Baseline Correction with Asymmetric Least Squares Smoothing

Paul H. C. Eilers Hans F.M. Boelens

October 21, 2005

AUTHORS' ADDRESSES: Department of Medical Statistics, Leiden University Medical Centre, P.O. Box 9604, 2300 RC Leiden, The Netherlands
Biosystems Data Analysis Group, Swammerdam Institute for Life Sciences (SILS), University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands.
E-MAIL: p.eilers@lumc.nl, hans.boelens@science.uva.nl

Abstract

Most baseline problems in instrumental methods are characterized by a smooth baseline and a superimposed signal that carries the analytical information: a series of peaks that are either all positive or all negative. We combine a smoother with asymmetric weighting of deviations from the (smooth) trend get an effective baseline estimator. It is easy to use, fast and keeps the analytical peak signal intact. No prior information about peak shapes or baseline (polynomial) is needed by the method. The performance is illustrated by simulation and applications to real data.

KEYWORDS: baseline, chromatography, spectroscopy, penalty, asymmetric least squares.

1 Introduction

Unstable baselines occur in many types of instrumental measurements. They can cause severe problems, especially when detection limits are approached[1, 2]. For this reason baseline correction is being used routinely. In many instrumental software systems it is a semi-manual process: the analyst indicates (on a computer screen) begin and end of "peak regions" and the software constructs a (piece-wise linear) connecting curve as an estimate of the baseline, which then is subtracted. This approach is subjective and time-consuming and of limited repeatibility[3].

There have been numerous proposals for automatic baseline correction. We limit ourselves to the case in which no models for signal (and background) are available. Assuming positive peaks, we can classify existing approaches as follows.

- Differencing and filtering. A drifting baseline generally shows a much smaller rate of change than the peaks of interest. Differencing of a signal amplifies the higher frequency components and hence suppresses baseline drift. Additional filtering may be necessary because high-frequency noise is amplified. Such an approach leads to bandpass filters that suppress both baseline drift and measurement noise. Peak shape models may also be used [4, 5].

- The maximum entropy approach [6] also tries to capitalize on the differences in frequency content by separating the signal into a slow (the baseline) and a fast part (the analyte signal).

- Multivariate, or polynomial, baseline modelling, taking advantage of areas where no signal, but only background, is present [7].

- Simulation of a running ball. An imaginary ball with a large radius is imagined running over (under) the signal. The trace of its lowest (highest)point is the estimated baseline [8]

- Borrowing ideas from robust statistics. Iterative weighting reduces the influence of the higher parts of peaks to essentially zero [9].

- Using an asymmetric goal function. The usual sum of squares of differences between data and a fitted curve is hard limited on the main part of the positive axis [10].

A more detailed comparison of some of these methods can be found in [11].

The purpose of this paper is to present an alternative and quite general approach for baseline estimation, which was only hinted at in connection to parametric time warping[12]. We use a (Whittaker) smoother to get a slowly varying estimate of the baseline[13]. In contrast to ordinary least squares smoothing, however, positive deviations with respect to baseline estimate are weighted (much) less than negative ones. The latter approach is known as asymmetric least squares (AsLS). There exist relatively few publications on the use of AsLS in regression modelling[14, 15]. The AsLS approach has also been used by us for smoothing [16], for background correction [17] in hyphenated chromatography and for finding new features in large spectral data sets [18].

Asymmetric least squares smoothing is attractive for baseline estimation: 1) it is fast, even for large signals; 2) the flexibility of the baseline can be tuned easily with one parameter; 3) the position of the baseline can be tuned with one other parameter. Given the two parameters, the computations are completely reproducible. Unfortunately we are not able to give and all-purpose recipe for automatic choice of the parameters for arbitrary signals, so human judgement will generally be needed. With a simulation we document performance in a situation where baseline and peaks are known.

2 Smoothing and asymmetric least squares

For completeness, we give a summary of the Whittaker smoother[13]. Let y be a signal, of length m, assumed to be sampled at equal intervals. Let z be another series, which should have the following properties: 1) being smooth, but 2) also being faithful to y. These two goals can be combined by minimizing the penalized least squares function

$$S = \sum_i (y_i - z_i)^2 + \lambda \sum_i (\Delta^2 z_i)^2, \tag{1}$$

where $\Delta^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) = z_i - 2z_{i-1} + z_{i-2}$. The first term in S measures the fit to the data, while the second term is a penalty on non-smooth behavior of z. The parameter λ tunes the balance between the two terms. A generalization introduces a vector w of weights and minimizes

$$S = \sum_i w_i (y_i - z_i)^2 + \lambda \sum_i (\Delta^2 z_i)^2. \tag{2}$$

The minimization problems leads to the following system of equations:

$$(W + \lambda D'D)z = Wy, \tag{3}$$

where $W = \mathrm{diag}(w)$ and D is a difference matrix: $Dz = \Delta^2 z$. This generally is a large system, as the number of equations is equal to m, the length of y. But it also is a very sparse system: only the main diagonal and two sub-diagonals above and below it are non-zero. Efficient storage and computation (both linear in m) are possible in Matlab or classic computer languages.

In a common applications relatively light smoothing will remove noise, while strong smoothing gives the slowly varying trend of a signal. Figure 1 shows (dotted line), for a simulated (chromatographic) signal (three Gaussian peaks on a sinusoidal baseline), the trend that would be obtained by strong smoothing. It is clear that this trend does not estimate the baseline. The signal deviates in both directions from the trend. For a good baseline, we should only see deviations in the positive direction; the analytical peaks of the signal should not be distorted.

When using this (or any other) smoother, the signs of the residuals $y - z$ don't matter: a positive residual gets the same weight as a negative one. We get interesting and useful results if we change this, and give much more weight to the negative residuals. We introduce a parameter p and compute weights as follows: $w_i = p$ if $y_i > z_i$ and $w_i = 1 - p$ otherwise. The goal is to find a solution, say $z(p)$, of (3) that conforms to this choice of weights. This is the principle of asymmetric least squares[16, 15], applied to smoothing.

The equations for the solution look intricate, because of the mutual interaction of weights and smooth curve, but it can be transformed into iterative application of two

easy computations. Let an approximate solution, say $\tilde{z}$ be given. A natural choice is the result of smoothing with uniform weights. Given $\tilde{z}$, it is trivial to compute new weights, say $\tilde{w}$. With these weights we solve (3) to get a new estimate of z. We repeat these steps until the weights don't change anymore. One can prove that the goal is convex and that the iterations always go downhill along the gradient, so that convergence will be obtained. In practice 5 to 10 iterations are sufficient. See the supporting material for mathematical details.

Implementation in Matlab is simple, as the following code fragment shows. To emphasize the basic simplicity of the algorithm, the number of iterations has been fixed to 10. In practical applications one should check whether the weights show any change; if not, convergence has been attained. Computation time is linear in length of y and is less than a second for a series of 1000 samples (on a 1000 MHz PIII computer).

```
function z = baseline(y, lambda, p)
% Estimate baseline with asymmetric least squares
m = length(y);
D = diff(speye(m), 2);
w = ones(m, 1);
for it = 1:10
   W = spdiags(w, 0, m, m);
   C = chol(W + lambda * D' * D);
   z = C \ (C' \ (w .* y));
   w = p * (y > z) + (1 - p) * (y < z);
end
```

There are two parameters: p for asymmetry and λ for smoothness. Both have to be tuned to the data at hand. We found that generally $0.001 \leq p \leq 0.1$ is a good choice (for a signal with positive peaks) and $10^2 \leq \lambda \leq 10^9$, but exceptions may occur. In any case one should vary λ on a grid that is approximately linear for $\log \lambda$. Often visual inspection is sufficient to get good parameter values. The speed of the algorithm allows interactive use: on a 1000 MHz PIII PC, using Matlab 6.5, it takes less than half a

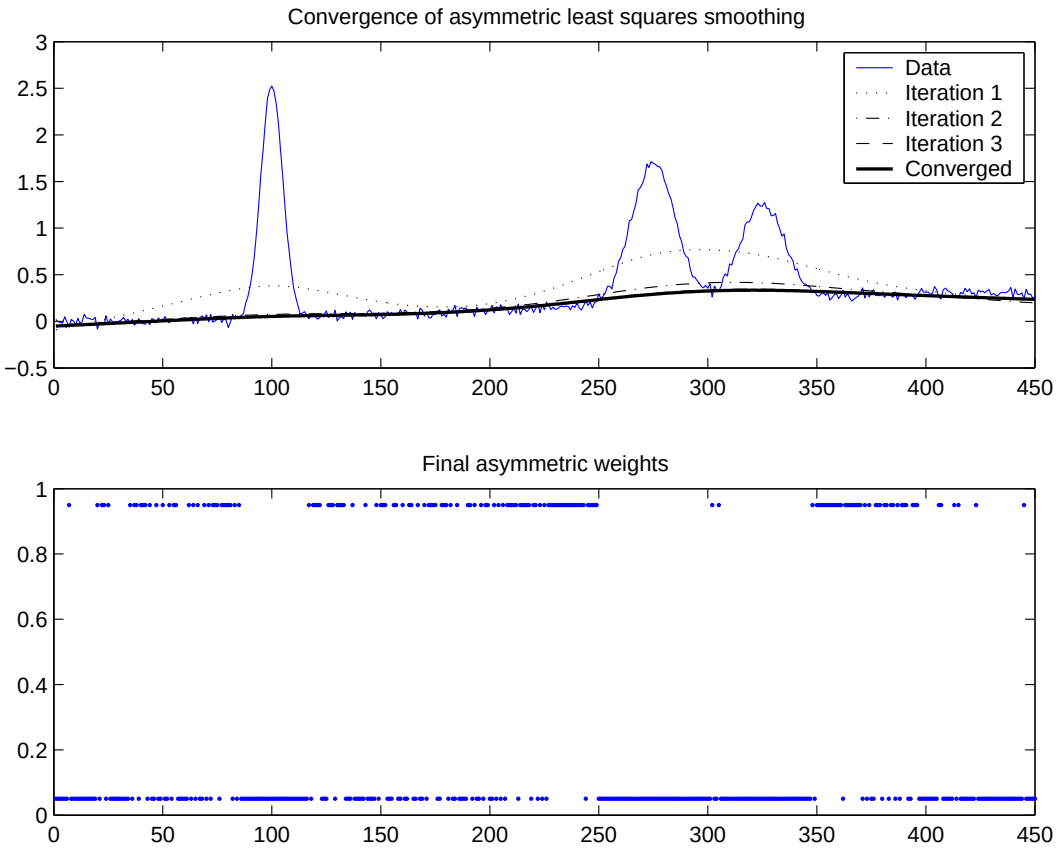

Figure 1: Simulated (chromatographic) data and iteratively estimated baseline by AsLS smoothing algorithm. The upper panel shows the first three iterations and the final result ($p = 0.05$ and $\lambda = 10^6$. The lower panel shows the final data weights.

second to estimate the baseline for a signal of length 1000.

In Figure 1 three iterations and the final result of the AsLS algorithm on the simulated chromatographic data are shown. Convergence occurred after 5 iterations. The initial estimate, which was the result of symmetric smoothing, is strongly improved with each new iteration. The figure also shows, in the lower panel, the final data weights. Note, that they are all small (0.05) at the peak positions.

To get an better impression of the potential performance of the algorithm, we computed the optimal values for parameters (p and λ), using the true baseline, say t, as a reference, i.e, we varied λ and p on a fine grid and minimize $Q = \sum_i (z_i - t_i)^2$. This is

unrealistic insofar that in reality we don't know t. But it gives us a figure for the best case performance. If Q can be made small, it is worthwhile to search for algorithms that yield (near) optimal parameter values from the data. On the other hand, if Q cannot even be made small if we have the actual baseline as our guide, the whole approach might better be discarded.

Figure 2 shows the simulated signal (upper right panel) the logarithm grid on which the parameters are varied (upper left plot). For each pair of parameters (λ, p) a baseline was estimated and the error, $RMSE = \sqrt{\sum_i^m (\hat{z}_i - t_i)^2/m}$, was computed. Contour lines and profile plots (left column of Figure 2) show clearly that one minimum of the $RMSE$ exists. For the values of λ and p at lowest $RMSE$ the baseline estimate is shown (fat solid line) in the upper right panel. The correspondence between the true and the estimated baseline is very good as shown in the center right panel (graph of difference $\hat{z} - t$). Compare the size of this difference with the size of the original baseline signal. It is also interesting to note that the peak signals, that carry the analytical information are hardly deformed (lower right panel graphs $x - (y - \hat{z})$).

The simulation shows that, with proper parameter values, a result can be obtained that is quite near to the true baseline, supporting the validity of the approach. In practice we have only the measured data to guide us in choosing good values of λ and p. We were not able to find a fail-safe cross-validation algorithm yet and therefore we report here our experience with relatively ad-hoc computations. They can help to set approximately optimal parameter values. Human judgement, however, still seems preferable for fine-tuning.

When relatively large "pure baseline" regions are present, guidance for a good value of the asymmetry parameter p can be obtained from the procedure that we sketch here. If p is chosen well, a histogram of the residuals $y - z$ will have two components: a more or less normally distributed peak near zero, representing noise around the baseline, and an asymmetric component on the positive axis, representing the peaks in the signal. When the peak of the "normal" distribution is at zero, the baseline will "cut through the middle of the noise" and this is what we want. One varies p to get this results.

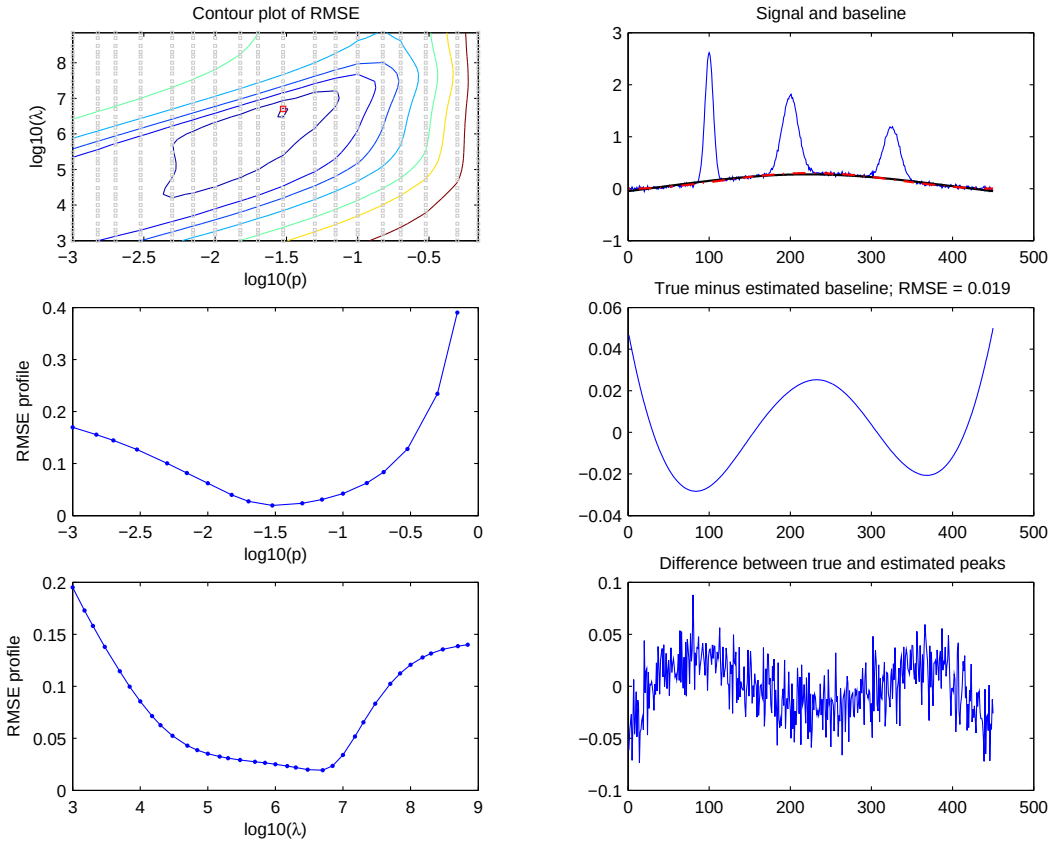

Figure 2: Best case performance of the algorithm when varying the parameters λ and p. Upper left panel: contour lines of RMSE; the light colored squares show the search grid that was used. Upper right: simulated data, true baseline (dashed line) and estimated baseline (solid line). Middle and lower left panels: profiles of RMSE at optimal λ and optimal p. Middle right panel: difference between estimated and true baseline. Lower right panel difference between true (chromatographic) peaks and the baseline corrected signal. (Matlab code to produce this plot is additional material)

.

Finding the optimal smoothing parameter λ is harder. We did some calculations to see if for a given p, a good value of λ might be found by asymmetric cross-validation. The idea is to leave out all the even observations (i.e those with i even) by giving them zero weights. The smoother will automatically compute interpolated values for them. These are compared to the real data by computing the asymmetrically weighted sums of squares of differences between interpolated and real values, say CV. One searches for the value of λ that minimizes CV. We remark that this only works well when the noise nearly white (uncorrelated). Otherwise too small values of λ will be found. We recommend to always check results by inspection.

This procedure seems to work reasonably well for a fixed value of p. If we try to include p in the cross-validation, we run into an unpleasant property of AsLS: the sum of asymmetrically weighted squared residuals depends strongly on p. A simple example illustrates this. Let $x_i = i/100$, $i = 1 \ldots 99$ be our "data", for which we compute the asymmetric mean $g_p = \sum_i w_i x_i / \sum_i w_i$, with asymmetric weights w. We also compute $WSS = \sum_i w_i (x_i - g_p)^2$ and $CSS = WSS / \sum_i w_i$. One sees that larger p gives larger WSS and CSS. This phenomenon wrecks cross-validation, as we experienced in our simulations. We could not find a theoretical principle or satisfactory empirical formula to calibrate WSS or CSS. As an ad-hoc solution we offer the following approach: compute WSS for the left-out data and divide it by as computed from WSS as computed from the left-in data. In general, we advice always to check by visual inspection.

Table 1: Asymmetric mean, sum of weighted squares (WSS) and corrected sum of weighted squares (CSS) for several values of asymmetry parameter p. The "data" are $i/100$, for $i = 1 \ldots 99$. The table shows that generally WSS and CSS increase with increasing p.

p	0.0010	0.0020	0.0050	0.1000	0.2000	0.5000
ALS mean	0.0354	0.0475	0.0707	0.2525	0.3350	0.5000
WSS	0.0304	0.0592	0.1410	1.8191	2.8746	4.0425
CSS	0.0098	0.0142	0.0190	0.0608	0.0726	0.0817

3 Applications

We now present a number of applications of the AsLS baseline correction procedure. The examples given are of different complexity and illustrate the performance and limitations of the algorithm.

Chromatographic signals generally are very well-behaved candidates for baseline estimation. Peaks are relatively sharp and often well isolated, with little noise. Figure 3 shows an example. The length of the signal is 3600 samples. The baseline is very smooth, leading to $\lambda = 10^5$, while noise is low, leading to a small value of p (0.01).

The mass spectrum in Figure 4 was obtained from human blood serum with a MALDI-TOF instrument, in the course of a proteomics project. Approximately 45000 data pairs (m/z, ion counts) were summarized in mass channels one Dalton wide. The influence of p is illustrated in figures 5 and 6. When the distribution of residuals is approximately centered at zero, the baseline runs in the middle of the noise band in areas without signal.

Figure 7 shows Raman spectra before and after baseline removal. The spectra exhibit a baseline disturbance (elevated baseline for low wavenumbers) that can be attributed to fluorescence of the reaction mixture. Also, a broad band-like feature can be seen centered around 1500 cm-1. This feature is caused by the glass wall of the reactor. Both these irrelevant spectral disturbances can be removed well by the ALS/baseline correction. Note that the settings of the baseline correction is the same for all three measured spectra.

Figure 8 shows data from an FTIR instrument [6]. The baseline here is quite variable, compared to the previous examples. A rather small value of λ (100) is needed to get a good fit (based on visual judgement).

When many peaks overlap and together form a "mountain range", baseline estimation is hard. A subjective choice of p and λ, guided by subject matter knowledge and experiences, will be the only reasonable choice then. An example is shown in Figure 9. Its hard to decide how much the baseline can be allowed to move upward under the two peak areas. In the lower spectrum the right part of the baseline is not well estimated.

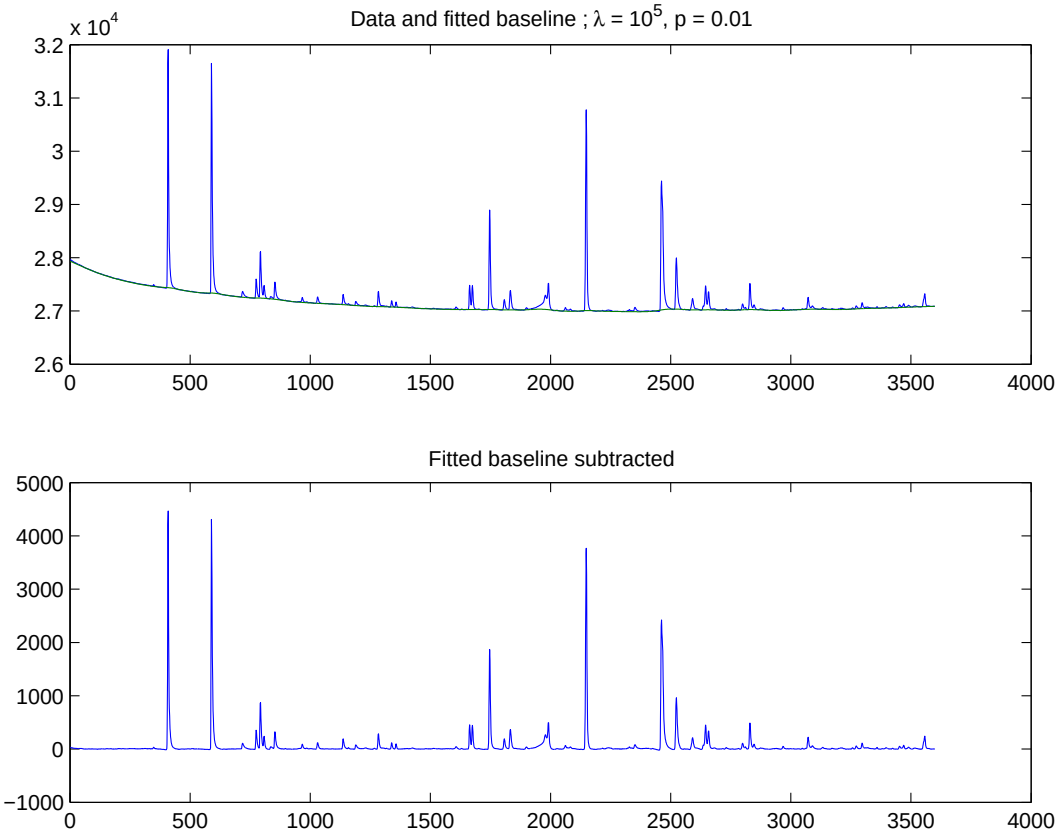

Figure 3: Baseline correction of a GC chromatogram. Top: GC chromatogram and estimated baseline; bottom: baseline corrected chromatogram.

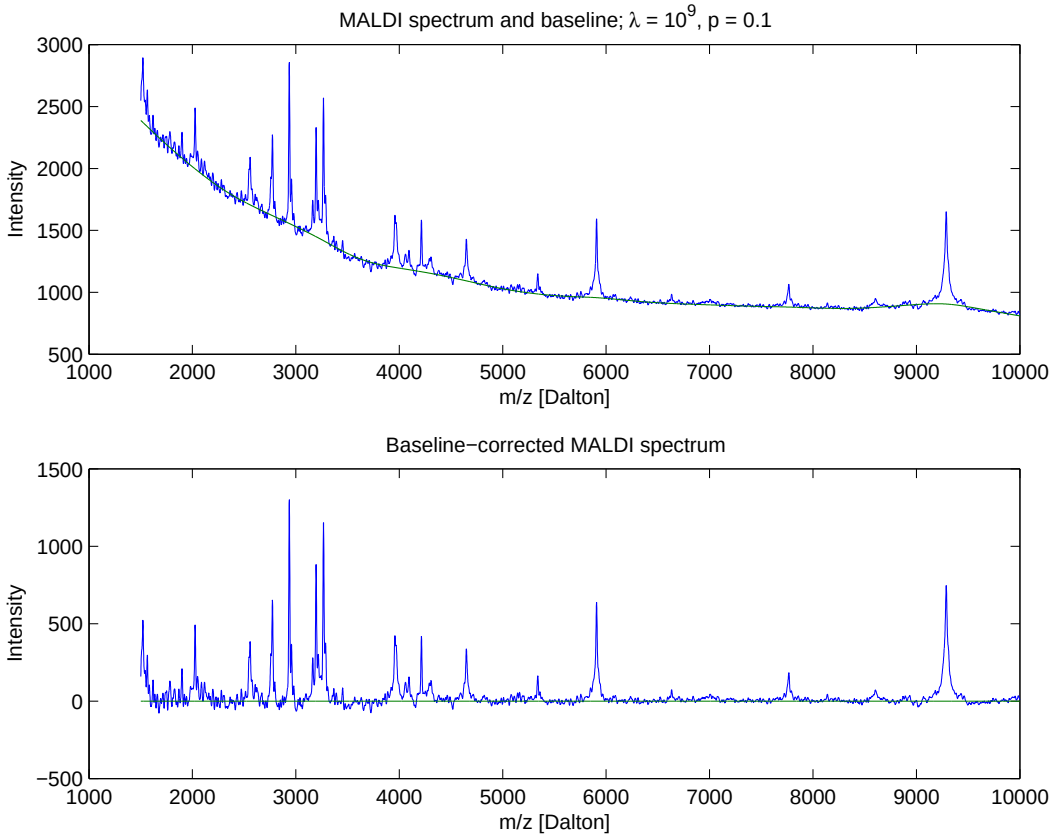

Figure 4: Baseline correction of a MALDI-TOF mass spectrum. Top: spectrum and estimated baseline; bottom: baseline corrected spectrum.

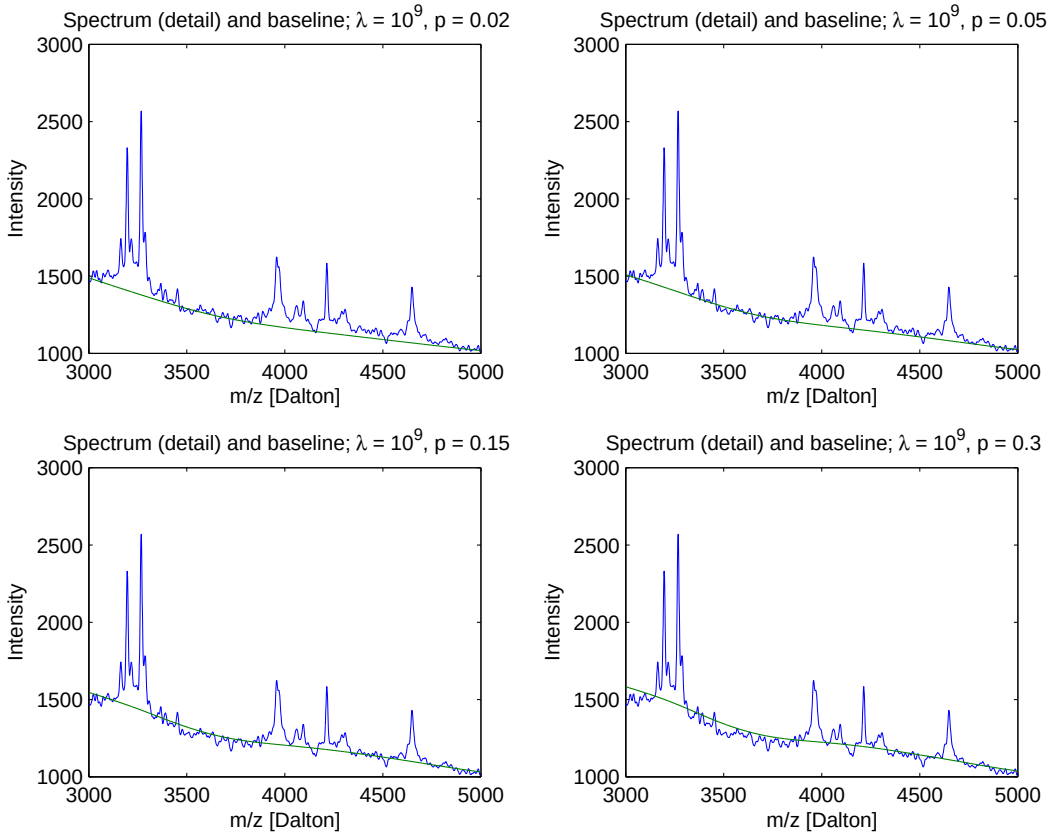

Figure 5: Detail of the MALDI-TOF mass spectrum with estimated baseline, illustrating the effect of the asymmetry parameter p.

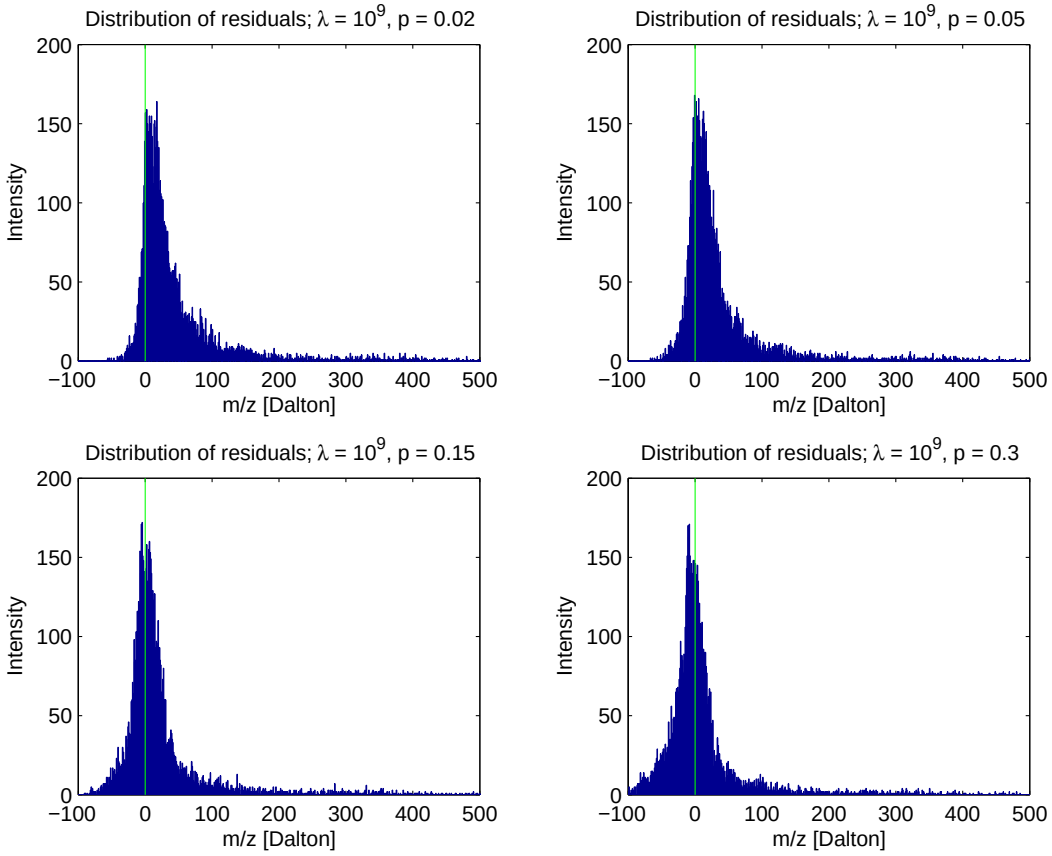

Figure 6: Illustration of how the asymmetry parameter p influences the distribution of the differences between spectrum and baseline. The peaked part of the distribution should be centered around zero.

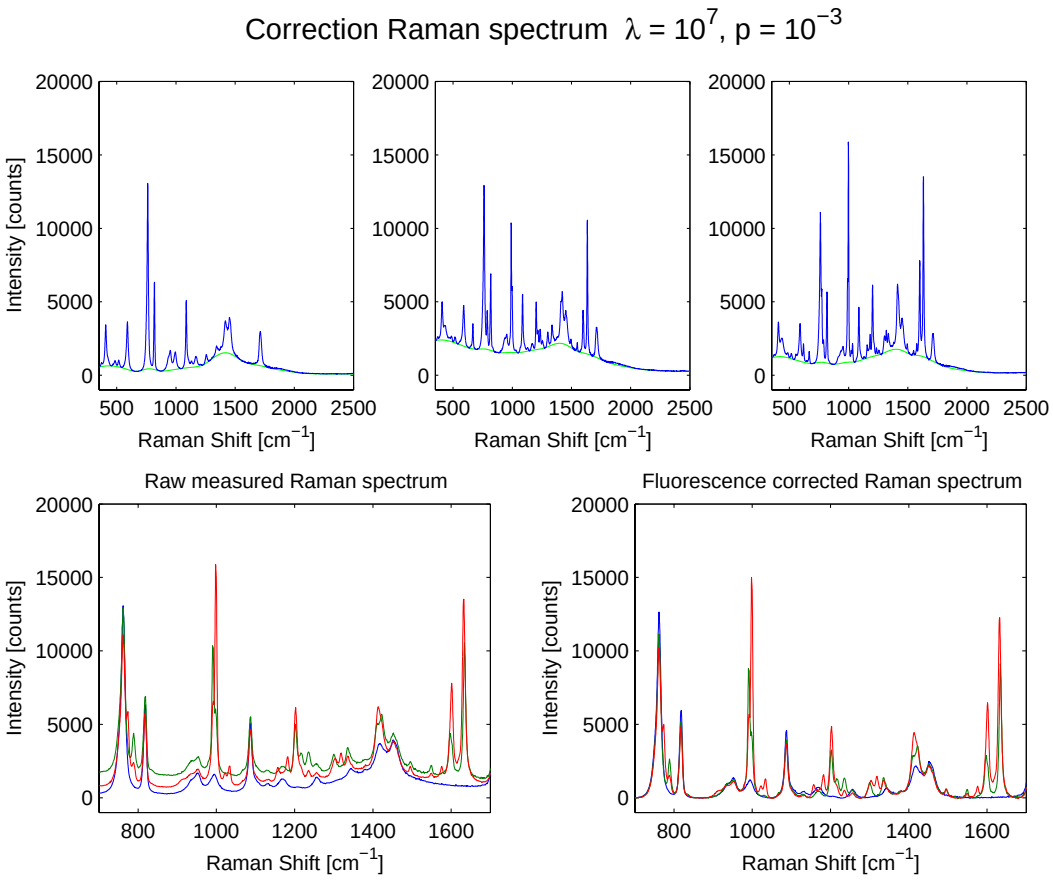

Figure 7: Baseline correction of Raman spectra that were measured during a solvent co-polymerisation reaction. Top plots: 3 measured spectra and estimated baseline; bottom plots: the three spectra in one graph, before (left panel) and after (right panel) baseline correction.

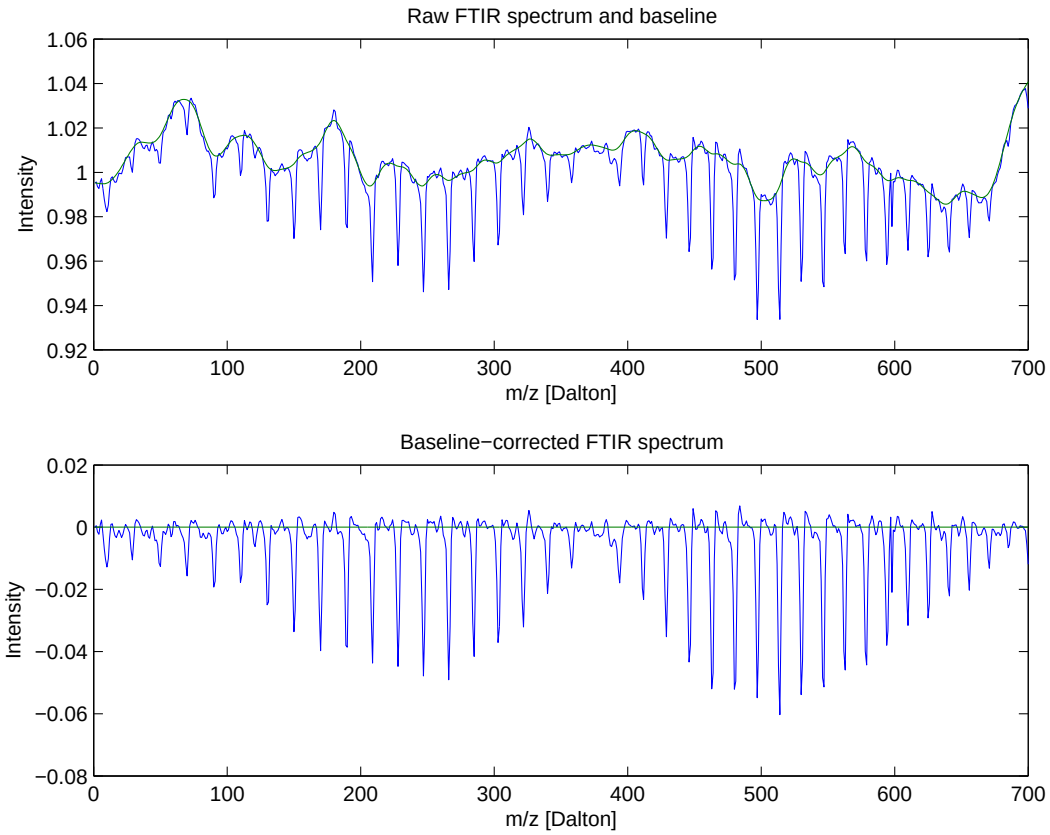

Figure 8: Baseline estimation for an FTIR spectrum. Top: spectrum and baseline; bottom: after baseline correction. The highly variable baseline dictates a small value of the penalty parameter: $\lambda = 100$ ($p = 0.9$).

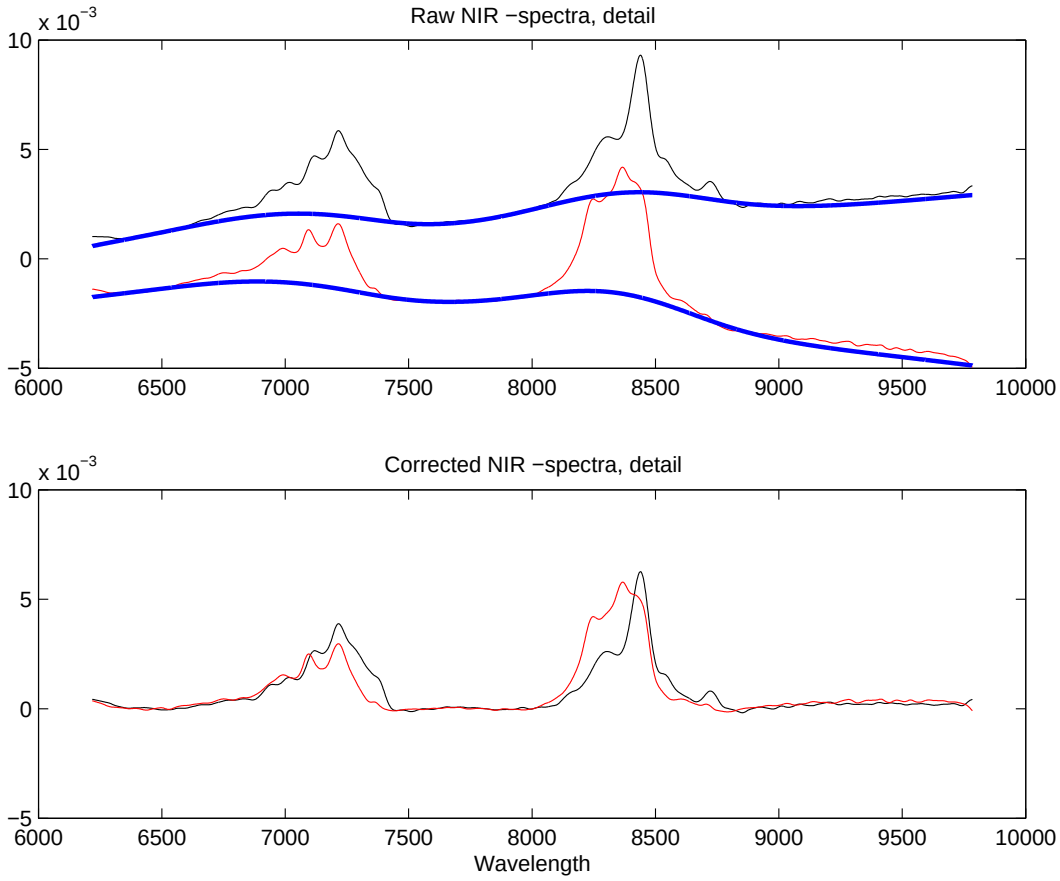

Figure 9: Two NIR spectra and baseline estimates. Parameter values: $p = 0.01$ and $\lambda = 10^7$.

We could get a better fit there by reducing λ, but then the baseline would be drawn appreciably more upward under the peaks.

4 Discussion

We have presented a fast, simple and effective algorithm for baseline estimation. Two parameters completely determine this baseline. A small simulation showed that a very good approximation to the real baseline for the optimal values of p and λ. Hence,

little distortion maybe expected of the analytical peak signal that should be retrieved. Unfortunately we were not yet able to present a fully automatic procedure to set the parameters to their "optimal" values. On the other hand, with narrow peaks, alternating with baseline sections with little noise, it is not hard to select suitable parameter values by visual inspection. In those cases sensitivity to the asymmetry parameter is small. Experience has already shown the value of the baseline correction algorithm in a variety of applications.

The Whittaker smoother is a natural choice, but asymmetric weighting can be combined with other smoothers too. In some cases it may be preferable to use P-splines[19], which combine a moderately-sized B-spline basis with a difference penalty. The advantage is that the size of the system equations in (3) is reduced to the size of the B-spline basis. This can be essential if sparse matrix software is not available and very long signals have to be processed.

P-splines can be extended to two and more dimensions with tensor products of B-splines and appropriate difference penalties. A fast algorithm for data on grids is available[20]. This can be adapted easily to an asymmetric least squares setting. Initial experience with GC-MS data, evolving spectra, and images have been positive. Results will be reported elsewhere.

More research is needed on data with broad peaks, where a baseline tends to be drawn into the peaks. The present difference penalty is a special, rather simple, way of encoding prior ideas about global smoothness. If reliable information is available about local variations in smoothness, this can be encoded in a variable penalty. To this end one can introduce penalty weights v and changes the penalty to $\lambda \sum_i v_i (\Delta^2 z_i)^2$.

There is much room for further improvements in other directions. The MALDI-TOF spectrum clearly shows that the strength of the noise diminishes with increasing values of m/z. This suggest that some form of additional weighting in the asymmetric sum of squares, to compensate for changes in the variance of the noise, my be beneficial. It also looks as if the flexibility of the baseline should be larger at low m/z. We are investigating these extensions in the framework of an explicit statistical model for

signals with pronounced baselines. The idea is to model a data series as the sum of three components: a smooth baseline, noise with a normal distribution and exclusively positive (negative) peaks having an asymmetric distribution on the positive (negative) axis [21]. Preliminary results have been encouraging, but more research is needed.

Another promising line of research is smoothing with an asymmetric L1 measure of fit, so-called quantile smoothing [22]. In an L1 norm the (weighted) sum of of absolute values takes the place of the (weighted) sum of squares. If the L1 norm is used in the penalty too, a linear program is obtained. The interior point algorithm allows efficient solution of this problem. A genomic application of quantile smoothing, replacing the smoothing spline with the Whittaker smoother, is described in [23].

Ruckstuhl et al. [9] report good results with an alternative deviation from symmetric least squares. They adapt an idea from robust statistics, a weighting function that essentially reduces the influence of observations far from the baseline to zero. This a combined with a local linear smoothing. This is highly non-linear problem and good starting values are essential to avoid getting stuck in a local minimum. Local regression can also have difficulties with broad peaks, because there may essentially be no data left (after weighting) to do regression on. In contrast, our penalty approach automatically interpolates even large areas and the convexity of the problem guarantees convergence to the global minimum.

Mazet et al.[10] use an explicit asymmetric goal function. For positive values the sum of squares is limited by a threshold. The baseline is modeled as a polynomial, which inherently limits its flexibility. There also is no guarantee that convergence to a unique minimum will occur.

An advantage of both the methods of Ruckstuhl et al. [9] and Mazet et al.[10] is that the small residuals $y - z$ are weighted symmetrically. This is not the case with our method: asymmetric weighting applies everywhere. But where residuals are small, we are on the baseline, not on the peaks, so we would like to have symmetric smoothing there. A fruitful area of research would be to develop hybrid methods, that starts with asymmetric smoothing, to come near the baseline with guaranteed convergence

and switches to an alternative weighting (symmetric for small residuals) in a second phase. This might also reduce the edge effect that sometimes occur with our method: the estimated baseline strays away from the signal at the edges, even when the edges apparently are part of the baseline (see Figure 1).

Acknowledgement We thank Peter Hamilton for providing the FTIR data, and Sijmen de Jong for providing the GC data. We also acknowledge the detailed comments of three anonymous reviewers that helped us to considerably improve our presentation.

References

[1] McNulty, D.A. ; MacFie, H.J.H. *J. of Chemom.* **1997**, *11*, 1-11

[2] Boelens, H.F.M. ; Kok, W.T. ; de Noord, O.E. ; Smilde, A.K. *Anal. Chem.* **2004**, *76*, 2656-2663

[3] Jirasek, A. ; Schulze G. ; Yu, M.M.L. ; Blades, M.W. ; Turner, R.F.B. *Appl. Spectrosc.* **2004**, *58*, 1488-1499

[4] Andreev, V.P. ; Rejtar, T. ; Chen, H. ; Mosovets, E.V. ; Ivanov A.R. ; Karger B.L. *Anal. Chem.* **2003**, *75*, 6314-6326

[5] Bogaert, B. ; Boelens, H.F.M. ; Smit, H.C. *Anal. Chim. Acta* **1993**, *274*, 71-85

[6] Phillips A.J. ; Hamilton P.A *Anal. Chem.* **1996**, *68*, 4020-4025

[7] Gemperline, P.J. ; Cho, J.H. ; Archer, B. *J. of Chemom.* **1999**, *13*,153-164

[8] Knee, M.A. and Annegarn H.J. *Nucl. Instrum. Methods Phys. Rese. Sect. B*, **1996**, *109*, 201-213

[9] Ruckstuhl A.F. ; Jacobson M.P. ; Field R.W. ; Dodd J.A. *J. Quant. Spectrosc. Radiat. Transfer* **2001**, *68*, 179-193

[10] Mazet, M. ; Carteret, C. ; Brie, D. ; Idier, J. ; Humbert, B. *Chemom. Intell. Lab. Syst.* **2005**, *76*, 121-133

[11] Schulze, G.S. ; Jirasek, A. ; Yu, M.M.L. ; Lim, A. ; Turner, R.F.B. ; Blades, M.W. *Appl. Spectrosc.* **2005**, *59*, 545-574

[12] Eilers, P.H.C. *Anal. Chem.* **2004**, *76*, 404-411

[13] Eilers, P.H.C. *Anal. Chem.* **2003**, *75*, 3631-3636

[14] Newey, W.K. ; Powell, J.L. *Econometrica*, **1987**, *55*, 819-847

[15] Efron, B. *Statist. Sinica* **1991**, *1*, 93-125

[16] Eilers, P.H.C. *Kwantitatieve Methoden* **1988**, *23*, 63-83

[17] Boelens, H.F.M. ; Eilers, P.H.C. ; Dijkstra, R. ; Fitzpatrick, F. ; Westerhuis, J.A. *J. Chrom. A* **2004**, *1057*, 21-30

[18] Boelens, H.F.M. ; Eilers, P.H.C. ; Hankemeier Th. *Anal. Chem.* **2005**, (in press)

[19] PHC Eilers P.H.C. ; BD Marx B.D. *Statist. Sci.* **1996**, *11*, 89-102

[20] Currie, I.D. ; Durbán, M. ; Eilers, P.H.C. *Stat. Methods* **2004**, *4*, 279-298

[21] Currie, I. ; Durbán, M. ; Eilers, P.H.C. In: Proceedings of the 18th International Workshop on Statistical Modelling, Leuven, **2003**; pp 97-102

[22] Koenker, R. ; Ng, P. ; Portnoy, S. *Biometrika*, **1994**, *81*, 673-680

[23] Eilers, P.H.C. ; de Menezes, R.X. *Bioinformatics*, **2005**, *21*, 1146-1153

List of Figures