Chromatographia (2011) 73:721–731
DOI 10.1007/s10337-011-1962-1

ORIGINAL

Comparison of Several Methods of Chromatographic Baseline Removal with a New Approach Based on Quantile Regression

Ł. Komsta

Received: 3 November 2010 / Revised: 29 December 2010 / Accepted: 4 February 2011 / Published online: 27 February 2011

Abstract The article is intended to introduce and discuss a new quantile regression method for baseline detrending of chromatographic signals. It is compared with current methods based on polynomial fitting, spline fitting, LOESS, and Whittaker smoother, each with thresholding and reweighting approach. For curve flexibility selection in existing algorithms, a new method based on skewness of the residuals is successfully applied. The computational efficiency of all approaches is also discussed. The newly introduced methods could be preferred to visible better performance and short computational time. The other algorithms behave in comparable way, and polynomial regression can be here preferred due to short computational time.

Keywords Baseline drift · Background drift · Splines · LOESS · Quantile regression · Quantile smoothing · Whittaker smoother

Introduction

The background drift is one of the important issues in chemometric data processing of chromatograms in "signal-like" manner. Such drift can significantly affect the performance of chemometric algorithms, similar to (or together with) a noise. While the denoising of chemometric signals is a well established topic in the literature, only several reports on the baseline problem exists. Moreover, the articles regarding baseline are spread in specialized journals, not only chromatographic ones, as the baseline estimation methods are universal and not restricted to particular (chromatographic, spectral etc.) signals. In addition, no comparative study of proposed methods in chromatographic context was performed, and no "holistic view" of the problem is published till now.

The literature can be traced back to late 1970s, when Pearson reported the first often cited baseline estimation method [1]. It is based on distinguishing of the data points to baseline points and peak points. The algorithm runs iteratively and checks which points lie in a specific interval related to the standard deviation of them. The process is repeated until convergence is reached; next the smooth curve is fitted to baseline points only. Although the algorithm is very computationally efficient (which was a critical requirement in these years), it requires choosing of two parameters (denoted u and v), convergence criterion, and finally a type of smooth curve fitted to estimated baseline points. As the wrong selection of these parameters can lead to unacceptable results, several manual experiments need to be performed and the analyst can still be unsure if the signals are very complex.

About 10 years later, Dietrich et al. [2] proposed another modified method of baseline estimation. The signal is transformed into the second derivative, as its absolute value (power spectrum) is significantly higher than zero in peaks region. The method was successfully applied to NMR signals with fifth degree polynomial as the fitted curve. This method requires the optimization of the differentiation process (for example, width of Savitzky–Golay filter); moreover, it can be unstable when the peaks are broad.

Another approach was presented next by Moore and Jorgenson [3]. They recommend the use of a running median filter with a very broad window. Although this method is successful and simple, it can only be successfully

Ł. Komsta (✉)
Department of Medicinal Chemistry, Medical University
of Lublin, Jaczewskiego 4, 20-090 Lublin, Poland
e-mail: lukasz.komsta@umlub.pl

applied to signals with narrow peaks separated by wide baseline segments. Broad peaks and peaks concentrated together lead to bad baseline estimation. The authors compared this method with the Butterworth and Chebyshev filters (applied twice to remove phase shifts), which perform very unpredictable in baseline filtering and their use should be avoided (there is no easy way to select cutoff frequency, and improper selection can result in distorted signal). The median approach was then extended by Friedrichs [4] to NMR spectra. Although good performance was achieved, this approach still cannot be used as universal baseline extractor, because it works only with narrow and well-separated peaks.

The 1990s resulted in another ways to extraction of baseline. Brown [5] proposed the use of Bernstein polynomials to NMR signals. Andrew et al. [6] applied Kalman filter to extract baseline drift in ultraviolet region in context of multivariate calibration. Golotvin and Williams [7] extended the method proposed by Friedrichs [4]. There were also some wavelet-based approaches to the problem in ICP-AES [9] or chromatography [10]. Both require the use of a specialized wavelet software and experimental choosing of a mother wavelet and a decomposition level. The main disadvantage, also noticed by authors, is the significance of these parameters—wrong a priori selection can fail completely the baseline extraction.

The late 1990s resulted in significant increase of computational power available to an average scientist and the iterative methods were proposed for estimation of baseline. They fit a smooth trend to the signal, which falls down to the baseline level after several iterations. Their use was probably not recommended earlier due to unacceptable computational time. Ruckstuhl et al. [8] proposed iteratively reweighted local regression (similar to LOESS algorithm), especially for coping with baseline in spectral data. Gan et al. [11] smoothed the signals with polynomial. After each iteration the estimated function was used as a threshold cutting original data until convergence. Daszykowski and Walczak [12] recommended the use of reweighted Whittaker smoother, discovered for chemometrics by Eilers [13]. A recent article published in this year [14] presents also this approach with ready-to-use software. The iterative methods deal greatly with baseline problem and are the subject of current article.

With the continuous improvement and widening of chemometric methods, the baseline filtering becomes more and more current topic. It is not very critical in the case of quantitative estimation (classical peak area computation), but becomes a critical part of processing procedure, when the chromatogram is treated as a unique signal-like vector (fingerprint) without any peak identification and evaluation [12]. In such approach, unsupervised chemometric methods (for similarity and clustering investigation) or supervised ones (for classification, discrimination, or extraction of some complex sample property from the whole chromatogram) are applied and baseline drifts can represent a majority part of whole variance in the processed dataset. Wrong removing of them can result in unpredictable errors, for example, calibration or discrimination can be impossible or clustering and similarity patterns between samples are strongly affected by random error from different baselines [12].

This article introduces two new automatic baseline methods. The first is based on quantile polynomial regression, and the second is based on quantile B-spline smoothing. The main and undoubtful advantage of new ideas is fully automatic processing of the signals. No parameters are set a priori, and ready-to-use routines are available under R. The proposed methods are compared with already proposed iterative approaches with optional newly proposed criteria for automaticity.

Theory

The Current Iterative Algorithms

The general iterative algorithm for baseline estimation is depicted on Fig. 1 and can be described as follows:

1. Fit a smooth curve to whole signal X without any weighting (all weights $W = 1$), obtaining fitted values f (Fig. 1a);
2. Modify the signal X (thresholding) or weights W (reweighting), according to obtained f;
3. Fit a smooth curve to modified signal X (or the same signal with modified weights W), obtaining next fitted values f (Fig. 1c);
4. If the fitted values f are good baseline estimate (convergence was achieved), stop (Fig. 1d). Otherwise go to step 2.

Before starting, several things must be set a priori. First, a kind of smooth curve fitted to the data must be chosen. This can be polynomial of selected order [11], locally fitted regression with selected span parameter (like LOESS [8]) or Whittaker smoother (penalized regression) [12, 14] with selected λ parameter. The curve can be also estimated using splines [15] with chosen degrees of freedom, but such spline application is not investigated in literature till now.

Then, there are two general ways of iterative modification. If the thresholding approach is used like in approach of Gan et al. [11], all parts of the signal above the fitted line are cut to this line (Fig. 1c)

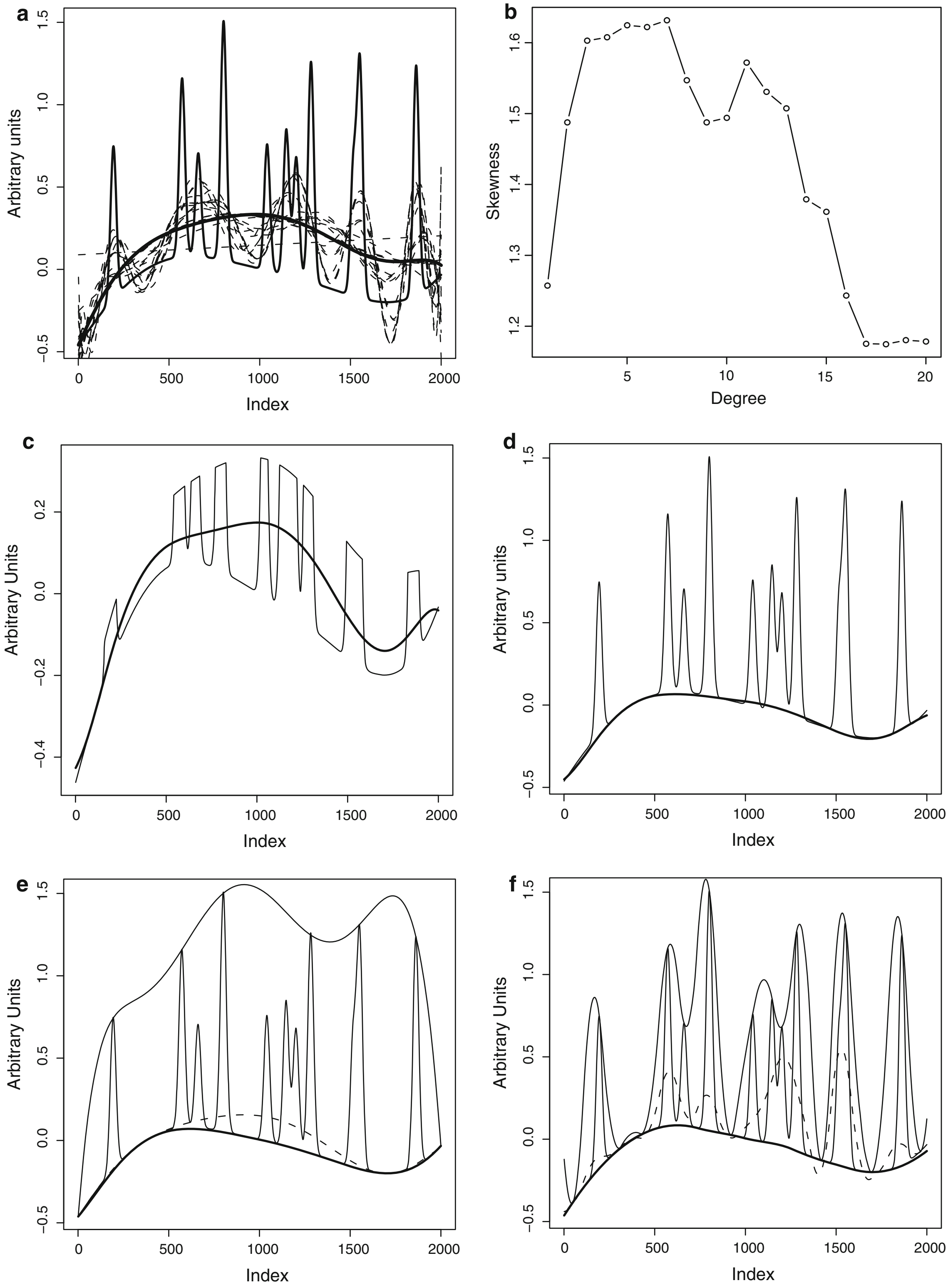


Fig. 1 The ideas of baseline estimation presented in the article: **a** curves of increasing flexibility (polynomial degree) fitted to the signal, optimal is depicted by *thicker line*, **b** plot of skewness versus flexibility—maximum indicates optimal one denoted on **a**, **c** second iteration when thresholding is used, **d** finally estimated baseline, **e** baseline estimated by quantile regression on optimal polynomial, **f** baseline estimated by splines. On **e** and **f**, regression against median (dashed line) and 0.99 quantile (solid outer thick line) is also depicted for illustrative purposes

$$x_i = \begin{cases} x_i & x_i \leq f_i \\ f_i & \text{otherwise.} \end{cases} \quad (1)$$

Each next fit results in a line lying in lower and lower part of the signal. After several iterations, only baseline remains.

The second way is to use weighted fitting (all curves mentioned can be used with weights). Daszykowski and Walczak [12] recommended setting all the weights above fitted line to small value (they will have almost no influence to the next fit); other points have weights set close to 1. The reweighting procedure has theoretically the same effect—subsequent iterations result in lower position of fitted line, until convergence is achieved.

No general remarks about required iteration number, nor strict convergence criteria were given in literature. If this algorithm is repeated to the convergence, the convergence criterion must be defined. On the other way, if predefined constant number of iterations is used, there is no recommendation how high this number should be.

The most important problem is the choice of "flexibility" of fitted curve (several flexibilities are shown in Fig. 1a). The degree of fitted polynomial, number of spline knots, span parameter of LOWESS, or λ parameter of Whittaker smoother should be set a priori, and wrong selection can underestimate baseline or harvest the peaks. Manual experiments with visual inspection of processed signals are most difficult part of baseline extraction in this way.

In this article, the following automatic method is proposed:

1. The curve of increasing flexibility is fitted to the signal, inspecting the skewness of the resulted residuals. The largest skewness (Fig. 1b) indicates optimum flexibility, modeling the baseline correctly, but not harvesting the peaks.
2. The procedure should stop if MSD (mean squared difference) between iterations is lower than 0.0001 of the mean of squared signal values. If the convergence is not achieved, algorithm should stop after 20 iterations.

Quantile Regression Methods

Another approach to baseline, never proposed in the literature before, is to use quantile regression methods. Whereas the classical methods of least squares approximate the conditional mean of the signal (and the fitted signal goes through the chromatogram, as on Fig. 1a), quantile regression estimates the median or other quantiles. Performing quantile regression against a small quantile (for example, 0.01, thick line on Fig. 1e and f) causes fitting a curve to the baseline, and the process does not need to be repeated iteratively. For comparative purposes, a quantile regression against median (dashed line) and 0.99 quantile (solid thick line) is also shown in Fig. 1e and f. The optimal complexity of the curve can be easily and successfully optimized against the coefficient known in classical regression as Akaike's information criterion (AIC). Details of the polynomial quantile regression are given comprehensively in the book of Koenker [16].

A spline quantile approach also exists [17], which was even further constrained [18] to preserve monotonicity, convexity, concavity, or other features. In this article, fitted baseline is unconstrained. Searching the literature did not bring any quantile version of LOESS, an unpublished discussion in internet about moving (running) quantile following classical LOESS was only found. However, the preliminary experiments shown that this approach would harvest the peaks too much in our case.

The following procedures (using R packages quantreg and cobs) were tested:

1. A standard polynomial regression for quantile 0.01: rq(y $\sim$ poly(x, degree), $\tau = 0.01$) with monitoring the AIC value. The degree with lowest AIC was fitted finally and treated as baseline
2. A spline quantile regression for quantile 0.01 with six knots of optimized (against AIC) position: cobs(x, y, $\tau = 0.01$)
3. A penalized version of previous approach, with 20 knots and penalty selected by SIC: cobs(x, y, $\tau = 0.01$, $\lambda = -1$)

Experimental

Synthetic Data

10000 very complex simulated signals of length 2000 were generated (random examples are shown in Fig. 2). Each signal consisted of:

1. 1–30 Gaussian peaks of width 1/15–1/30 of whole signal, with varying height and a slight probability to be overlapped
2. known significant baseline consisting of several sines (of varying frequency and phase) and several polynomial components of degree up to 4.

The procedure used to generate the signals is available upon e-mail request from the corresponding author.

Comparison of the Iterative Algorithms Without Any Criteria

The generated signals were first subjected to baseline extraction with all possible combinations of parameters:

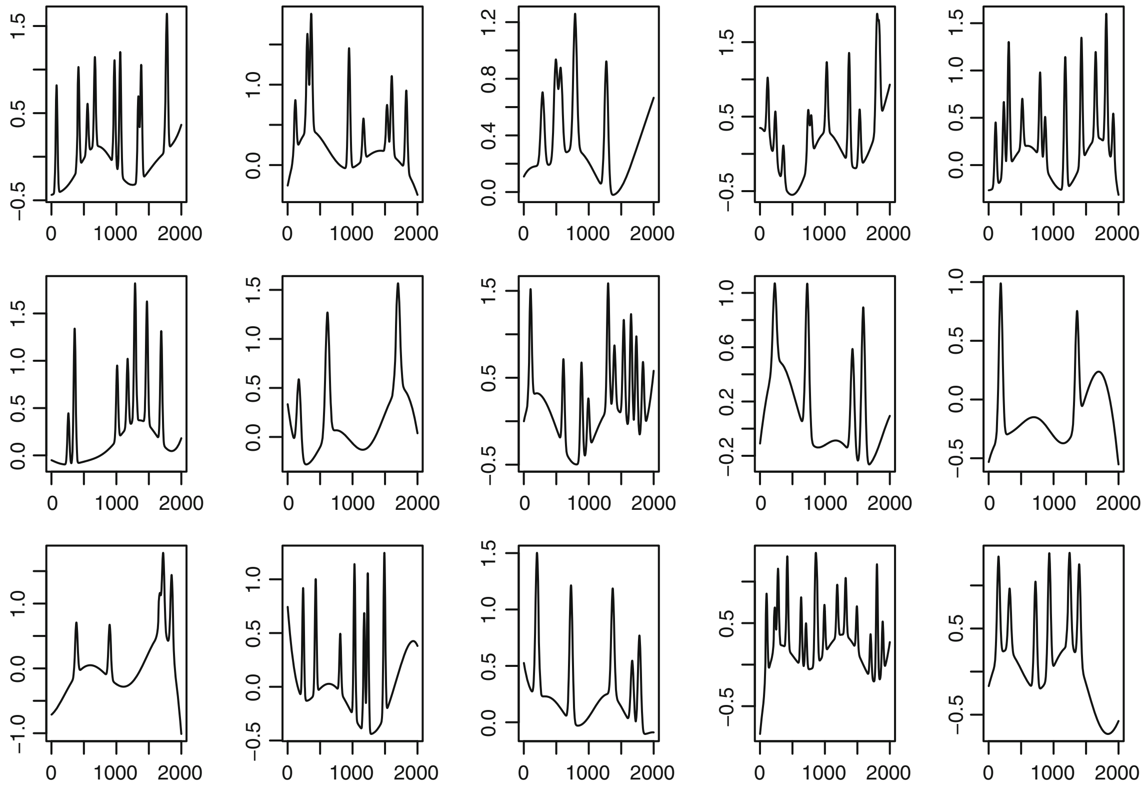


Fig. 2 Examples of artificial signals used in the study

1. Curve: polynomial (degree 1–20), spline (degrees of freedom 1–40), LOESS (span 0–1), Whittaker ($20 < \lambda < 1000$)
2. Iteration mechanism: thresholding, hard reweighting
3. 1–20 forced iterations of each approach

Comparison of the Iterative Algorithms with the Proposed Flexibility and Convergence Criteria

The generated signals were next subjected to baseline extraction with automatically selected complexity parameter (proposed in this paper), against maximum skewness of the residuals (Fig. 1b). The procedure was stopped if MSE between iterations was lower than 0.0001 of squared signal mean.

Baseline Estimation with a Newly Proposed Approaches

The generated signals were then subjected to newly proposed algorithms:

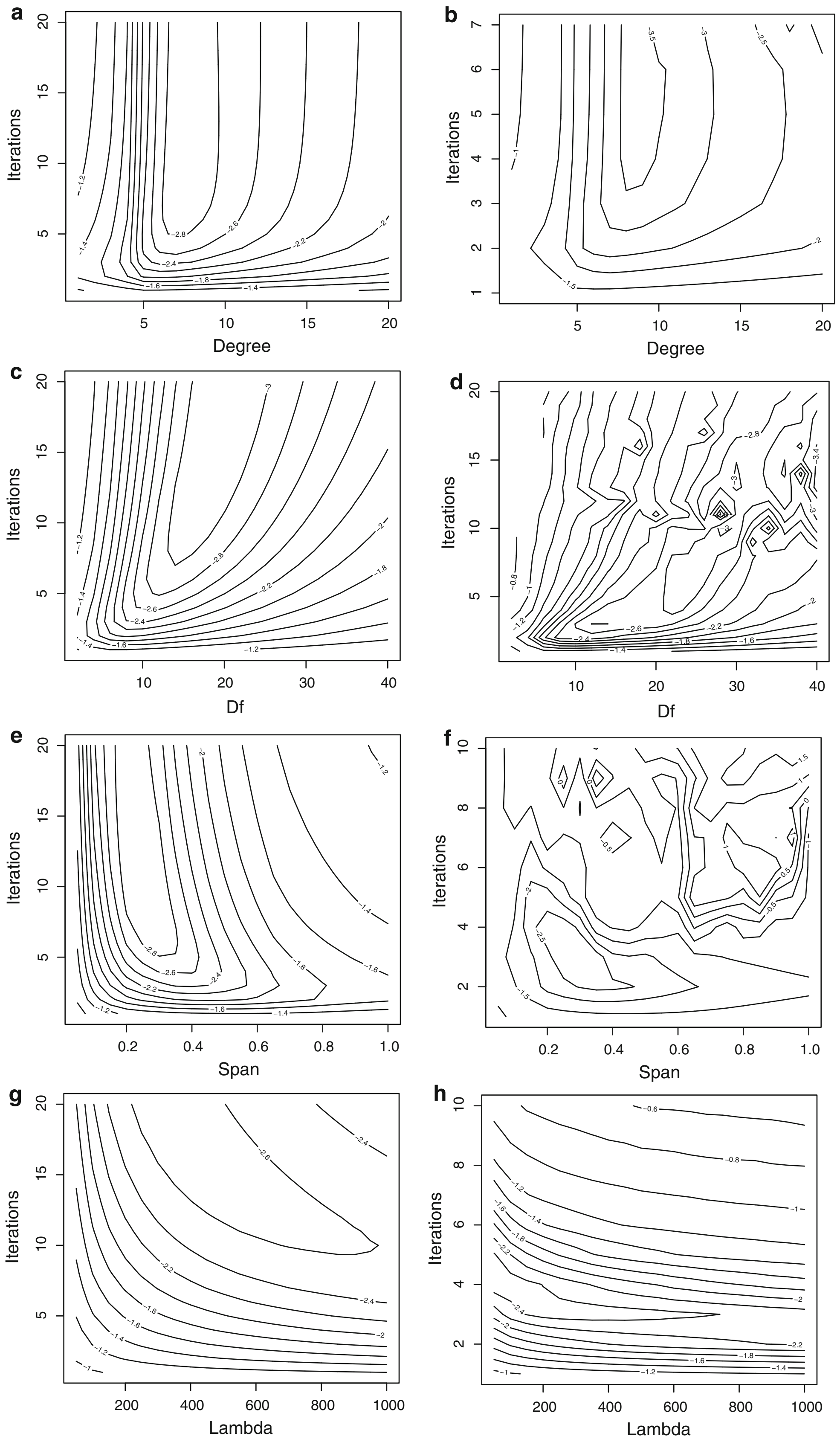
1. Quantile polynomial regression with $\tau = 0.05$, with optimal degree chosen with AIC criterion.
2. Quantile B-spline regression with maximally 20 knots (optimized by cross validation) with λ parameter found with SIC criterion.

All computations were done under GNU R 2.10.1 on computational cluster. The comparative thresholded and reweighted algorithms were implemented using built-in fitting functions: lm, lowess, smooth.spline. The Whittaker smoother was ported from Matlab code given by Eilers in his appendix [13]. For quantile regression, built-in functions of packages cobs and quantreg were used.

Real Chromatographic Data

The densitometric dataset consisted of thin layer densitograms of essential oils, each chromatographed five times. The essential oils (Sabana Oil, Warsaw, Poland and Avicenna Oil, Wrocław, Poland) were purchased in a local drugstore. They were diluted in methanol (analytical grade, POCH, Gliwice, Poland) (1:100) and applied to the Silica Gel GF254 plates (Merck, Darmstadt, Germany) in 20 µL amounts as 8 mm bands by means of Desaga AS-30 applicator. The plates were developed vertically under saturated conditions in Desaga tanks to the 15 cm distance. The temperature was conditioned at 24 °C, and mobile

Fig. 3 The decimal logarithm of MSE (mean square error) between estimated and real baseline for artificial signals without automatic criteria, computed by (**a**–**b**) polynomials, (**c**–**d**) splines, (**e**–**f**) LOESS, and (**g**–**h**) Whittaker. Estimation was done with thresholding (**a**, **c**, **e**, **g**) and reweighting (**b**, **d**, **f**, **h**)



phase was toluene–ethyl acetate (93:7), according to the European Pharmacopoeia 6.0. After the development, the plates were dried in the ambient temperature, sprayed with anisaldehyde solution R according to the same pharmacopoeia, and heated in 105 °C within 10 min.

After cooling, the plates were immediately scanned by Desaga CD-60 densitometer at 550 nm with 2 mm slit width and 0.4 mm slit height. This resulted in 85 signals (17 oils, 5 densitograms of each), each signal of length 1850 points.

Results and Discussion

The first, preliminary experiments with artificial signals brought the very important conclusion. If the signal has several narrow peaks only, separated by wide baseline segments, almost all algorithms perform very well. The only requirement is to use the curve with sufficient flexibility. There is almost no risk that the baseline will harvest the peaks, because such situation would occur only in the case of enormous increasing of the curve flexibility.

On the contrary, the signals with consecutive, often overlapping peaks, with the baseline drift almost hidden in relatively short segments between them, are very difficult to process. The curve used and the whole method (including number of iterations) are then very important. In general, it is easier to estimate too low flexibility, than too high. The polynomials with degree lower than 5 almost always underestimate the baseline.

The investigated signals were designed to be rather difficult to process. They have varying number of peaks, some part of peaks is overlapped and the baseline is modeled as the mixture of sine and polynomial components. It must be also noted that no artificial noise was used in the study, although its addition was used by Gan et al. [11]. The presence of the noise is insignificant in the case of thresholding methods. However, the reweighting methods applied to noisy signal result first in underestimation of the baseline (the estimated baseline is located below the noise), and in the next iterations the baseline "falls off" from the signal. Therefore, the best practice is to perform the denoising (smoothing) before the baseline extraction (even use oversmoothed signal version only for baseline approximation).

As the pure baseline is known in the case of artificial signals, the performance of the method could be measured. It was done by calculating the MSE (mean square error) between estimated baseline and the real one. Due to very high range of obtained MSE, the results are presented in the decimal logarithmic scale. Based on the signal scale, the value lower than −2.5 can be treated as sufficient baseline approximation.

The first part of the study was to investigate the behavior of the algorithms when curve flexibility is set a priori and iterations are forced without any convergence criterion, as no such study was carried out in the literature till now. Figure 3 shows the contour plot (map) of obtained MSE values against two parameters: curve flexibility (degree of polynomial etc., depending of method used) and iteration number. The following conclusions can be obtained in the case of analyzed dataset:

1. The polynomials (Fig. 3a and b) worked well in range of degree from 7 to 15. In the case of thresholding, six iterations are sufficient. Next iterations do not improve the solution, but there is no risk of divergence. In the case of reweighting, very low value −3.5 is reached with polynomial of 7–9th degree in 4–6 iterations. Above seven iterations, there is a small risk of divergence.
2. The splines (Fig. 3c and d) work well with 15–25 degrees of freedom. After seven iterations of thresholding, the good baseline approximation is achieved, and small improvement is obtained up to 20 iterations. In the case of reweighting, 3–5 iterations are sufficient, next the algorithm diverges and error value increases rapidly.
3. The LOWESS (Fig. 3e and f) fits baseline well in span range 0.2–0.4. Four iterations are sufficient in the case of thresholding. The reweighting method performs here poorly; only three iterations give acceptable results. Next iterations bring a divergence.
4. The Whittaker smoother (Fig. 3g and h) performs well in thresholding approach in wide range of λ values. However, at least ten iterations are needed. Reweighting method works well in slightly less wide range (200–800), but only 3–4 iterations are good value, another iterations result in failure.

This implies that convergence criterion (if used) must be very carefully set, especially when reweighting is applied. Wrong selection of convergence criterion will result in falling of the estimated baseline (divergence) and introducing very large error and unacceptable results. Based on above experiments with artificial and real signals, the proposal is to stop iterations when mean square difference between two consecutive baseline estimators is lower than 0.0001 of the signal power value (mean squared value of all the signal samples). It must be underlined here that too restricted convergence criterion can be never reached and algorithm can lead to falling down of the baseline from the signal. The condition "if MSE between iterations is lower than 0.0001 of squared signal mean" seems to be the best compromise between quality of the baseline and low risk of omitting convergence and falling into the divergence.

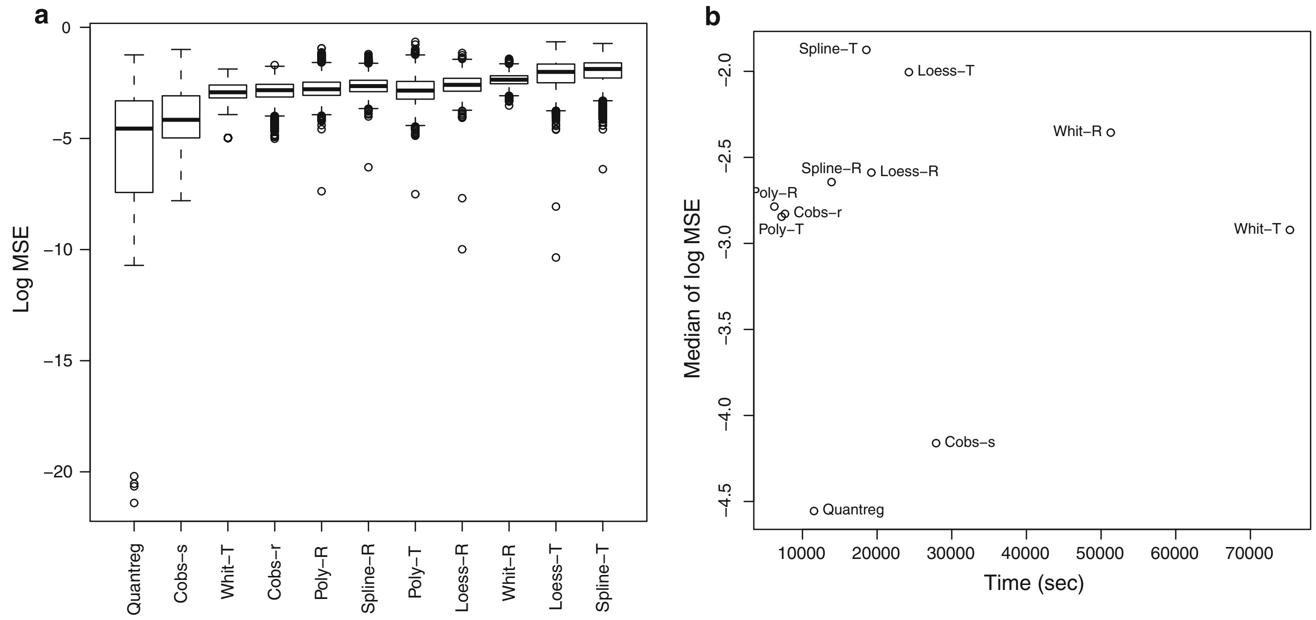


Fig. 4 Boxplot of decimal logarithm of MSE between real and estimated baseline for all the investigated methods with automatic criteria (**a**) and correlation of efficiency with computational time for all signals (**b**). *R* means reweighting, *T* means thresholding

The literature does not contain any recommendations of curve complexity and so far the analyst must experimentally set it against visual inspection of the results. The method proposed in this article, based on the skewness of the residuals (Fig. 1b), seems to be a good alternative to manual choice. Applying automatically chosen flexibility (against skewness) and iteration number (against convergence criteria) gives possibility to apply these algorithms automatically without manual setting of these two parameters. It is very important, as (as seen from Fig. 3) results can be wrong if wrong flexibility is chosen or wrong number of iterations applied. In the next investigation, four methods, both with thresholding and reweighting (total eight approaches) were next tested on the same artificial signals with automatic choice of flexibility and automatic stop after proposed convergence. For each signal, optimal curve flexibility was independently chosen against the skewness, and iterations were stopped after MSD lower than 0.001 of signal power. The results are compared with newly proposed methods based on quantile regression and quantile smoothing.

The comparison of mean squared error (MSE) between estimated and real (known in the case of artificial signals) baseline is depicted as the boxplot in Fig. 4a. The computational time is correlated with obtained error in Fig. 4b. The following conclusions can be made:

1. Reweighted algorithms (denoted with "$-r$") are computationally more efficient than the same approaches based on thresholding (denoted with "$-t$"), because of less iteration number is needed to convergence.
2. The quantile regression and quantile smoothing approach with automatically chosen lambda (cobs-s) are visibly better than the other algorithms, performing similarly. Although the spread of MSE values are somewhat higher, no worse values than other algorithms are met over all artificial signal dataset.
3. The quantile regression is the winner of compromise between efficiency and the computational time. The cobs-s algorithm is also efficient, but needs three times more of the computational time.
4. The algorithms based on polynomials, both reweighted and thresholded, are most efficient among the other algorithms.
5. The Whittaker smoother is less computationally effective due to creation of large matrices at each run. Use of several sparse matrices packages in R did not prove any significant advantage (both in our implementations and airPLS package [14]).

The study of quantile regression baseline estimation on experimental dataset was compared with Whittaker smoother approach implemented in airPLS package, after proper denoising of densitometric real signals [19]. The results are depicted in Figs. 5 and 6. The similarity between signals was depicted (Fig. 6) by Principal Component Analysis. The unprocessed signals (Fig. 6A) show very disturbed clustering, and similarity between several densitograms of the same oil is strongly affected

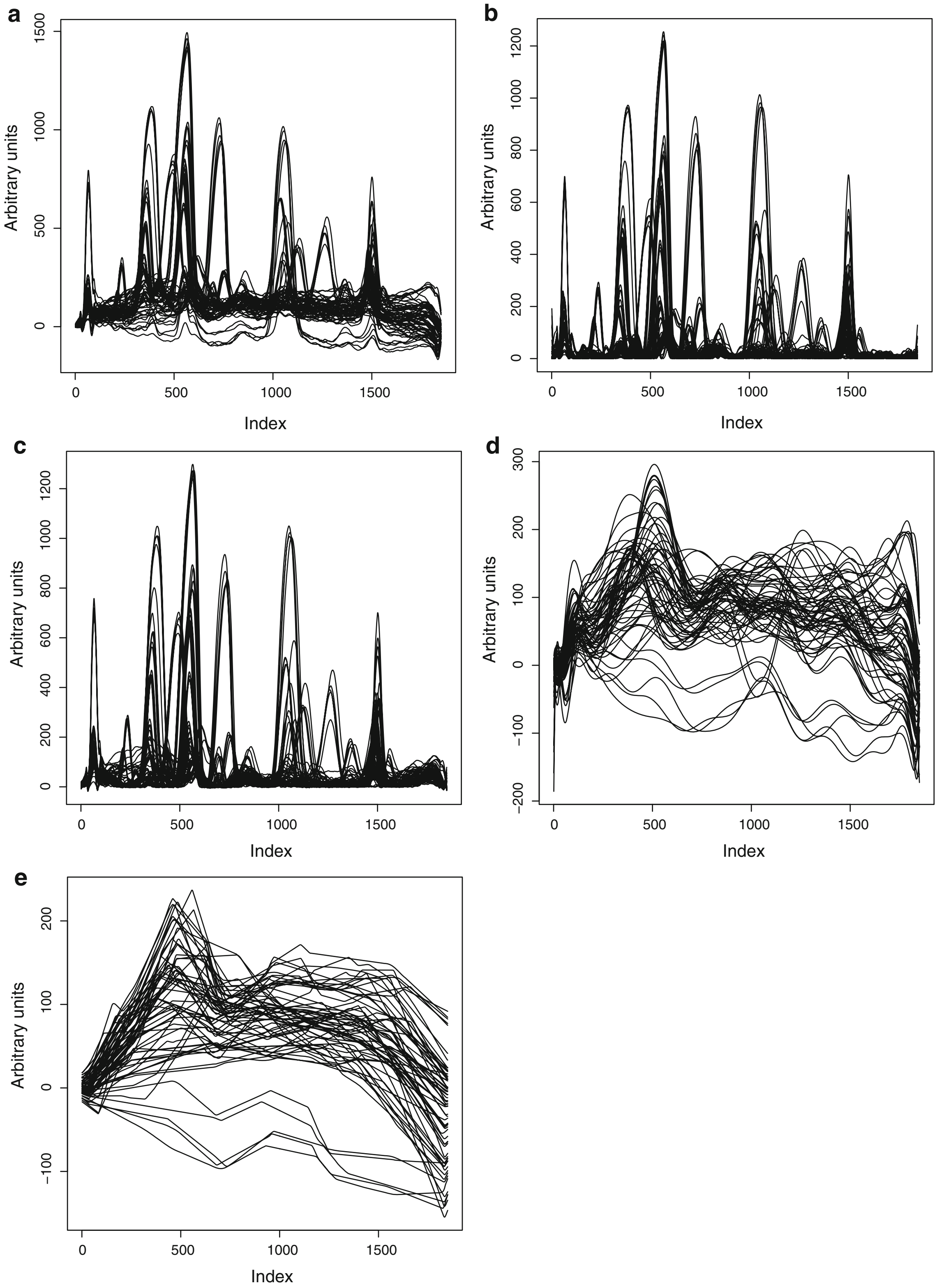


Fig. 5 Densitometric signal dataset: original (**a**), baseline filtered by quantile regression (**b**), baseline filtered by airPLS (**c**) and estimates of the baselines from quantile regression (**d**) and airPLS (**e**)

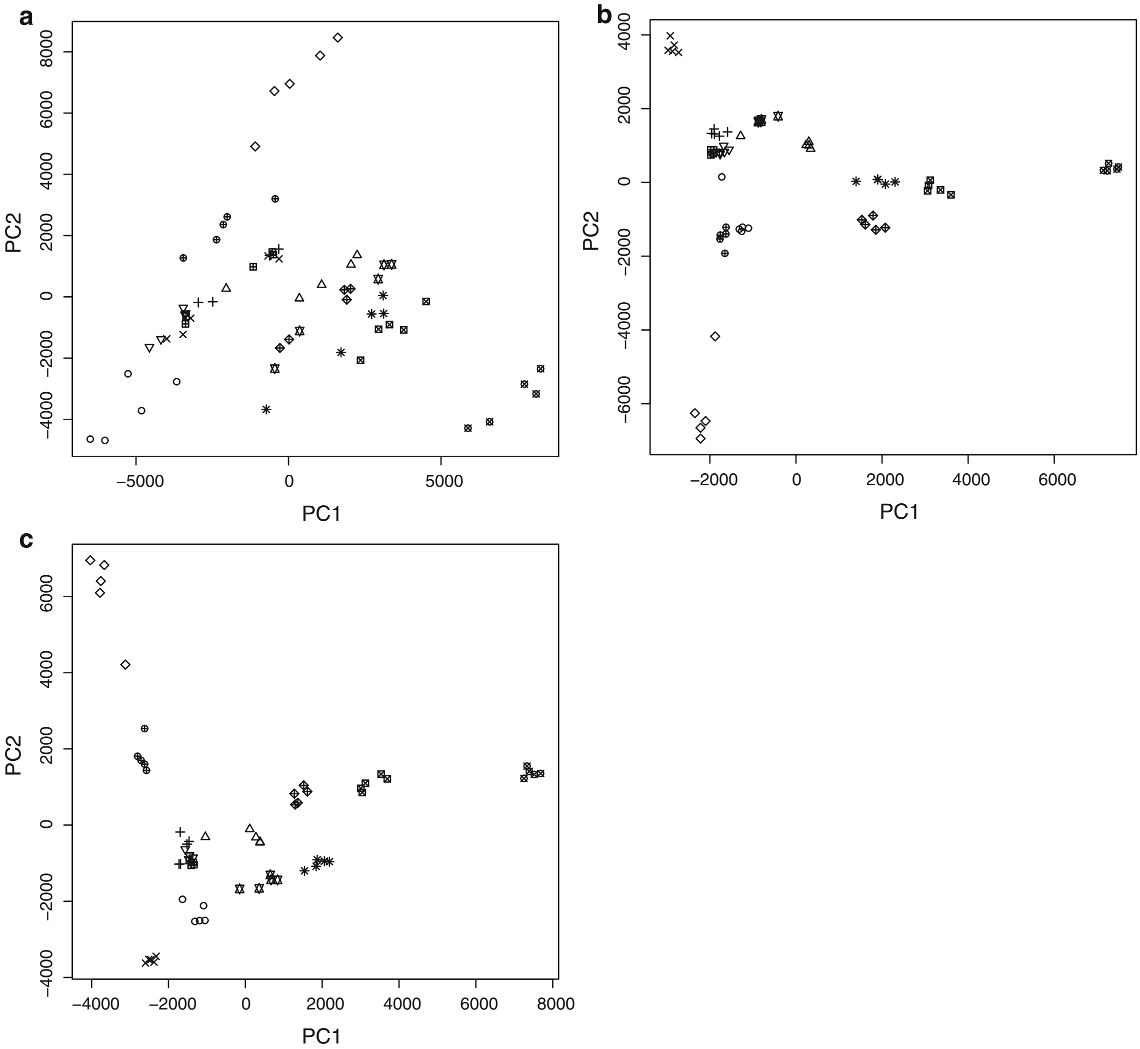


Fig. 6 PCA of the densitometric signals shown in Fig. 5: before baseline removal (**a**), after baseline removal with quantile regression (**b**), after baseline removal with airPLS (**c**). The *symbols* denote the species of plants from the oils were taken

by random variation of baseline. On the contrary, after baseline filtering (Fig. 6b and c) the densitograms of the same oil are very similar and whole dataset clusters visibly against the type of the essential oil. Therefore, it can be concluded that both algorithms perform well and the signal variability created by baseline drifts was almost removed. However, airPLS algorithm fits non-smooth baseline, and differences between samples of the same origin in PCA plot (Fig. 6b) are slightly larger. Moreover, several signals were not processed accurately with airPLS, and positive baseline is still visible. Quantile regression dealt with these issues. Comparing the computational time for this dataset (40 s for quantile regression and 350 s for airPLS, where the difference enhances with longer signals) the new quantile regression approach can be recommended.

Conclusion

The quantile regression and spline quantile smoothing with automatically chosen penalty can be preferred baseline estimation methods due to visible better speed and similar performance with the other approaches. Quantile regression is undoubting winner both in performance and computational efficiency. The other algorithms behave in comparable way, and polynomial regression can be here preferred due to short computational time. The new method of selecting optimal complexity, based on the residual skewness, works well with all approaches. As the baseline methods are universal, the applicability of proposed approach is not limited to TLC with densitometric detection (as presented illustrative example), but to all chromatographic (and even non-chromatographic) signals.

Acknowledgments The author would like to gratefully acknowledge the computational support of Interdisciplinary Centre for Mathematical and Computational Modeling, Warsaw University, Poland, under the computational Grant No. G36-8.

Appendix—R Code

The quantile regression baseline estimation can be performed using following code (using rq from quantreg package):

```
baseline <- function (x)
{
    aics <- rep(NA, 20)
    g <- 1:length(x)
    for (k in 1:20) {
        fit <- rq(x ~ poly(g, k), tau = 0.01)
        aics[k] <- AIC(fit)
    }
    optk <- which.min(aics)
    fit <- rq(x ~ poly(g, optk), tau = 0.01)
    b <- fitted(fit)
    return(b)
}
```

References

1. Pearson GA (1977) J Magn Reson 27:265–272
2. Dietrich W, Rüdel CH, Neumann M (1991) J Magn Reson 91:1–11
3. Moore AW, Jorgenson JW (1993) Anal Chem 65:188–191
4. Friedrichs MS (1993) J Biomol NMR 5:147–153
5. Brown DE (1995) J Magn Reson 114:268–270
6. Andrew KN, Rutan SC, Worsfold PJ (1999) Anal Chim Acta 388:315–325
7. Golotvin S, Williams A (2000) J Magn Reson 146:122–125
8. Ruckstuhl AF, Jacobson MP, Field RW, Dodd A (2001) J Quant Spectrosc Radiat Transf 68:179–183
9. Ma XG, Zhang ZX (2003) Anal Chim Acta 485:233–239
10. Chau FT, Leung AK (2000) Application of wavelet transform in processing chromatographic data. In: Walczak B (ed) Wavelets in chemistry. Elsevier, Amsterdam, p 207
11. Gan F, Ruan G, Mo J (2006) Chemom Intell Lab Syst 82:59–65
12. Daszykowski M, Walczak B (2006) Trends Anal Chem 11:1081–1096
13. Eilers PHC (2003) Anal Chem 75:3631–3636
14. Zhang ZM, Chen S, Liang L (2010) Analyst 135:1138–1146
15. de Boor C (1978) A practical guide to splines. Springer, New York
16. Koenker RW (2005) Quantile regression. Cambridge University Press, Cambridge
17. Ng P (1996) Comput Stat Data Anal 22:99–118
18. Ng P, Maechler M (2007) Stat Model 7:315–328
19. Komsta Ł (2009) Anal Chim Acta 641:52–58