) and multivariate data analyses were performed. Each chromatogram was loaded into a MatLab workspace as a vector composed of the FID-signal collected over the duration of the GC run. The chromatograms were appended into a matrix where each row was the chromatogram from a single sample. The algorithms for COW and ROBPCA were downloaded from the literature (9,10). The algorithm for RSVD was kindly provided by A. Belousov (11).

Pre-processing of data

Pre-processing of the chromatograms prior to PCA is necessary to remove variations unrelated to chemical compositions. The pre-processing consists of baseline correction, normalization, and peak alignment using COW.

Baseline shift removal

Because baseline shifts affect both the warping and the normalization, a baseline correction is necessary. Furthermore, PCA cannot separate variance due to peak misalignment from variance due to baseline shifts. Hence, the baseline correction was essential. All chromatograms were individually baseline-corrected by subtracting the average signal for the last 1300 s and first 150 s, respectively, from the full chromatogram.

Normalization

Normalization to a constant area was used to compensate for differences in the amount of injected sample for Data set 1 (gas chromatograms of FAMEs), taking advantage of the unspecificity of the FID. Data set 2 (gas chromatograms of ATDs) was normalized by dividing each chromatogram by the injected amount of sample, giving informational value to the total amount of volatiles produced. In both cases, normalization was necessarily applied after baseline adjustment in order to give meaningful results.

Chromatographic alignment by COW

The aim of COW was to align two chromatographic profiles by piecewise linear stretching and compression, also known as warping, of the time axis of one of the profiles relative to the other. The chromatograms are subdivided into segments that were iteratively stretched and compressed by interpolation. The optimal alignment is the solution that maximizes the correlation between corresponding segments in the sample and the reference chromatogram. The number of data points each segment is allowed to change (maximal warping) is determined by the so-called slack parameter and depends on the peak shift to correct. According to Nielsen et al. (2) the optimal alignment will be achieved when the segment length is in the region of the number of data points making up the sharpest peak in the chromatogram.

The optimal chromatographic alignment settings in this study were selected as the segment length and slack that maximizes the first singular value as proposed by Christensen et al. (5). Combinations of segment lengths from 10 to 60 data points, and increments of 5 and slacks between 1 and 5 were tested to find the best settings. The optimal settings were based on the evaluation of the whole data set for the data from gas chromatograms of FAMEs and of 30 randomly selected samples for the data from gas chromatograms of ATDs.

PCA

The classical PCA method is not robust against outliers because of the least squares criterion. This means that even one single outlier in the data set can have an arbitrarily large effect on the model and lead to wrong interpretation and conclusions.

Different approaches have been proposed for making a robust version of PCA. They can be grouped as follows: (*i*) techniques that replace the classical covariance matrix by a robust covariance estimator (6,12,13) as the minimum covariance determinant (MCD) (14). Unfortunately, these approaches are limited to relatively low-dimensional data and are computational costly. (*ii*) Another group is methods that use projection pursuit (PP) techniques (15–20). PP searches for structure in high dimensional data by projecting these data into a lower-dimensional space that maximizes a robust measure of spread called the projection index. These methods can handle situations where the number of variables exceeds the number of samples. (*iii*) A combination of (*i*) and (*ii*) called ROBPCA (7) is used, which should yield more accurate estimates than the raw PP algorithm. The final group (*iv*) involves adjustments to the internal computations of the singular value decomposition (SVD) algorithm by replacing the least squares criterion with a robust estimate (8,21,22). These RSVD methods can handle high-dimensional data and element-wise outliers. Element-wise outliers exist where one or several individual data elements in otherwise good rows are corrupted.

In this study, the classical least square PCA will be compared with the two robust versions, ROBPCA and RSVD. Both robust methods can handle situations with more variables (columns) than samples (rows), are computationally feasible, and have shown good performance in other studies (17,23).

ROBPCA

The ROBPCA approach combines PP with robust covariance estimation in lower dimensions (7). The ROBPCA method can be divided into three major steps. First, the data, stored in an $n \cdot p$ data matrix X, were pre-processed by reducing their data space to the affine sub-space spanned by the n observations. This was performed by SVD of the column mean-centred X, without loss of information. In the next step of the ROBPCA algorithm, PP was used for initial dimension reduction ($k << p$). A measure of "outlyingness" was computed for each data point. The h data points with smallest outlyingness were then retained, the covariance matrix of this h-subset computed, and the number of principal components to retain (k) selected. In the last step of the ROBPCA algorithm, the re-weighted MCD estimator is then applied to this lower dimensional data space to find a robust center and covariance estimator of the projected samples. Finally, these estimates were back-transformed to the original space, and a robust estimate of the location of X and of its scatter were obtained.

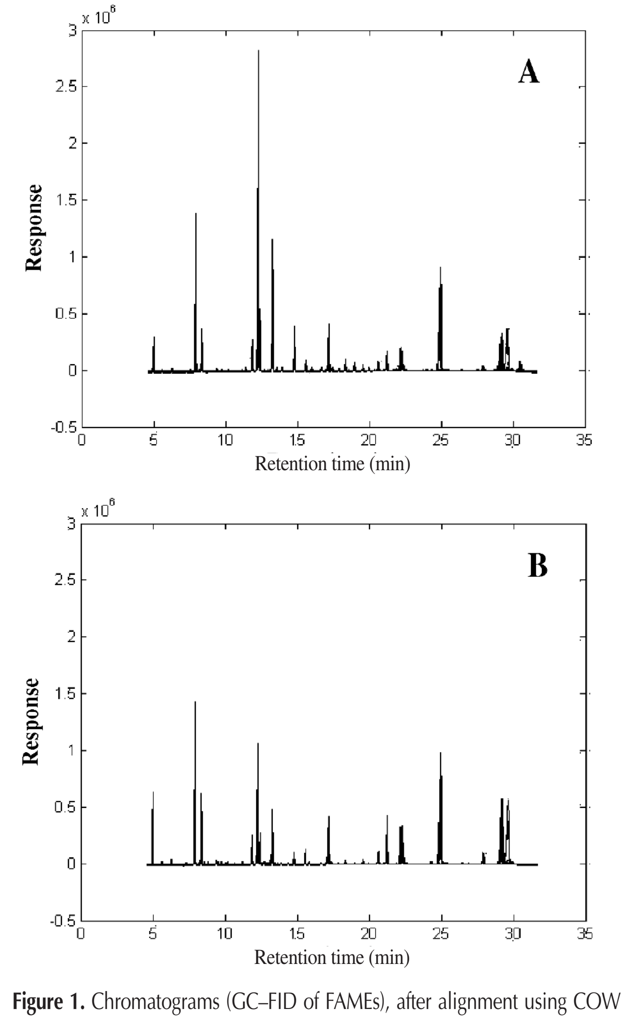

Figure 1. Chromatograms (GC–FID of FAMEs), after alignment using COW with a segment length of 15 data points and a slack of 3 points, of samples from fish fed on diets containing vegetable oil (A) or fish oil (B).

Robust singular value decomposition

This method, called RSVD (8), was based on the alternating least squares algorithm for SVD proposed by Gabriel (24). In this algorithm, the minimization problem was solved with criss-cross regressions, which involves iteratively computing dyadic (rank 1) fits using least squares regression. The original Gabriel–Zamir SVD algorithm is then rendered robust by substituting the non-robust least squares regression with a robust estimator, which in this case, was the alternating L1-norm (the sum of absolute residuals).

Results and Discussion

Data set 1 (GC–FID of FAMEs)

Optimal warping parameters

Figure 1 shows the aligned chromatograms appearing from fish whose feed contained mostly vegetable oil or pure fish oil, respectively. In all chromatograms, the same fatty acids appear, but with different concentrations, reflecting the different feed types. Fish fed vegetable oil contained higher amounts of 18:1 (n-9), 18:2 (n-6), and 18:3 (n-3) than did fish fed fish oil. On the other hand, fish fed fish oil contained the highest amount of 14:0, 16:0, 16:1 (n-7), 18:4 (n-3), 20:4 (n-3), 20:5 (n-3), 20:1 (n-9), 22:1 (n-11), 22:5 (n-3), and 22:6 (n-3). The relatively high amount of long chain polyunsaturated fatty acids in the fish fed vegetable oil is due to small amounts of fish meal in the feed.

The peak identified around 24.7 to 27.3 min in the un-warped data is due to an internal standard in some of the samples. This peak is isolated from the other peaks, and for that reason, it is possible to exclude the part of the chromatogram from the data analysis allowing samples both with and without an internal added standard to be included in the data matrix. If the part containing the standard was retained, severe artefacts in both the normalization step and in the following PCA modelling would occur.

The warping parameters segment length and slack were considered optimal when maximizing the first principal component from a PCA model fitted to the warped data. Combinations of segment lengths of 10 to 60 data points with increments of 5 data points and slacks between 1 and 5 were tested. Furthermore, the mean relative difference together with the maximal decrease and

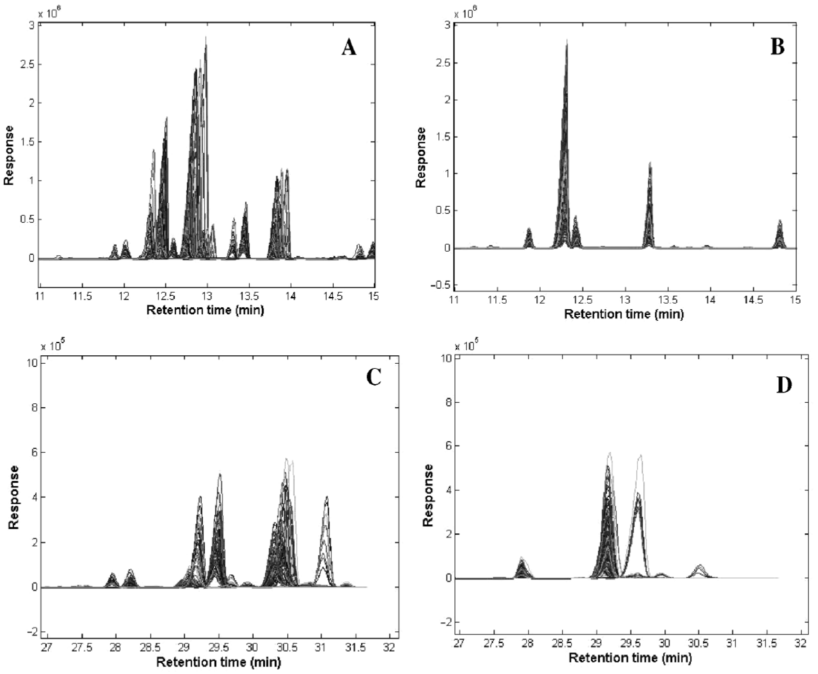

Figure 2. The effect of warping on two selected regions of the chromatograms (GC–FID of FAMEs): before warping (A and C); and after warping (B and D). For warping, COW was used with a segment length of 15 data points and a slack of 3 points.

increase in area difference between the un-warped and the warped chromatograms were calculated for all tested settings to evaluate the warping effect on the chromatogram profiles.

The explained variance for a one-component model increased from 30.6% (un-warped and un-centred data) to 87.2%, attained with a segment length of 15 data points and a slack of 3. The absolute area of the chromatograms after warping was changed, on average, with 4.4% compared with the original chromatograms with a maximal decrease in the area of 12.0%, and a maximal increase in area of 6.0%. These changes in area were due to interpolation when warping the data. Four of the samples experienced a decrease in area of more than 10% compared with the original chromatograms. When comparing the raw data of these samples with the standard chromatogram, it appeared that they had large shifts in retention times, resulting in the maximum warping allowed. In Figures 2A and 2B, the effect of warping was illustrated on a selected region of the chromatograms where the improvement by warping was pronounced. However, in the last part of the chromatograms, the improvement was not that good (Figures 2C and 2D). This misalignment was caused by larger shifts in retention time in the last part of the chromatograms and might be addressed by modifying the COW algorithm. The chromatograms might be split into several segments along the retention time axis and different warping parameters used for each of these segments.

Alternatively, misalignment may be dealt with by using RSVD, a method that only excludes outlying elements. This means that it was not necessary to exclude whole samples because of misalignment in some part of the chromatograms because the properly aligned parts of the chromatograms are still available for analysis.

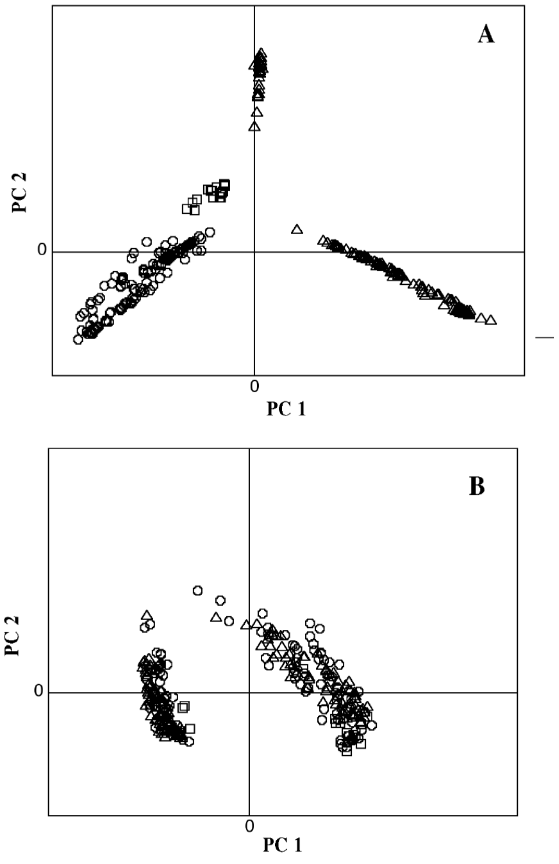

Figure 3. PCA scores: PC2 versus PC1, without warping (A) and with warping (B). The samples are marked according to frozen-storage time: 0 months (□), 4 months (○) and 24 months (△).

Principal component analysis

To investigate the effect of warping on the results obtained from PCA modelling, PCA was first applied to the mean-centered un-aligned data set. The score plot of PC1 versus PC2, from the model fitted to the un-aligned data, is presented in Figure 3A. Four distinct groups appear: three groups matching the storage period and time of analysis and one group where all samples belong to the same storage period, measured on the same day. Because of the experimental design, a confounding effect between storage period and time of analysis was unavoidable; it was, therefore, difficult to conclude if the grouping was due to storage period or time of analysis. When looking at the un-aligned chromatograms from samples stored for 24 months, a clear shift in retention time between the two groups appear, indicating that the clustering seen in Figure 3A was due to shifts in retention time, rather than chemical differences between the samples. Similar results were obtained when comparing the chromatograms for two groups separated along PC1. In Figure 3B the corresponding score plot of PC1 versus PC2, for a PCA

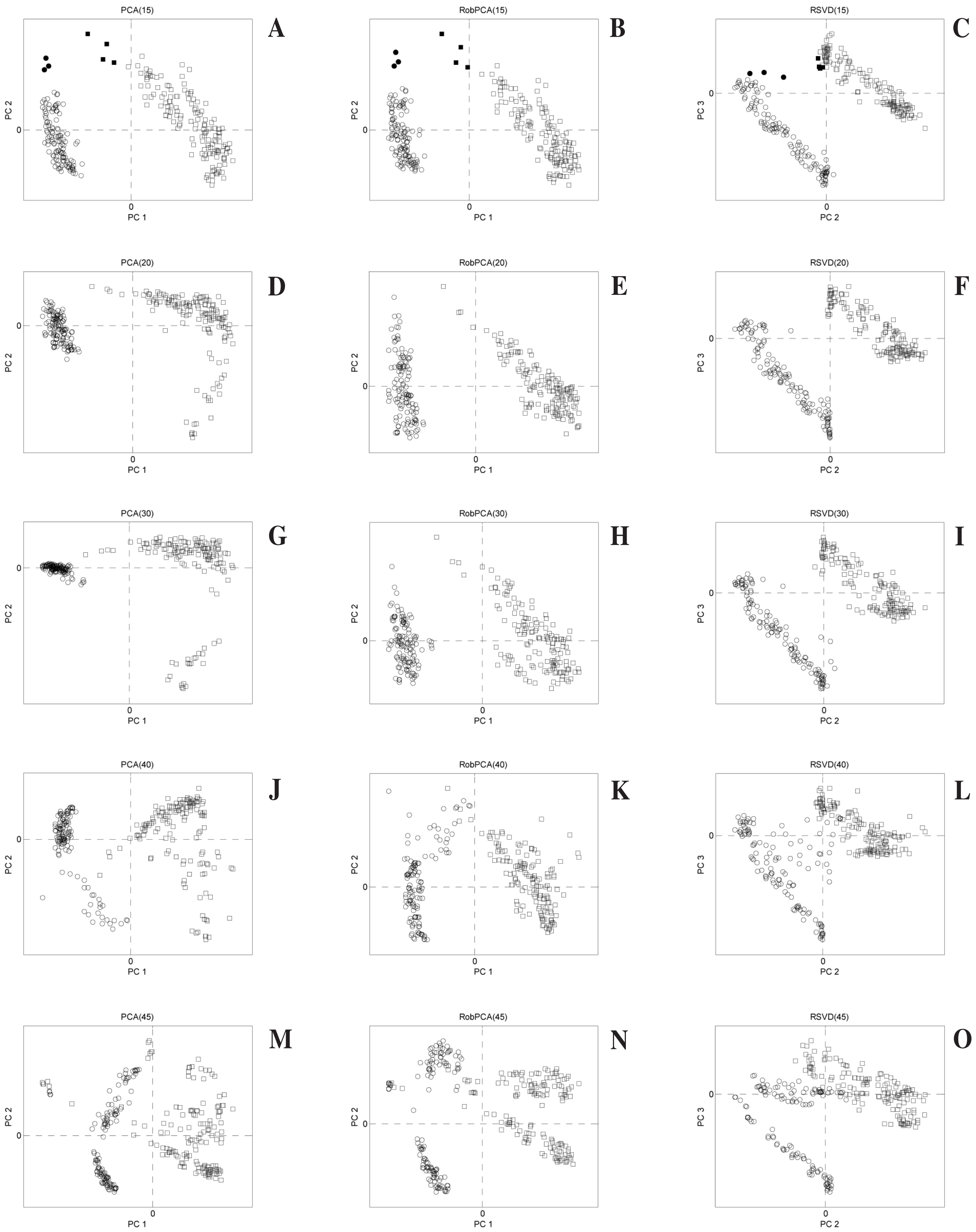

Figure 4. PCA scores: PC2 versus PC1 for classical PCA (column 1: A, D, G, J, and M) and for ROBPCA (column 2: B, E, H, K, N), and PC3 versus PC2 for RSVD (column 3: C, F, I, L, O). The chromatograms (GC-FID of FAMEs) were aligned by warping with the slack kept constant at 3 and varying segment lengths: 15 (A–C), 20 (D–F), 30 (G–I), 40 (J–L), and 45 (M–O) data points. The samples are marked according to oil type in the feed: vegetable oil (○) and fish oil (□). A few "extreme" samples are marked with filled symbols (A–C).

model fitted to the warped data, shows two groups only. These groups cannot be ascribed to a storage period or time of analysis, but they correlate to changes in the fatty acids profile caused by the different oils in the feed.

Furthermore, the loading plots for PC1 (37.6%) and PC2 (14.8%), from the un-warped data, showed complicated patterns, with many regions resembling the first derivative. This is typical for data distorted to a high degree by shifts in retention time (1). The shifts in retention time not only affect the first PC but also the subsequent components.

Thus, it was concluded that the pattern for the unaligned data was due to misalignments in the un-warped chromatograms rather than to chemical differences of the samples. For a reliable interpretation of the PCA model, alignment of the chromatograms is essential.

As illustrated in Figure 3, warping the chromatograms clearly improves a PCA, but it may be difficult to obtain optimal warping for all samples, especially in unsupervised situations. In that case, using robust PCA methods on the warped data may be helpful and provide better results than does the traditional PCA method, based, as it was, on least squares estimates. Moreover, even with perfectly aligned data, outliers may occur because of instrumental instability, etc. In this situation, the use of a robust methods was also of advantage.

In the score plot of PC1 versus PC2, both from traditional PCA and ROBPCA, clusters for each of the two treatments (vegetable or fish oil) were observed (Figure 4, first row). For both methods, PC1 scores discriminated between fish oil and vegetable oil, whereas PC2 scores displayed the variance between individuals in each group. Samples of fish fed vegetable oil were characterized by a high concentrations of 18:1 (n-9), 18:2 (n-6), and 18:3 (n-3), as their peaks in the chromatogram were positively loaded in PC1, and lower concentrations of 14:0, 16:0, 16:1 (n-7), 20:4 (n-3), 20:5 (n-3), 22:1 (n-11), and 22:6 (n-3), with peaks highly negatively loaded in PC1. The opposite results were obtained for samples of fish feed with fish oil.

In neither of the two models (traditional PCA and ROBPCA) was PC2 correlated to the experimental design, but this was primarily due to biological variation within the groups and to artefacts, such as a suboptimal baseline correction. No other groupings where found in higher order PCs. The difference in baseline was especially pronounced for the extreme samples with high score values in PC2 in both traditionally PCA and ROBPCA (filled symbols).

An even better class separation was obtained with elementwise robust PCA (Figure 4). No centering of the data was built in this RSVD algorithm, as was the case for ROBPCA, meaning that the first PC explained the centering of the data and was, for that reason, not interesting. PC2 and PC3 explained 60.0% and 22.1%, respectively, of the variance when PC1 was excluded, and these PCs are both relevant for the clustering. The same fatty acids, as found from the two previous models, were responsible for the clustering in Figure 4.

The explained variance in the first PC increases with ROBPCA 77.8%, compared with traditional PCA, 69.7%. The cumulative variance of the two components from RSVD, associated with the clustering due to different oils in the feed, was estimated to 82.1%. The variance was concentrated in the robust models, as a result of excluding outlying samples or outlying elements from the modelling step leading to increased class separation and reduced within-class variation.

In the former paragraphs it was illustrated that for well warped data, the results obtained with traditional PCA and ROBPCA were fairly good, even though the result can be improved by using the robust SVD method. Now, it will now be interesting to compare the PCA methods with decreasing data quality to investigate how well the data need to be aligned in order to yield acceptable results according to clustering. The data quality was based on the explained variance for the different warping parameters tested, fitting a one component model (PCA) to the normalized, but un-centred data (5). The slack was kept constant at 3, and the segment length was increased from 15 to 50 data points. The explained variance for a one component model when evaluating the warping parameters was: segment 20, 86.0%; segment 30, 84.4%; segment 40, 79.6%; segment 45, 72.4%; and segment 50, 67.0%.

The score plots in Figure 4 illustrate the effect of reduced data quality on the three different principal component analysis procedures. Results obtained for data warped with a segment length of 50 data points are not displayed, as they were similar to the results obtained with data warped with a segment length of 45 data points. A clustering according to different types of oil in the feed was observed for all three methods for data of high quality, although the clearest clustering was obtained with the two robust methods. With decreasing data quality (i.e., 79.6% explained variance and below in this case) the plot gets more unclear regardless of which PCA method was used to analyze the warped data. This clearly illustrates that data, and thereby the warping, need to be of a certain quality to obtain reliable results. The robust methods can not remedy problems with large shifts in retention time.

Data set 2 (GC–FID of ATDs)

Optimal warping parameters

Figure 5 shows aligned chromatograms for samples stored at –20°C and –80°C. The profiles and the total amount of oxidation products depend strongly on the storage temperature, as would be expected. The number of peaks and their areas are much higher for samples stored at –20°C than for those stored at –80°C (The storage time was 24 months in both cases).

The highest obtained explained variance for a one component un-centred PCA model was 80.1%, attained with a segment length of 20 data points and a slack of 3. In comparison, the explained variance for a one component model of un-warped and un-centred data was only 65.8%.

Principal component analysis

The score values of PC1 and PC2 from both traditional PCA and ROBPCA, as well as of PC2 and PC3 from RSVD, are shown in Figure 6. The samples are marked according to their storage temperature. For all three models, PC1 scores (PC2 for RSVD) turned out to be reasonable in storage temperature. The scores went from one sign to the other related to storage temperatures from –80°C or –30°C to –20°C. The clearest grouping according to storage temperature, –80°C or –30°C versus –20°C was observed with RSVD. No big difference in PC1 scores was observed between classical PCA and ROBPCA. Three outlying samples were separated from the other samples along PC2 (PC3 for RSVD). With ROBPCA, the three outliers are excluded from the modelling step and are placed closer to the other samples. Additionally, the variation accounted for by PC2 scores (PC3 for RSVD) was due to variation within each storage time, reflecting the biological variation. It was not possible to identify other patterns in the data by plotting other combinations of principal components.

The explained variance for PC1 and PC2 was 62.1% and 19.4%, respectively, for classical PCA and 76.3% and 11.%, respectively, for ROBPCA, resulting in a slightly higher explained variance for a two component model when applying ROBPCA. For RSVD, the explained variance for PC2 and PC3 was 6.0% and 22.1%, respectively. The low explained variance was a result of the presence of the outlying samples; only the first principal component was associated with a common variation between all samples, whereas the following components were primarily associated with the outlying samples. PC5 from RSVD accounted for 23.5% of the explained variance and was only caused by the three outlying samples (results not shown).

The chromatographic profiles of the three outliers were almost identical. A comparison of the chromatograms from the three outliers with the other samples stored at –30°C showed that the profile from the outliers were outstanding from the other chromatograms, with some peaks reaching higher or lower intensities, whereas other peaks were missing or only found for the three outliers. The full data vectors of these samples may, therefore, be regarded as outliers. This can also explain why the robust SVD method was not able to handle these outliers efficiently. All elements from the sample ought to be excluded, but the method can "only" handle up to 50% outlying elements in each data vector. The data set was not perfectly warped, meaning that all peaks are not perfectly warped and outlying elements exists. This is why different groupings are observed

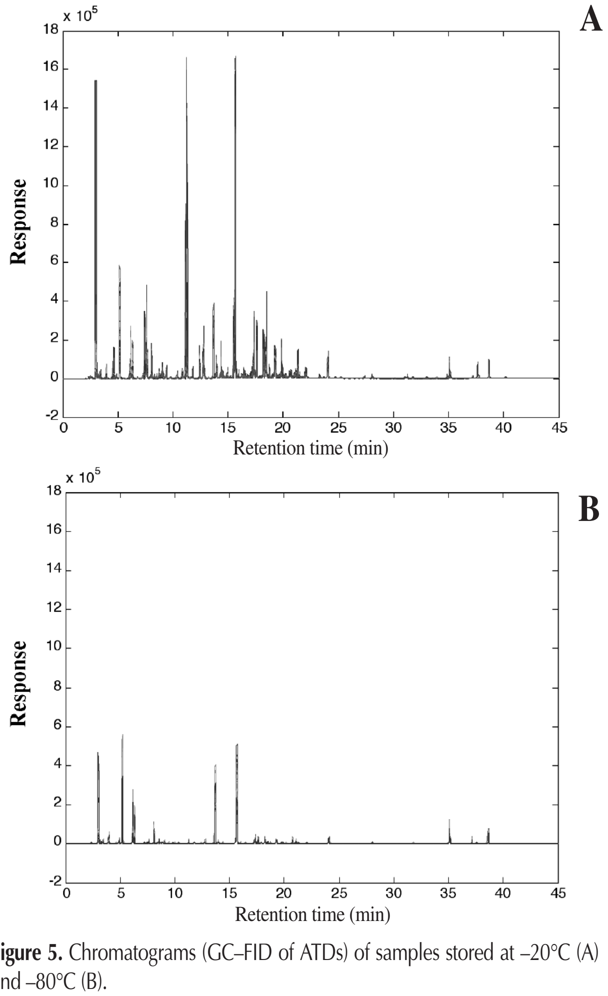

Figure 5. Chromatograms (GC–FID of ATDs) of samples stored at –20°C (A) and –80°C (B).

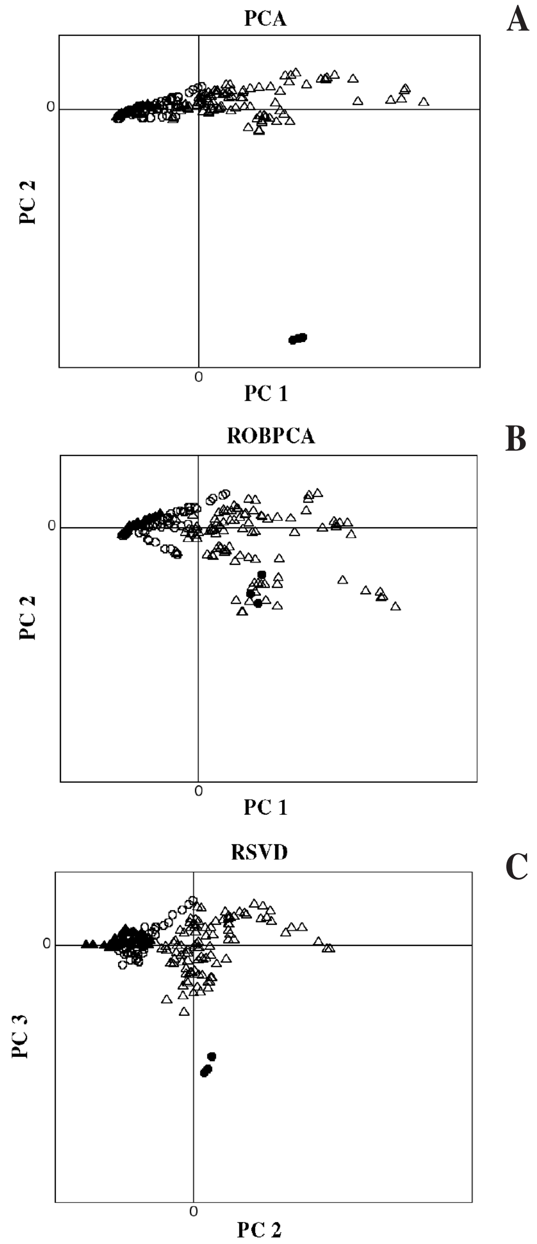

Figure 6. PCA scores: PC2 versus PC1 for classical PCA (A), ROBPCA (B), and PC3 versus PC2 for RSVD (C). The chromatograms (GC-FID of ATDs) were aligned by warping with a slack of 3 and a segment length of 20 data points. The samples are marked according to storage temperature: –20°C (△), –30°C (○), and –80°C (▲). Three outliers (all –30°C samples) are marked with filled circles.

between ROBPCA and RSVD in the actual situation: in ROBPCA, the entire sample is excluded from the modelling step, leaving out the three outliers completely and thereby assigning PC2 to another, perhaps more interesting, variation.

Conclusion

In designed experiments where one looks at a whole set of chromatograms at a time, multivariate data analysis is a useful alternative to classical peak selection and area calculation procedures. Alignment of the chromatograms is necessary and may, to a large extent, be done by automatic procedures. In situations where only suboptimal alignment is obtained, or other situations where outlying measurements occur (e.g., because of bad baselines or errors in sample amount injected) robust algorithms are to be preferred in order to keep the outliers from severely interfering with the multivariate models. Situations where only some part of the chromatograms are not properly aligned are best dealt with by using element-wise robust methods (e.g., RSVD). When the outliers are due to features throughout the chromatogram, sample-wise robust methods (e.g., ROBPCA) perform the best.

Acknowledgment

This work was supported by a grant to SFM from the Danish Ministry of Food, Agriculture, and Fisheries. Gas chromatograms of FAMEs and of ATDs collected by dynamic headspace were kindly provided by the lipid group (C. Jacobsen) of the authors' institute.

References

1. G. Malmquist and R. Danielsson. Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods. *J. Chromatogr. A* **687:** 71–88 (1994).
2. N.-P.V. Nielsen, J.M. Carstensen, and J. Smedsgaard. Aligning of single and multiple wavelength chromatographic profile for chemometric data analysis using correlation optimized warping. *J. Chromatogr. A* **805:** 17–35 (1998).
3. K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, and R.E. Synovec. Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis. *J. Chromatog. A* in press (2007).
4. G. Tomasi, F. van den Berg, and C. Andersson. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemomet.* **18:** 231–41 (2004).
5. J.H. Christensen, G. Tomasi, and A.B. Hansen. Chemical fingerprinting of petroleum biomarkers using time warping and PCA. *Environ. Sci. Technol.* **39:** 255–60 (2005).
6. P. Rousseuw and A.M. Leroy. *Robust Regression and Outlier Detection.* John Wiley & Sons, New York, NY, 1987.
7. M. Hubert, P.J. Rousseuw, and K. Vanden Branden. ROBPCA: a new approach to robust principal component analysis. *Technometrics* **47:** 64–79 (2005).
8. D.M. Hawkins, L. Liu, and G.S. Young. "Robust singular value decomposition", National Institute of Statistical Sciences, Technical Report 122. Research Triangle Park, NC, 2001.
9. http://www.models.kvl.dk. Date accessed (January 2006).
10. http://www.wis.kuleuven.ac.be/stat/robust.html. Date accessed (January 2006).
11. A. Belousov. Westfalischen Wilhelms University. Personal Communication (2006).
12. P.L. Davies. Asymptotic behavior of S-estimators of multivariate location and dispersion matrices. *Ann. Statist.* **15:** 1269–92 (1987).
13. C. Croux and G. Haesbroeck. Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika* **87:** 603–18 (2000).
14. P. Rousseeuw. Least median of squares regression. *J. Am. Statist. Assoc.* **79:** 871–80 (1984).
15. G. Li and Z. Chen. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. *J. Am. Statist. Assoc.* **80:** 759–66 (1985).
16. J.S. Galpin and D.M. Hawkins. Methods of L1 estimation of a covariance matrix. *Comput. Statist. Data Anal.* **5:** 305–19 (1987).
17. Y. Xie, J. Wang, Y. Liang, L. Sun, X. Song, and R. Yu. Robust principal component analysis by projection pursuit. *J. Chemometrics* **7:** 527–41 (1993).
18. C. Croux and A. Ruiz-Gazen. A fast algorithm for robust principal components based on projection pursuit. COMPSTAT, Physica-Verlag, Heidelberg, Germany, 1996, pp. 211 – 216.
19. M. Hubert, P.J. Rousseeuw, and S. Verboven. A fast method for robust principal components with applications to chemometrics. *Chemometrics Intell. Lab. Syst.* **60:** 101–11 (2002).
20. C. Croux and A. Ruiz-Gazen. High breakdown estimators for principal components: the projection-pursuit approach revisited. *J. Multivariate Anal.* **95:** 206–26 (2005).
21. L. Liu, D. Hawkins, S. Ghosh, and S. Young. Robust singular value decomposition analysis of microarray data. *P. Natl. Acad. Sci. USA* **11:** 13167–72 (2003).
22. C. Croux, P. Filzmoser, G. Pison, and P.J. Rousseeuw. Fitting multiplicative models by robust alternating regressions. *Stat. Comput.* **13:** 23–36 (2003).
23. M. Hubert and S. Englen. Robust PCA and classification in bioscience. *Bioinformatics* **20:** 1728–36 (2004).
24. K.R. Gabriel and S. Zamir. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics* **21:** 489–97 (1979).

Manuscript received January 13, 2006.