in NEURAP'95, 8th Intern. Conference on Neural Networks and their Applications, March 20-23, 1996, Marseilles, France,
pp. 26-32 , Presses de l'Imprimerir SEDIMM, Marseilles

Constrained Topological Mapping for Baseline Construction

Bärbel Herrnberger & Horst–Michael Groß

Technical University of Ilmenau
Department of Computer Science and Automation, FG Neuroinformatik
98684 Ilmenau, Germany
{bahe,homi}@informatik.tu–ilmenau.de

Abstract. This paper addresses the use of self–organizing maps for baseline construction in chromatograms. Unlike local techniques, the problem is seen in terms of global optimization: a straight and smooth path including sampled points with high significance for baseline membership is to be found. For their smoothing capabilities, and for reproducing the probability density function of the input, self–organizing maps allow for balancing between these demands accomplishing a kind of nonparametric weighted regression. The significances are determined from feature extraction and feature fusion at a local scale. Applying global optimization, robustness is achieved in two ways: First, the result will align to the position of the majority of significant points, and, second, a single false decision on the local level won't be able to change the course of the baseline completely ensuring comparability of peak measurements in similar chromatograms. Comparability, however, is essential for routine analysis which is the intended field of application.

1 Motivation

1.1 Baseline Definition

Chromatographic separation is a widely used technique for quantifying complex mixtures of chemical substances. The chromatographic process decomposes the mixture into a sequence of their individual components resulting in an intensity vs. time profile – the chromatogram – as shown in figure 1. Substances can be distinguished by the positions of their corresponding peaks while the peak area indicates the amount of a substance.

An ideal chromatogram consists of well–separated Gaussian peaks. In reality, peak overlap, drifts of varying sign, negative peaks, and ruptures of the kind shown in figure 2 can be found. For peak measurement, and, therefore, for exact quantification, chromatograms have to be corrected such that the profile of all detected substances is obtained. A curve called baseline is searched for ignoring peak overlap, separating negative peaks from positive ones, and following drifts and ruptures.

For more details on chromatography, see e.g. [1].

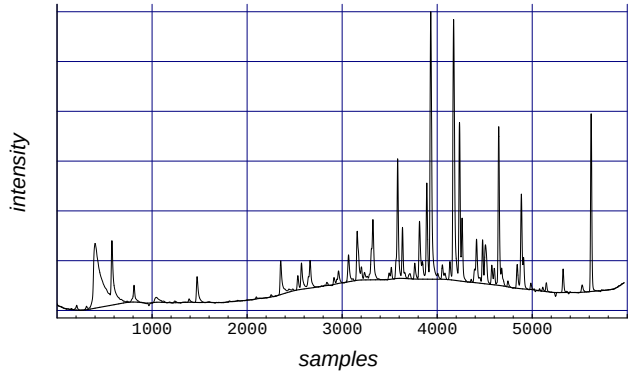

Fig. 1. Chromatogram and manually assigned baseline

1.2 Other Approaches

Constraints on slope and drift Most automatic strategies (e.g. [2]) identify the construction of a baseline which is described as a mainly horizontal and straight curve, with estimating slopes. Following this, baseline candidates are those points having small or zero slope. Having in mind the properties of real chromatograms, minima between overlapping peaks and the tips of negative peaks will be marked baseline members by mistake. This problem is usually handled by introducing a threshold upon drift. Though including additional data points for the decision, the strategy remains local retaining the following shortcomings:

1. Problems in using binary decisions: Threshold criteria are critical in two ways: First, including a data point into the baseline in one chromatogram and excluding it in another one generally results in significant differences in the course of the constructed baselines such that peak measurements in two similar chromatograms will possibly not be comparable. Second, adaptation problems occur: With the threshold being too low, the baseline cannot follow the chromatogram, otherwise, it runs into groups of overlapping peaks.

2. Adequacy of slope: Though baseline is a preferably horizontal signal, at a local level, differing orientations are allowed in principle. That is, even data points with higher slope can be baseline members [3].

Spectral analysis If baseline was a low–frequent signal, low–pass filtering should produce a baseline. Spectral analysis stresses the global aspect of a baseline, assuming low and high frequency parts of the spectrum separated from each other, which is not given here. Because an ideal peak of Gaussian shape contains low frequencies, too, low–pass filtering for extracting baseline will fail because even ideal chromatograms will be deformed. With sharp peaks, the

in NEURAP'95, 8th Intern. Conference on Neural Networks and their Applications, March 20-23, 1996, Marseilles, France, pp. 26-32 , Presses de l'Imprimerir SEDIMM, Marseilles

deformation is diminished. As in the case of ruptures, the baseline can even contain high frequencies.

In conclusion, there are problems with both purely local and unspecific global methods for baseline construction. The proposed conception introduced in the next section will combine local and global aspects of a baseline.

2 Conception

Local decisions on baseline membership without a close interplay with the course of the whole chromatogram don‘t seem to be reasonable. Considering this, a system of two stages is proposed. In the first and local stage, each data point is assigned a continuous value considered as its degree of baseline membership or significance. The second and global stage applies an optimization strategy for balancing these significances and for producing the final curve, simultaneously.

Expecting a single feature not sufficient for a decision on baseline membership, the local stage assigns to each of the data points a vector of features and fuses them for obtaining the significances. Globally, piecewise linear weighted regression is performed by constrained topological mapping with the weights given by these significances.

When describing a baseline as a mainly straight signal containing data points of high significance, self–organizing maps seem to be appropriate because of their topology preserving and smoothing capabilities as well as their ability to reproduce the probability density function of the input. Following the nodes of a one–dimensional map, the sequence of weights gives the knots of a piecewise linear path through the data. If probability is identified with significance, the network nodes shall shift their weights towards significant points in input space.

Obviously, for the performance of the proposed strategy, the choice of adequate features and their sensible fusion is essential. Nevertheless, the global stage does not expect all points actually belonging to baseline having significantly higher membership values than non–baseline points. Rather, besides significance, the position of an individual sampled point in relation to the position of the other significant points is considered. The self–organizing map balances significances such that points having high significances can be excluded from baseline, while points with low significance can be included according to the position of the majority of the significant data points; thus, significances are corrected implicitly. Single inappropriate significances should not have a large influence on the result.

The plan of the paper is as follows: Section 3 starts with describing the global stage; the determination of the significances is introduced in section 4. Results, problems, parameters as well as a comparison to smoothing splines is given in section 5.

3 Baseline Construction

3.1 Basic Idea

The idea is to interpret the chromatogram as a superposition of two (deterministic) signals: the baseline searched for and a disturbing signal given by (positive and negative) peaks.

Thus, baseline construction can be considered as a problem of approximating an unknown function of one variable (the baseline course) $y = f(x)$ by identifying the dependent variable y and the independent variable x with intensity and time, respectively (see figure 1).

Besides other model–free algorithms such as smoothing splines (see e.g. [6]), self–organizing maps have been proven for performing function approximation [7][8][9]. In their ability to handle multi–valued functions, in automatic knot positioning [7][8], as well as in numerical stability, they are superior to these methods.

3.2 Constrained Topological Mapping

Generally, in order to approximate an unknown function of $M-1$ variables

$$y = f(x_1, \ldots, x_m, \ldots, x_{M-1}) + \varepsilon \qquad (1)$$

in a given domain by a sequence of N vectors $(\hat{y}_n, x_{n,1}, \ldots, x_{n,m}, \ldots, x_{n,M-1})^T$, $n = 1, \ldots, N$, input samples

$$\boldsymbol{v}_k = (y_k, x_{k,1}, \ldots, x_{k,m}, \ldots, x_{k,M-1})^T$$

will be given, and, therefore, each of the N nodes of the self–organizing map carries a weight vector

$$\boldsymbol{w}_n = (w_{n,1}, \ldots, w_{n,m}, \ldots, w_{n,M})^T.$$

Here, $w_{n,1} = \hat{y}_n$ represents the dependent variable, and

$$\boldsymbol{w}_n^* = (w_{n,2}, \ldots, w_{n,M})^T$$

represents the independent variables

$$\boldsymbol{v}_k^* = (x_{k,1}, \ldots, x_{k,M-1})^T,$$

respectively.

For approximating single–valued functions of one or more variables, topology preserving [10] in the space of the independent variables is necessary that is not a priori given by applying the original algorithm of Kohonen. To ensure topology–preserving, the following constraints (constrained topological mapping, [7]) have to be put on weights and weight adaptation:

3. Initialization of the weights such that a topological order in the space of independent variables is given.

2. During the learning process, the preservation of that topological order is accomplished by searching for the best–matching node in the space of independent variables only.

in NEURAP'95, 8th Intern. Conference on Neural Networks and their Applications, March 20-23, 1996, Marseilles, France, pp. 26-32 , Presses de l'Imprimerir SEDIMM, Marseilles

Thus, weight adaptation for each of the N nodes in the map will be by

1. Finding the best–matching node c

$$c = c(\boldsymbol{v}_k) = \arg\{min_{n=1}^{N} \left\| \boldsymbol{w}_n^* - \boldsymbol{v}_k^* \right\|\} \quad (2)$$

4. Adapting all weights including the dependent variable

$$\boldsymbol{w}_n(t) = \boldsymbol{w}_n(t\text{–}1) + \Delta_k \boldsymbol{w}_n(t) \quad (3)$$

$$\Delta_k \boldsymbol{w}_n(t) = \eta(t)\, h_{cn}(t)\, (\boldsymbol{v}_k - \boldsymbol{w}_n(t\text{–}1)) \quad (4)$$

with h_{cn} being a monotonously decreasing neighbourhood function and $\eta(t)$ a learning rate, both being gradually shrunk. Clearly, the smoothness of the resulting curve depends on the final width of h_{cn}.

Allowing the adaptation of all the weights, there is correspondence to principal curves [11][12] minimizing the summed squared error in both the dependent and the independent variables.

3.3 Weighted Regression

Self–organizing maps reflect the probability distribution of the input: nodes accumulate in areas of high input density. If probability or input density is identified with significance, the resulting curve is expected to have its course through the most significant data points, that is, nodes should preferably shift their weights towards these points. Referring to [12], presenting input samples according to their probabilities, or modulating the learning rate by these probabilities produces qualitatively comparable results. With significances s_k given, (4) becomes

$$\Delta_k \boldsymbol{w}_n(t) = s_k\, \eta(t)\, h_{cn}(t)\, (\boldsymbol{v}_k - \boldsymbol{w}_n(t\text{–}1)) \quad (5)$$

that is, the degree of weight change and the significance value are directly proportional.

4 Local Processing: Obtaining Significances

For the local stage, the following conceptions hold:

Handling the chromatogram as a 2D image This is the basis of how an expert evaluates a chromatogram. There, a baseline can be described as a preferably horizontal signal touching data points at the lower border of the chromatogram, whereby, at a local level, differing orientations are allowed in principle. Referring to an image, positional relations between sampled points can be considered. Besides slope (see section 1), two features are computed: distance to points at the lower data border, and local point density.

Feature fusion The features assigned to each point are fused to a continuous value, its significance for baseline. In order to avoid false decisions at this stage, threshold functions leading to binary significances are not applied. Rather, a rule describing the dependency of the significances from the local features is given, which is handled in fuzzy logic terms. First, the features are identified with linguistic variables, then, appropriate membership functions for the corresponding linguistic terms are derived. The parameters of the membership functions and the type of inference is obtained from a cost function upon a set of pre–classified data points and their local features.

4.1 Local Features

Distance The idea is to introduce some blurring into the line image: each point is dilated[1] to an area called the structuring element. The curve resulting from an erosion at the lower border of this chromatogram area (using the same structuring element) is employed as the reference for the following distance measurement, where a feature $m_c \in [0, 1]$ describing the vertical distance to its corresponding point on that curve is assigned to each sampled point[2].

With line images as the chromatogram, there is no need for performing closing on the image itself, and, therefore, no need for scaling. With rectangular structure elements, dilatation is equivalent to a ‘running minimum', and erosion to a ‘running maximum‘, respectively. Both can be performed with low computational effort.

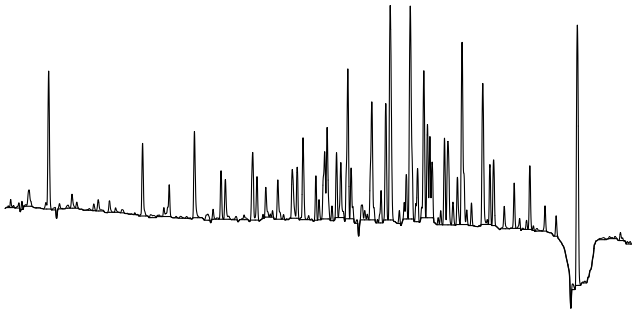

Fig. 2. Closing by a rectangular structuring element. Both negative peaks and baseline disturbance are included in the result.

Point Density If the chromatogram is seen as a scatterplot, there is – in general – a higher point density at the baseline. Because of a preferably horizontal baseline course, an anisotropic Gaussian filter is applied with the width larger than the height.

The obtained features m_d are shown in figure 2. Though the computation of this feature needs a 2D representation of the chromatogram, there is, again, low computational effort: point density has to be determined only in certain positions in the image (the positions the data points project to).

Here, scaling will be needed but adjusting a significant peak to a given height/width ratio and applying the resulting scaling factor to the whole chromatogram seems to be sufficient. Scaling is considered a less critical factor because of

1. For an overview of image processing methods see e.g. [4].
2. All features are normalized to the feature extrema of the chromatogram the data point belongs to – features are relative.

in NEURAP'95, 8th Intern. Conference on Neural Networks and their Applications, March 20-23, 1996, Marseilles, France, pp. 26-32 , Presses de l'Imprimerir SEDIMM, Marseilles

the continuous filter surface, i.e., small changes in the filter size will produce small density changes

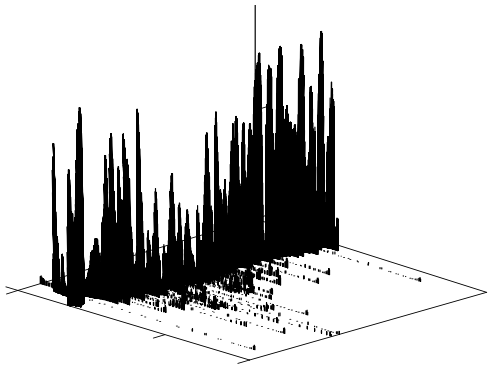

Fig. 3. Local point density

4.2 Feature Fusion

The significance of a sampled point for baseline should be high, if there is a small slope m_s, a small distance m_c, and a high point density, m_d.

The feature analysis supporting this hypothesis is given in figure 4 showing the histograms of the individual features for baseline and non–baseline points, separately.

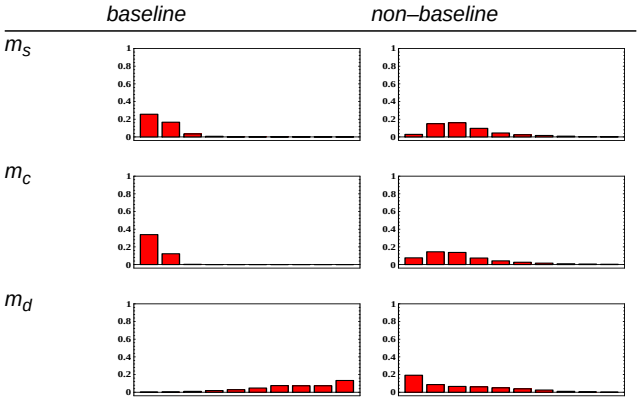

Fig. 4. Feature histograms (nonlinear transformation of m_s and m_c into [0,1]). Most but not all of the baseline points meet the conditions above. Values are divided into h=10 classes with class i collecting values from the interval $[(i-1)/h, i/h]$; $i = 1, \ldots, h$.

With principal forms given for the membership functions for both the linguistic terms ‘high‘ and ‘small‘,

$$S(x, a, \delta) = \begin{cases} 0, & x \le a - \delta \\ 2\left(\frac{x - a + \delta}{2\delta}\right)^2, & a - \delta < x \le a \\ 1 - 2\left(\frac{x - a + \delta}{2\delta}\right)^2, & a < x < a + \delta \\ 1, & x \ge a + \delta \end{cases} \quad (6)$$

and

$$Z(x, a, \delta) = 1 - S(x, a, \delta)\ , \quad (7)$$

the kind of feature fusion is obtained from the following steps:

1. Optimizing the parameters a (inflection point) and δ (width) of both functions according to a cost function using a training set of feature vectors $(m_s, m_d, m_c)^T$ together with their classification as a baseline or a non–baseline member.

2. Specifying an operator for fusing the individual conditions stated in the rule, again derived from a cost function.

Optimizing Membership Functions Optimization is intended to sharpen the discrimination between baseline and non–baseline points. For each of the linguistic terms $M \in \{M_s, M_c, M_d\}$ bound to the individual features, the parameter set $\boldsymbol{p} = (a, \delta)^T$ was determined by minimizing the cost function

$$C(M) = \frac{1}{2K} \sum_{k=1}^{K} (1 - M_k)^2 b_k + M_k^2 (1 - b_k)$$

where $M = \mu(m, \boldsymbol{p})$, $M_k = \mu(m_k, \boldsymbol{p})$ (8)

upon a set of K manually pre–classified data points from various chromatograms of different forms. (A sampled point is classified as a baseline member, if it fits a baseline drawn by an expert.) Here, $\mu(m, \boldsymbol{p})$ is the membership function describing M, and M_k is the value of μ applied to feature m of data point k, m_k. Baseline membership b_k is coded binary ($b_k = 1$ indicating baseline members, $b_k = 0$, otherwise). This function will penalize both non–baseline points with high membership values and baseline points having small ones. The optimization result is shown in figure 5.

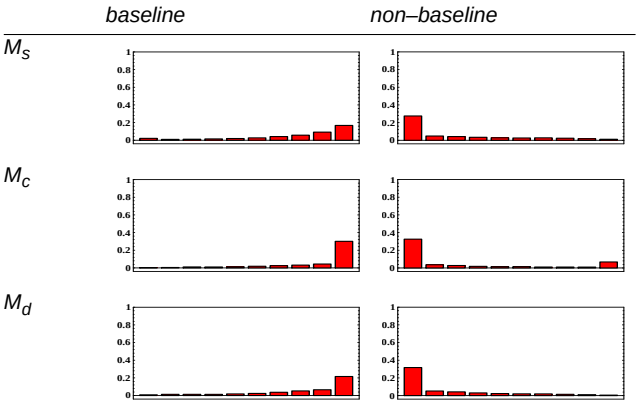

Fig. 5. Histogram of fuzzified features. Discrimination has been sharpened, but not all of the baseline points have got significantly higher membership values than non–baseline points, yet.

Fusion The kind of feature fusion is again derived from (8), with the argument M being replaced by an operator $op(M_s, M_c, M_d)$ and M_k replaced by $s_{k_{op}} = op(M_{s,k}, M_{c,k}, M_{d,k})$, respectively. The value s_k will be taken for the significance of point k for baseline.

in NEURAP'95, 8th Intern. Conference on Neural Networks and their Applications, March 20-23, 1996, Marseilles, France, pp. 26-32 , Presses de l'Imprimerir SEDIMM, Marseilles

From the set of non–parametric average operators *op* (minimum, geometric mean, arithmetic mean, dual–of–geometric–mean, maximum) chosen to cope with inconsistencies within the features, geometric mean was found to produce minimal costs. See figure 6 for results.

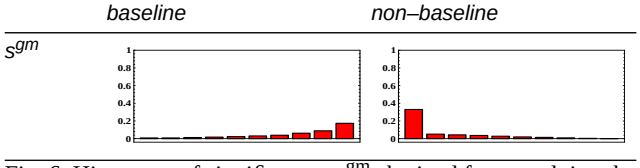

Fig. 6. Histogram of significances s^{gm} obtained from applying the geometric mean operator.

5 Application

5.1 Network, Parameters, Results

Approximating a function of one variable, which is the case in baseline construction, the self–organizing map consists of a chain of nodes with $M = 2$ weights each. As the best matching node will be computed in only one dimension, again, there is no need for scaling.

The weights are initialized such that their projections into the input space form a straight line connecting the first and the last sampled point of the chromatogram. In the course of weight adaptation, each input or sampled chromatogram point excerts a force on this line with the strength of deformation depending on the size of its significance.

As done in [7], shrinking the learning radius σ of the neighbourhood function

$$h_{cn}(t) = \exp\left(-\frac{(c-n)^2}{2\sigma(t)^2}\right) \tag{9}$$

is bound to shrinking the learning rate η

$$\eta(t) = \eta_0\left(\frac{\eta_1}{\eta_0}\right)^{t/t_{max}}, \tag{10}$$

by

$$\sigma(t) = \eta(t)\frac{S_0}{\sqrt{2}}. \tag{11}$$

In (9), c and n denote the indices of the best matching and an arbitrary node in the chain, respectively; in (10) and (11), η_0 and η_1 give initial and final values of the learning rate, t and t_{max} are the number and the maximal number of learning steps.

The qualitative course obtained is mainly dependent on the ratio N/S_0.

Results are shown in figure 7.

Matching sampled points There is a conceptual problem inherent in the regression technique used: Because of minimizing the distance (in the dependent variable) between the individual sampled points and the resulting curve, or, more specifically, minimizing the sum of squared errors in connection with a smoothness constraint (12), points of high significance will not necessarily be included into the resulting curve – a qualitative course of a baseline is obtained. However, 'qualitative' does not really mean a limitation of the method if one keeps in mind that the chromatogram itself is subject to random disturbances.

The problem could be tackled by putting a threshold on the distance of individual data points to the obtained curve and by connecting only points with distances below threshold for baseline construction. This threshold could be given absolutely (but depending on the intensity range of the chromatogram), or – avoiding an additional parameter – by statistical methods.

Contraction Because of the vector quantisation properties of self–organizing maps, a contraction of the weights towards the weighted centre of gravity of the input data is unavoidable. That contraction can be suppressed by excluding the first and the last network node from weight adaptation.

5.2 Constrained Topological Mapping vs. Smoothing Splines

A short comparison of the performance of smoothing splines and constrained topological maps is given emphasizing the choice of constrained topological maps for baseline construction.

Smoothing splines minimize a cost function C consisting of two contrary conditions: The first term in (12) forces the resulting curve to include all data points, while the second is for obtaining a straight curve:

$$C(p) = R(p) + \frac{1}{p}I(p) \tag{12}$$

$$R(p) = \frac{1}{K}\sum_{k=1}^{K}(\hat{y}_k - y_k)^2$$

$$I(p) = \int_{x_1}^{x_k}[f_S(t,p)]^2 dt$$

with $\hat{y}_k$ being the approximation of the sampled data point y_k at position x_k. Furthermore, f_S is the overall spline function consisting of a sequence of polynomials valid in succeeding intervals.

For weighted splines employing weights s_k, $R(p)$ becomes

$$R(p) = \frac{1}{K}\sum_{k=1}^{K}s_k(\hat{y}_k - y_k)^2 \tag{13}$$

With equal knot positions and number of nodes in the map, the smoothing parameter p balancing the influence of both $R(p)$ and $I(p)$, and the final learning radius of the constrained topological maps show comparable effects. That is,

in NEURAP'95, 8th Intern. Conference on Neural Networks and their Applications, March 20-23, 1996, Marseilles, France, pp. 26-32 , Presses de l'Imprimerir SEDIMM, Marseilles

with a large learning radius or a small parameter p, smoothing predominates.

Caused by the iterated procedure, weight adaptation in self–organizing maps demands extended computational effort. This disadvantage is compensated for by numerical stability: The computation of splines comes down to an inversion of a system matrix. With small parameters p, with large data vectors, and with using weights, the invertability of this matrix is not necessarily guaranteed but has to be determined on the data at hand, which, in turn, is time consuming.

Last, using splines, there will be a problem in handling zero weights which are allowed by feature fusion. There, data points with zero weights have to be eliminated from the data vector leading to the computation of non–equidistant splines resulting in increased computational effort, again. By eliminating single data points, the probability for an overshooting of the resulting curve rises, such that additional efforts have to be taken.

6 Discussion

In this approach, baseline construction in chromatograms has been stated from a global point of view. A baseline was defined as a straight path including data points of high significance. Significances have been obtained locally by extracting a couple of features and by fusing them. However, a clear decision on baseline membership by these significances will not be obvious in general. Therefore, a global stage is considered for producing the final curve correcting the significances according to global demands, implicitly.

To accomplish this, weighted nonparametric regression is performed by constrained topological mapping. In the present application, there are some important reasons for preferring this method to ‘classical‘ ones, e.g. smoothing splines.

Obviously, the result depends on the appropriateness of the significances obtained in the local stage, and therefore, on the adequacy and on the parameters of feature extraction.

Results could be further improved if the parameters of feature fusion would be optimized on a training set containing data points of similar chromatogram forms – at the expense of losing generalization.

The proposed strategy leads to robustness in the following sense: Because the resulting curve aligns to the position of the majority of points evaluated as significant, no single false decision on the local level will be allowed to disturb the curve or will be able to change the course of the baseline completely. Considering global aspects, comparability of peak measurements in similar chromatograms is obtained, which is essential for routine analysis, the intended field of application of the proposed strategy.

References

[1] A. Braithwaite and F.J. Smith. Chromatographic Integration Methods. Chapman & Hall, 1992

[2] Hewlett Packard Company. HP3350 User Reference Manual, 1987

[3] N. Dyson. Chromatographic Integration Methods. RSC Chromatography Monographs. The Royal Society of Chemistry, 1990

[4] R.M. Haralick and L.G. Shapiro. Computer and Robot Vision. Addison–Wesley, 1993

[5] H.–J. Zimmermann. Fuzzy set theory and its applications. Kluwer Academic Publishers, 1991

[6] P. Diercks. Curve and Surface Fitting with Splines, Clarendon Press, 1993

[7] V. Cherkassky and H. Lari–Najafi. Constrained Topological Mapping for Nonparametric Regression Analysis. Neural Networks 4:27–40, 1991

[8] V. Cherkassky and F. Mulier. Self–Organizing Networks for Nonparametric Regression. In: V. Cherkassky, J.H. Friedman, and H. Wechsler (eds.) From Statistics to Neural Networks, pp. 188–212, Springer, 1993

[9] T.J. Hastie and R.J. Tibshirani. Nonparametric Regression and Classification; Part I – Nonparametric Regression. In: V. Cherkassky, J.H. Friedman, and H. Wechsler (eds.) From Statistics to Neural Networks, pp. 62–69, Springer, 1993

[10] Th. Martinetz and K. Schulten. Topology Representing Networks. Neural Networks 7(3):507–522, 1994

[11] T. Hastie and W. Stuetzle. Principal Curves. Journal of the American Statistical Association, 84(406):502–516, 1989

[12] H. Ritter, Th. Martinetz, and K. Schulten. Neural Computation and Self–Organizing Maps. Addison–Wesley, 1991

in NEURAP'95, 8th Intern. Conference on Neural Networks and their Applications, March 20-23, 1996, Marseilles, France, pp. 26-32 , Presses de l'Imprimerir SEDIMM, Marseilles

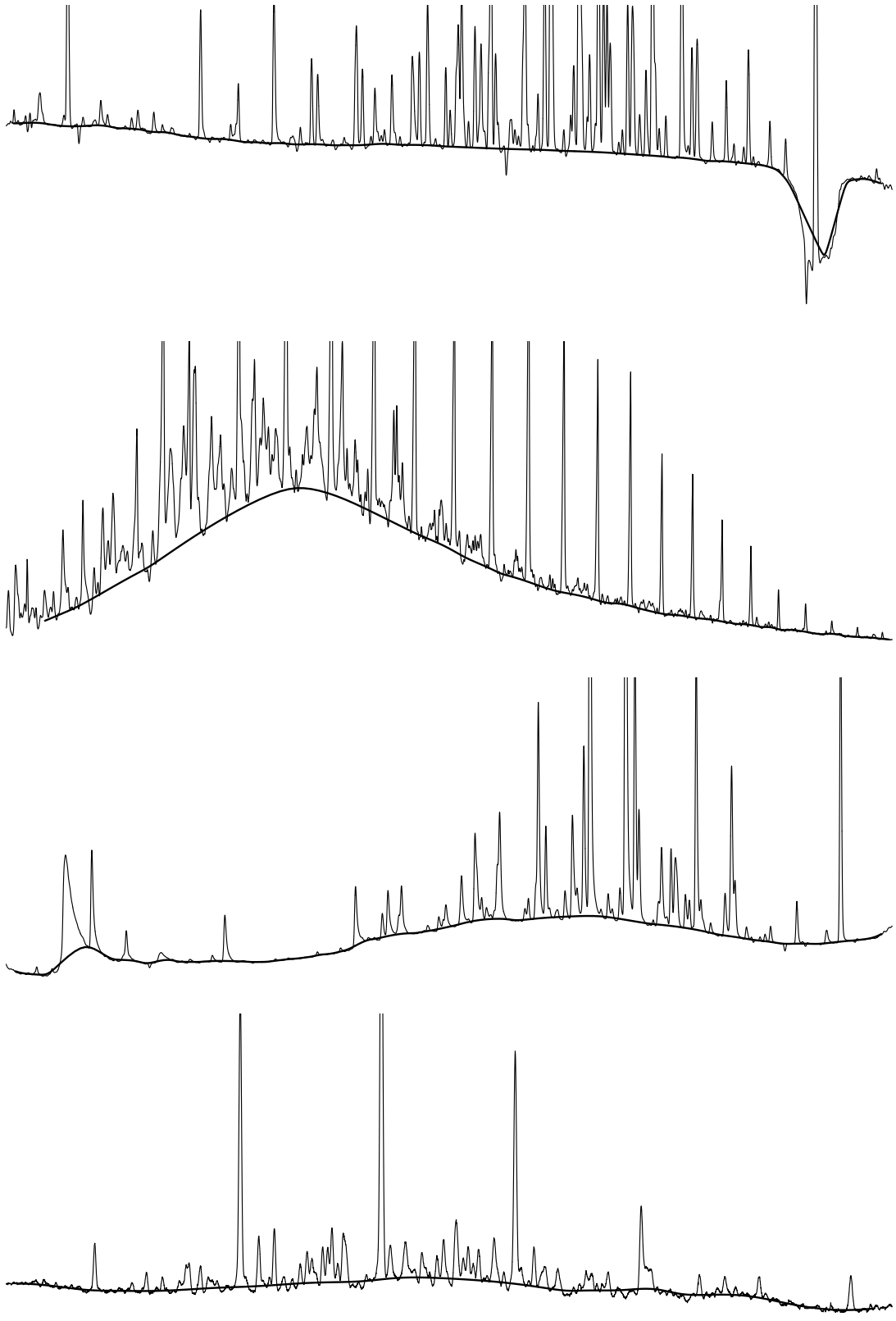

Fig. 7. Result of the proposed algorithm in four chromatograms of different form and with different number of sampled points. Parameters: $N = 300$, $\eta_0 = 1,0$, $\eta_1 = 0,05$, $t_{max} = 15K$, $S_0 = (80, 80, 80, 200)$