Supplementary material

An automated method for baseline correction, peak finding and peak grouping in chromatographic data"

Lea G. Johnsen*[ab], Thomas Skov[a],

Ulf Houlberg[b] and Rasmus Bro[a]

[a] *Dept. Food Science, University of Copenhagen, Denmark.*
Tel: 0045 3533 3222; E-mail: rb@life.ku.dk
[b] *Chr. Hansen A/S, Bøge Alle 10-12, 2970 Hørsholm, Denmark.*
Tel: 0045 4574 7474; E-mail: dklgj@chr-hansen.com

Comparison of peak finding algorithms

The four peak finding algorithms used in this section is the algorithm used in FastChrom and the d0, d2, and d2r algorithms used in the peak finding algorithm in PLS_toolbox (Eigenvector). The d0 version uses smoothened data as underlying basis for the peak finding, while d2 and d2r uses the smoothened second derivative as basis. The difference between d2 and d2r is that d2r is using an estimation of noise to validation of the found peaks.

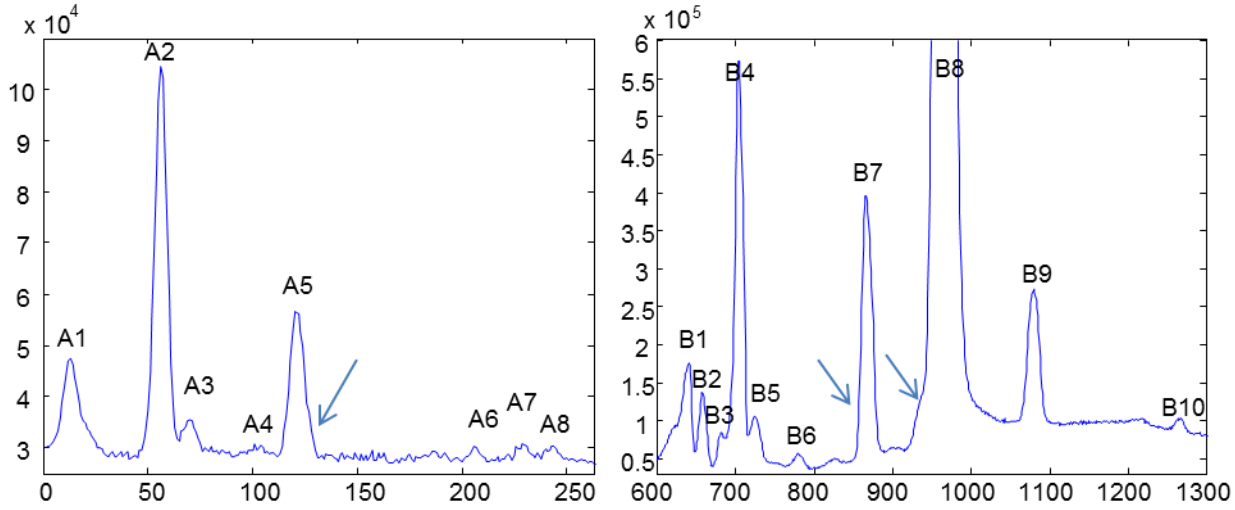

Figure S1. Two different chromatograms with different signal-to-noise levels. The peaks are numbered, and in **Table S1** it is indicated which peaks that can be found by the different peak finding algorithms. The arrows indicate shoulders on peaks.

Evaluation of how the four peak finding algorithms perform on the two chromatograms is shown in **Figure** S1. It is seen that they are all able to detect the large and well behaving peaks. However, only the algorithm incorporated into FastChrom and the d0 version are able to detect the low signal-to-noise peaks. On the other hand, only the d2 and d2r versions are able to detect the shoulders indicated with arrows in the figure. Based on these observations the d2 and d2r versions are the most suitable algorithms if the goal is to resolve overlapping peaks. However, this is not the case here, and therefore the algorithm incorporated into FastChrom or the d0 version seem like the most appropriate methods since these algorithms are better at detecting small peaks.

Table S1. Indication of which peaks it is possible to find with the three tested peak finding algorithms. FC; the algorithm implemented in FastChrom. d0, d2 and d2r; three different versions of the peak finding algorithm available in PLS_toolbox (Eigenvector).

	A1	A2	A3	A4	A5	A6	A7	A8	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
FC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
d0	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X
d2		X			X				X	X	X	X	X		X	X	X	
d2r		X							X	X	X	X			X	X	X	

Comparison of peak heights in known and obtained signal

Figure S2 shows the deviation between the peak heights in the known (simulated) signals and the signals obtained after removal of baseline contribution with the four different methods (AIMA, Quantile, FastChrom, and ALS).

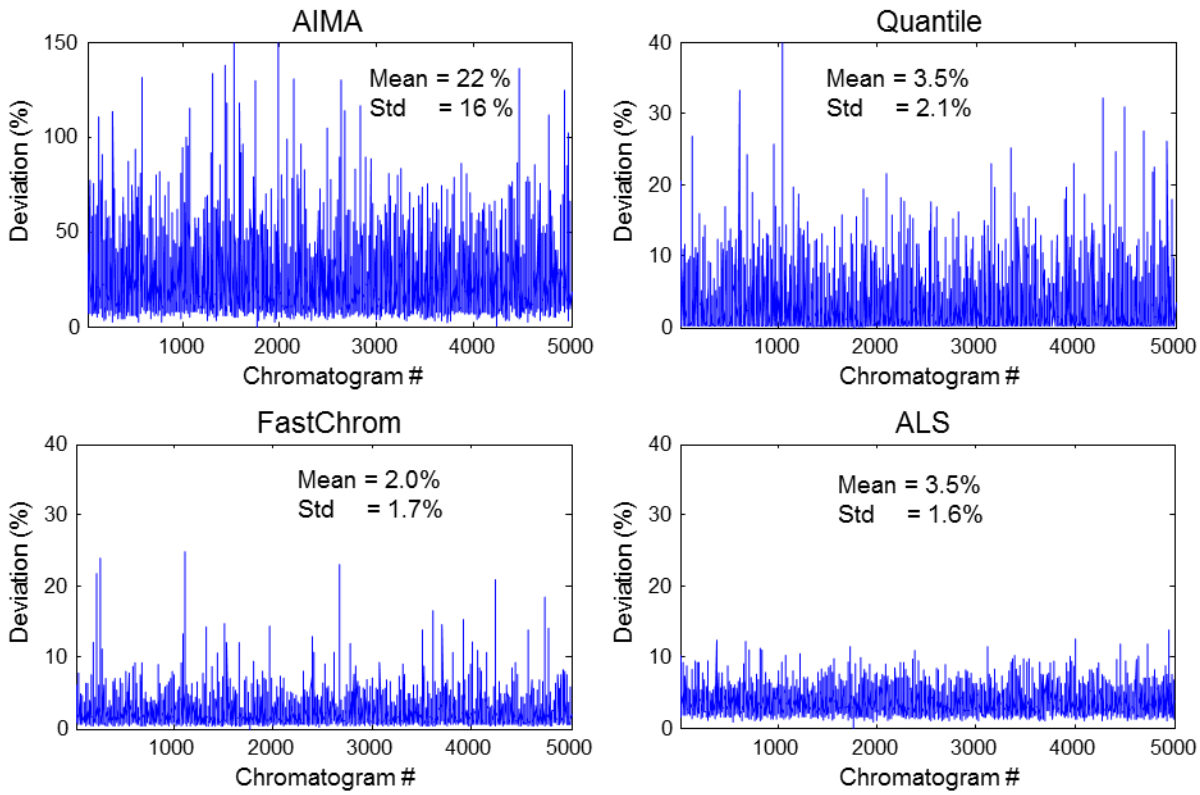

Figure S2. The mean deviation between peak heights in known and obtained signal in each of the 5000 simulated chromatograms. The deviation is indicated as percentage.

Complete FastChrom with ALS

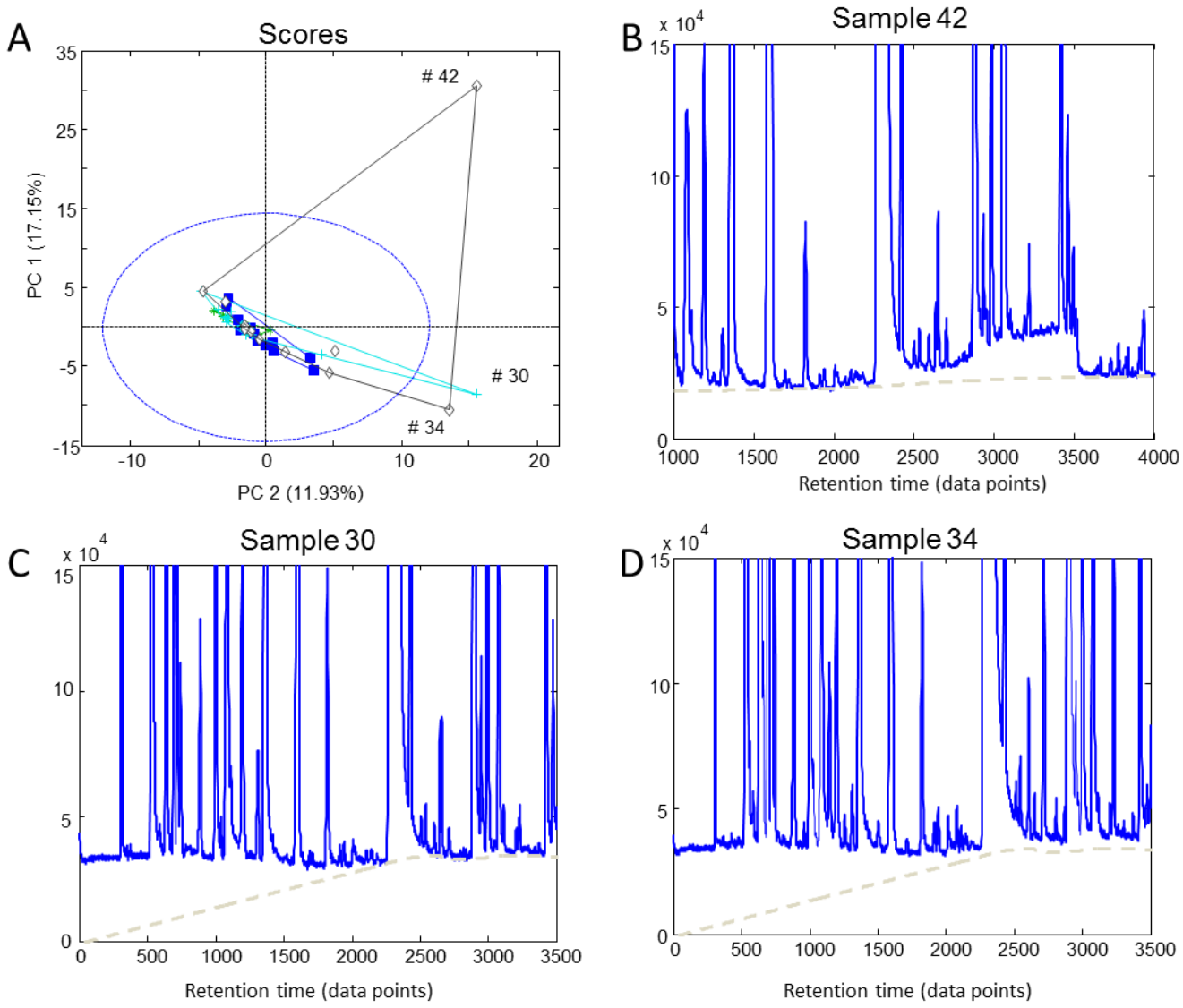

Figure S3. A: PCA calculated on the peak heights obtained by applying FastChrom with ALS baseline estimation. B-D: The estimated baseline (dotted lines) and chromatograms (full lines) of the outlier samples from A. The examples show that several peaks are influenced by the inadequate performance of ALS, and hence a PCA is severely affected.

Figure S3 shows a score plot from a PCA calculated on the peak heights found by FastChrom with ALS as baseline estimation technique. The score plot indicates that three samples are deviating a lot from the remaining samples. As shown in the paper this is not the case in the score plot from a PCA calculated on peak heights found by FastChrom with the newly developed FastChrom baseline estimation. Figure S3 B to D shows the ALS estimated baselines for the three deviating samples. It is clear to see that the baseline estimation not is performing optimally for these samples and that several peaks are affected by the non-optimal baseline estimation. These observations indicates that it is merely wrong estimations of peak heights that are the reason for these samples to appear as

deviating and not differences in the chemical composition. The same problem has been observed when using Quantile regression, but not for the FastChrom baseline estimation.