Anal Bioanal Chem (2009) 395:213–223
DOI 10.1007/s00216-009-2940-4

ORIGINAL PAPER

A solution to the 1D NMR alignment problem using an extended generalized fuzzy Hough transform and mode support

Erik Alm · Ralf J. O. Torgrip · K. Magnus Åberg · Ina Schuppe-Koistinen · Johan Lindberg

Received: 10 April 2009 / Revised: 24 June 2009 / Accepted: 25 June 2009 / Published online: 22 July 2009

Abstract This paper approaches the problem of intersample peak correspondence in the context of later applying statistical data analysis techniques to 1D ^{1}H-nuclear magnetic resonance (NMR) data. Any data analysis methodology will fail to produce meaningful results if the analyzed data table is not synchronized, i.e., each analyzed variable frequency (Hz) does not originate from the same chemical source throughout the entire dataset. This is typically the case when dealing with NMR data from biological samples. In this paper, we present a new state of the art for solving this problem using the generalized fuzzy Hough transform (GFHT). This paper describes significant improvements since the method was introduced for NMR datasets of plasma in Csenki et al. (Anal Bioanal Chem 389:875-885, 15) and is now capable of synchronizing peaks from more complex datasets such as urine as well as plasma data. We present a novel way of globally modeling peak shifts using principal component analysis, a new algorithm for calculating the transform and an effective peak detection algorithm. The algorithm is applied to two real metabonomic ^{1}H-NMR datasets and the properties of the method are compared to bucketing. We implicitly prove that GFHT establishes the objectively true correspondence. Desirable features of the GFHT are: (1) intersample peak correspondence even if peaks change order on the frequency axis and (2) the method is symmetric with respect to the samples.

Keywords Metabolic profiling · NMR · Peak detection · Image processing · Hough transform · Synchronization · Alignment

Electronic supplementary material The online version of this article (doi:10.1007/s00216-009-2940-4) contains supplementary material, which is available to authorized users.

E. Alm · R. J. O. Torgrip (✉) · K. M. Åberg
Dept. of Analytical Chemistry, BioSysteMetrics Group,
Stockholm University,
106 91 Stockholm, Sweden
e-mail: ralf.torgrip@anchem.su.se

R. J. O. Torgrip · K. M. Åberg · I. Schuppe-Koistinen · J. Lindberg
AstraZeneca R&D Södertälje, Safety Assessment,
Molecular Toxicology,
151 85 Södertälje, Sweden

Introduction

The objective of this paper is to demonstrate a method capable of establishing intersample correspondence between spectral peaks in data obtained using ^{1}H-nuclear magnetic resonance (NMR) spectroscopy, typical of biological samples. The problem of noncorrespondence is well known—it is a fact that peaks do not remain in the same intersample position on the frequency axis. Varying chemical and physical sample properties induce peak shifts that make multivariate (or univariate) analysis of the raw data matrix confusing or pointless. It is not even likely that all peaks retain their relative order; this phenomenon originates from the fact that different analyte protons have different *shift sensitivity* to varying sample properties such as pH, salt content, temperature, etc. We hypothesize that the concept commonly referred to as the *matrix effect* and its effect on peak location are in fact deterministic, not random—a hypothesis that the presented work supports and fully exploits.

Traditionally, the first approach to remedy the encountered synchronization problem is to minimize the variation of physical and chemical parameters by controlling the physicochemical properties of the sample using, e.g., buffers and isothermal analytic conditions, but this does not fully remedy the problem.

The second approach used is to preprocess the data in such a way so that the influence of peak shifts is removed. Suggested solutions are algorithmic of nature and include, inter alia, bucketing, warping, and alignment approaches. Bucketing [1–3] is the most straightforward of these techniques and uses a piecewise integration of preselected fixed spectral segments. The recognized problems with bucketing are that (1) several unrelated peaks can end up in the same bucket and (2) a single peak can be split between buckets. Attempts to resolve the bucketing problems have been made by, e.g., dynamically selecting the global bucket boundaries, but the fact remains that this technique destroys the information contained in a high-resolution NMR spectrum.

Warping [4–7] is another approach to solve the correspondence problem; this class of techniques is most widely used for aligning chromatographic data. Warping works by establishing a transfer function that operates on the time or frequency axis of the sample to be warped. The transfer function maps points of the target and sample axis to reach correspondence. After the transfer function is established, the axis of the sample is transformed by insertion, deletion, or interpolation to reach a warped (synchronized) spectrum. Examples of warping techniques are correlation-optimized warping [8, 9] and dynamic time warping [10, 11]. These algorithms generally work well on chromatographic data and simple NMR spectra but perform unsatisfactorily in crowded regions of complicated spectra and fail when peaks change places. These methods also rely on the choice of a target spectrum, i.e., the methods are *not* symmetric with respect to the order of which spectra are processed. Furthermore, warping parameters are subjected to manual selection resulting in results that may vary considerably depending on biased choice.

Another class of alignment techniques such as peak alignment using reduced set mapping [12–14] is based on reducing continuous spectra to peak lists using peak detection. These peak lists are then matched over samples using an appropriate choice of algorithm, e.g., a tree search. The alignment class of algorithm does not use a continuous warping function. Both the warping and alignment class of algorithms are *incapable* of establishing correspondence when peaks change order.

The Hough-based algorithm presented in this work belongs to the alignment family of techniques as it uses a sparse peak list and establishes correspondence without the use of a transfer function. The presented method *is* capable of assigning correct correspondence even when peaks change order.

In a previous paper, we introduced the generalized fuzzy Hough transform (GFHT) as a way to establish correspondence by finding shift patterns associated with physical and chemical sample properties [15]. The Hough transform is originally an image analysis algorithm [16–22], so in this context the NMR dataset is treated as an image comprising a samples × bins matrix (pixels). From this image, using peak detection, a new matrix **X** is constructed wherein pixels where a peak maximum has been detected is assigned the value 1, while the rest are given the value 0. In its most simple form, the GFHT for NMR data can be described as follows. One clearly assignable peak is chosen for an entire dataset. The intersample peak positions are recorded as a *shift pattern.* This shift pattern multiplied by an expansion parameter α (considered as an expansion of the shift pattern along the frequency axis) is used as a model to describe the shifts of peaks throughout the entire dataset. The GFHT is used analogously as in image analysis to find parameterized shapes in an image, i.e., the Hough is iterated through predetermined values of the parameter α while recording the Hough score h which measures how well the current parameters and shift pattern describe the peak shifts. The Hough score is recorded in a matrix **H** (denoted the Hough transform space) designed to encompass the parameter span and its resolution. In the NMR context, each maximum in the Hough transform space corresponds to a parameter set that matches the positions of a peak throughout the *entire* dataset (all samples). The success of this matching process is dependent on the observation that the peak shifts can be described by this single-parameter model. In the previous paper, we showed that for plasma ^{1}H-NMR data the single-parameter model was adequate, but for the more complexly shifting urine data the results were not satisfying.

In this work, the GFHT approach is taken several steps further:

- To reach a solution where more complex data (in terms of peak shifts) such as urine spectra also can be aligned, we have incorporated a multicomponent peak shift model (MCSM) of the peak shifts based on principal component analysis (PCA) [23, 24]. The MCSM is derived from a selection of *a number of model peaks* whose individual peak shift patterns are collated and used as the basis of a PCA model of latent peak shift patterns. Linear combinations of the MCSM components are now used to test for matches of *several different* peak shift patterns in the NMR data.
- Because the incorporation of the MCSM significantly adds to the computational complexity of the GFHT transform by adding dimensions to the Hough indicator tensor (HIT), we present a more efficient algorithm, i.e., a list implementation of the algorithm, for performing the calculations.
- We show that naïve sample classification can be used to find peaks that are specific for a group of samples by partitioning the HIT.

- Since the GFHT is dependent on peak detection, we have included the peak detection method used in this work.

We demonstrate the extended GFHT alignment using two already-published ^{1}H-NMR datasets of different origin, size, complexity, and acquisition mode. We show that the GFHT establishes peak correspondence and that the presented new additions make the extended GFHT a powerful alignment technique fully capable of aligning complex ^{1}H-NMR datasets such as the ones encountered in bioanalysis.

Method

We traverse the extended GFHT method by discussing the data used, followed by an elaboration on the peak detection algorithm, the implementation of PCA to establish the MCSM, and end with a section demonstrating a faster way of calculating the GFHT score, which is used to find the parameters (α_i) for corresponding peaks.

Datasets

Briefly, the *Arabidopsis* dataset comprises manually designed samples made to mimic the metabolome of the plant *Arabidopsis thaliana* [25]. The *Arabidopsis* set contains 24 64 k spectra recorded on a Varian 600-MHz instrument using 2D ^{1}H–^{13}C HSQC acquisition. The samples contain 27 compounds, 24 biologically relevant molecules, and three nonbiological standards. Seven of the biologically relevant compounds are varied to mimic six different phenotypes of *A. thaliana*; the rest are kept constant. The concentrations of the seven varying compounds are used as reference ("ground truth") in the validation of the GFHT method. The *Arabidopsis* set is considered as "controlled" with respect to physicochemical parameters. The samples were titrated to an observed pH of 7.400 (±0.004) and the data contain no true biological variation. The *Arabidopsis* data still exhibit peak shifts, indicating that peak shifts in H-NMR data are hard to avoid.

The second dataset is a rat urine dataset collected during a toxicity study of the metabolic impact of ethionine [12]. The ethionine set comprises 336 64 k spectra collected on a Bruker 600-MHz instrument using NOESY acquisition. One of the dosing groups, the high single dose (five rats sampled twice per day for 7 days totaling 35 spectra) is used to visualize the metabolic impact of the toxin. In the validation of the ethionine, the class labels are arranged into two groups; the high single dose in one group (dosed days only) and the rest of the samples in another group to more easily interpret the GFHT validation results in terms of PCA score plots.

Bucketed data were created using a bucket size of 0.04 ppm with removal of the internal standards, resulting in 256 buckets for both datasets. The full experimental procedures and details for the *Arabidopsis* and ethionine sets are described in the Electronic Supplementary Material.

Peak detection

In its original image analysis application, the Hough transform space is more easily interpreted when calculated on a sparse feature (edge)-detected image. The same holds true for the GFHT application to ^{1}H-NMR data—a sparse peak-detected matrix is required for the algorithm to yield distinct maxima in the HIT (denoted Hough indicator array in the previous paper). Any peak detection algorithm could potentially be used with the GFHT but the results will depend on the completeness of the peak lists generated. In this paper, we present a *naïve zero-area filter* that is used for peak detection. The filter is created without any prior assumptions about the data.

The filter is derived from the internal TSP/DSS standard peak of the (phase-corrected) spectrum to be peak-detected but any baseline-separated peak of modest intensity could also be used. Using a real peak to derive the filter shape is preferred over using a theoretical lorentzian peak (or any other peak shape depending on spectral preprocessing) because to some extent the filter derived from the real peak can compensate for global phenomena such as bad shim and bad phasing. To construct the filter, set the original data matrix as $\mathbf{Z}$. A segment around the TSP/DSS peak in each spectrum ($\mathbf{z}$) is extracted and the second derivative with reversed sign of this segment, normalized to a sum of zero, constitutes the filter $g_{\mathrm{norm}}(\mathbf{z})$.

$$\mathbf{g}(\mathbf{z}) = -\frac{d^2 y}{dz^2} \tag{1–2}$$

$$\mathbf{g}_{\mathrm{norm}}(\mathbf{g} > 0) = \mathbf{g}(\mathbf{g} > 0)\,\frac{-\sum_{\mathbf{g}(\mathbf{z})<0}\mathbf{g}(z)}{\sum_{\mathbf{g}(\mathbf{z})>0}\mathbf{g}(\mathbf{z})}$$

The filter and the spectrum are convolved. After this filtering pass, a Lorentzian is fitted to each of the local maxima in the convolved vector by least squares. If the fitted Lorentzian does not have a maximum peak value greater than three times the noise standard deviation, the peak is discarded. The algorithm used for peak detection is:

1. Calculate the noise standard deviation in an empty part of the spectrum
   (e.g., −3.5 to −0.5 ppm for ethionine).

2. Cut out a section, 0.06 ppm wide, around the internal standard (TSP/DSS) peak at 0 ppm. Process this section by a, e.g., Savitzky–Golay [26] second derivative with reversed sign and adjust the resulting filter to a sum of zero by multiplying the positive part of it by an appropriate factor (Eq. 2) resulting in g_{norm}
3. Convolve the entire spectrum ($\mathbf{z}$) with g_{norm}.

$$\mathbf{z}_{\mathrm{c}}(n) = \sum_{k=-\infty}^{\infty} \mathbf{z}(n-k) \cdot \mathbf{g}_{\mathrm{norm}}(k) \tag{3}$$

4. Detect all local maxima in the convolved spectrum ($\mathbf{z}_{\mathrm{c}}$) and record their position and sample number into a list.
5. For each maximum in the list, fit a Lorentzian plus a linear baseline model to the raw data ($\mathbf{z}$) to a window around the maximum using least squares. The objective function for the peak-fitting step is:

$$e(k,m,b) = \sum_{x=x_0-i}^{x_0+i} \left(y(x) - \left(k(x-x_0) + m + \frac{ab}{(x-x_0)^2 + a^2} \right) \right)^2 \tag{4}$$

where a, the peak shape, is derived from the TSP/DSS peak and held constant throughout the spectrum; x_0 is the peak mode position; $y(x)$ the intensity and $2i+1$ the width of the local segment in data points. m is a constant baseline component. Discard peaks with S/N less than three.

6. The positions of the remaining detected peaks are inserted as ones in $\mathbf{x}$ or (for the updated way of calculating the HIT) stored in a list together with the maximum value of the fitted Lorentzian intensity and the spectrum they were found in. This yields a list of dimensions three times the number of detected peaks.

Figure 1 depicts the extraction, derivation of the filter, convolution, and peak detection results. Depending on the noise level, this algorithm typically reveals between 1,000 and 1,500 detected peaks per urine ^{1}H-NMR spectrum (600-MHz instrument). The peak detection is able to detect most shoulders and overlapping peaks. However, the algorithm does not perform well for peaks with a shape that heavily deviates from the shape of the internal standard peak, e.g., spinning sidebands and urea.

Principal component analysis of shift patterns

The extended GFHT alignment method described in this paper is based on the previous paper but extends the shift pattern analysis by the addition of principal component analysis revealing *underlying* (*latent*) *shift patterns*; we denote this model as MCSM. To establish the MSCM of the shift patterns, the shift pattern of *several* (typically around ten to 15) easily assignable peaks are selected; their peak position was recorded and arranged into a matrix (samples × peaks). After mean centering, a PCA of the peak location matrix yields a few significant components where the score vectors constitute the underlying shift pattern, see Fig. 2.

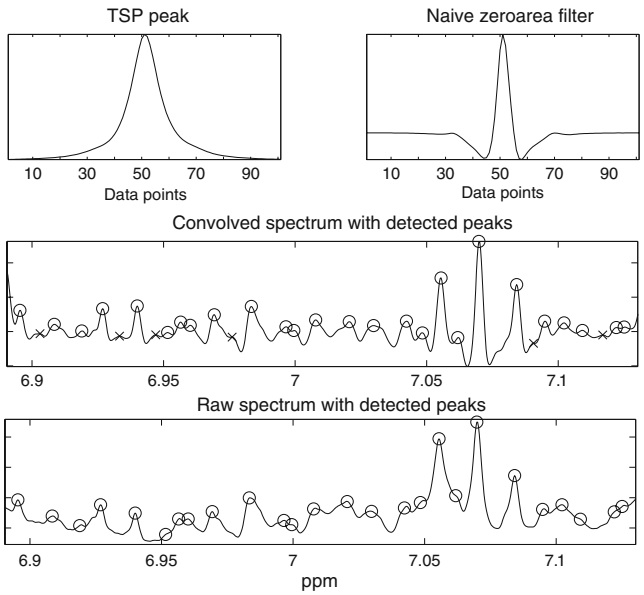

Fig. 1 Naive zero-area filter peak detection of a small section of a ^{1}H-NMR spectrum of rat urine. *Top left panel*: extracted TSP peak. *Top right panel*: the derived zero-area filter shape. *Middle panel*: spectrum convolved with the filter shape. (*circles*) Indicates detected peaks and (*x*) indicate possible peaks that were detected by the filter but then discarded in the lorentzian fitting step. *Bottom panel*: peaks detected in the spectrum

The relative magnitudes of the corresponding eigenvalues (or explained variance) indicates the rank of the underlying shift phenomena and hence the number of latent shift phenomena occurring in the data. The success of the extended GFHT method depends on the assumption that there are relatively few (one to five) significant latent shift components; otherwise, the resulting size of the HIT will pose a computational obstacle.

Calculating and interpreting the Hough transform space

The MCSM is used by the extended GFHT to search the data $\mathbf{X}$ for peak position matches using linear combinations of the significant score vectors as match patterns. The model for the location of a peak in all samples is:

$$\boldsymbol{\delta} = k + \alpha_1 \mathbf{s}_1 + \ldots + \alpha_K \mathbf{s}_K \tag{5}$$

Where $\boldsymbol{\delta}$ is a vector of peak locations tested in the HIT; k is the average location of the peak; K is the rank of the

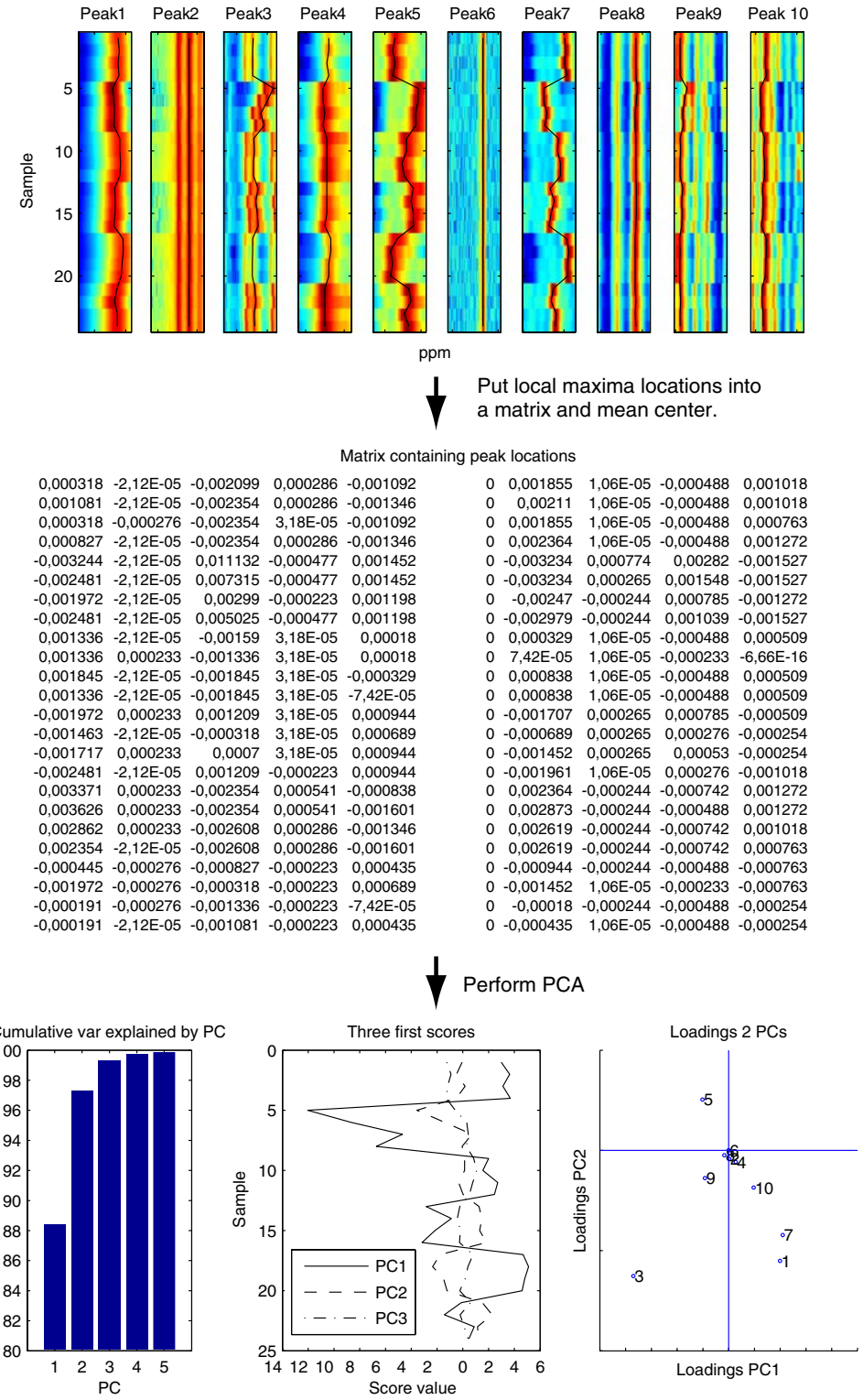

Fig. 2 PCA of the shift patterns derived from ten peaks in the synthetic *Arabidopsis* dataset. The *top row* shows heat maps of the ten peaks, with local maxima (peak detected) marked with a *black line*. The positions of the peak maxima form a 24×10 matrix of parts per million values which are analyzed with PCA (*bottom panel*). The cumulative variance-explained plot (*bottom left panel*) shows that two or three PCs describe the peak shifts well. The scores (*bottom middle panel*) constitutes the MSCM and are further used as a model for all peak shifts. The loadings (*bottom right panel*) show the magnitude of the first two shift vectors (scores) that is needed to explain the shift of the ten peaks

MCSM model; $\alpha_1 ... \alpha_K$ are the shift pattern expansion parameters and $\mathbf{s}_1 \ldots \mathbf{s}_K$ are the corresponding score vectors (latent shift pattern). Next, the range and resolution of $\alpha_i \mathbf{s}$ to be tested is user-defined and the calculation of the Hough score for all combinations of α_i for all positions where peaks are present is performed. The initial MSCM model gives an indication about the magnitude of the $\alpha_i \mathbf{s}$ and the resolution is usually set around 20–50 steps between $\min(\alpha_i)$ and $\max(\alpha_i)$. The calculated Hough scores are stored in the HIT.

Table 1 GFHT parameters used for the ethionine and synthetic *Arabidopsis* datasets

Dataset	Spectra	Model peaks[a]	PCs (K)	Variance explained by model (%)	α range	HIT size (**H**)[b]	Peaks aligned[c]
Ethionine	336	10	3	93.7	[−30, 30]	$65,536 \times 41 \times 21 \times 21$	839
Arabidopsis	24	10	2	98.7	[−100, 100]	$65,536 \times 51 \times 21$	356

[a] The number of manually selected peaks making up the MCSM

[b] The dimensionality of the HIT is $K+1$; the α_1 step size is, e.g., $(30-(-30))/41=1.46$ (ethionine)

[c] The number of detected maxima in the HIT

Calculating the GFHT

The definition of the Hough transform from the feature-detected data matrix **X** to the indicator tensor **H** is:

$$h_{k,l,m,\ldots} = f\left(\boldsymbol{\alpha}_{l,m,\ldots}, k\right) \tag{6–8}$$

$$f\left(\boldsymbol{\alpha}_{l,m,\ldots}, k\right) = \sum_i \sum_j x_{ij} \exp\left[-\frac{1}{2}\left(\frac{j-k-\boldsymbol{\alpha}_{l,m,\ldots} \bullet \mathbf{s}_{i\bullet}}{\sigma}\right)^2\right]$$

$$\boldsymbol{\alpha}_{l,m,\ldots} = \left[(\mathbf{a}_1)_l, (\mathbf{a}_2)_m, \ldots\right]$$

k is a position along the variable axis (ppm). $\mathbf{a}_1$, $\mathbf{a}_2$, etc. are vectors containing evenly spaced values of the Hough parameters for the principal components 1, 2, etc. $\mathbf{s}_{i\bullet}$ is row i of the shape matrix **S** that is the scores from the principal component analysis where each column is scaled to unit standard deviation. σ is a fuzzy parameter which is user-defined. $\sigma=2$ data points has been used throughout this work. An example: size(**X**)=($i\times j$), rank(MCSM)=K ($K=2$ as example) and we choose the following resolution on the alphas; length($\mathbf{a}_1$)=L, length($\mathbf{a}_2$)=M, we (have) get the following sizes **S** ($i\times K$), h ($1 \times 1 \times 1$), **H**=($j\times L\times M$), $\boldsymbol{\alpha}(1\times K)$ and k is traversed from 1:j; l is traversed from 1:L and m is traversed from 1:M.

Naïve partitioning and a new algorithm for calculating the GFHT score

First, consider the natural partitions of a typical ^{1}H-NMR dataset for metabolic profiling. There are often two or more groups of samples involved in these kinds of studies, e.g., one group dosed with a candidate drug and one control group or one group with a lesion and one healthy group. This partitioning can be exploited by separately calculating the GFHT with a local HIT for each of the sample groups and using the maximum value in the local HIT to update the global HIT. By dividing the HIT, we assign higher weights to peaks that are only present in a specific group. These peaks will now be detected and aligned although they are not present in all samples, i.e., these peaks will not be regarded as noise peaks. This is a useful feature when looking for biomarkers for a certain condition. If there is only one class label in the sample set, all spectra can be treated as a single class. The use of this partitioning can be seen in Fig. 6.

Equations 6–8 are the mathematically strict way of defining the GFHT transform, but in practice the GFHT can be calculated in an alternative way that is faster by using a peak list instead of the large feature-detected matrix **X**. This modification does not alter the solution. The complexity of the indicator tensor and consequently also the number of calculations grows exponentially with the number of shift patterns (K) and linearly with the chosen resolution of the parameters (α_i). Because of the discussed algorithm complexity issue, it is desirable that the efficiency of the algorithm improves. The improved algorithm is cast as follows:

1. Arrange your detected peaks for the whole dataset in a peak list; each peak should have the entries *sample*

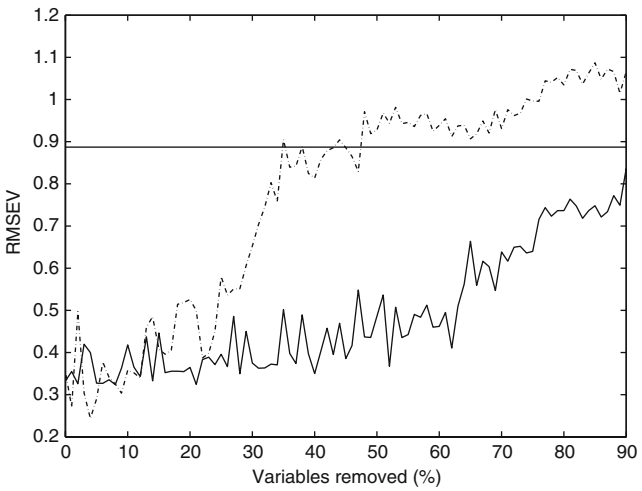

Fig. 3 Validation results for PLS1 models of the seven constituents with variable concentrations in the *Arabidopsis* dataset. Y is autoscaled. The dataset was divided into a calibration set comprising 19 samples and an external validation set comprising five samples; 7×91 PLS1 models were built for Bucketed (*dashed curve*) and GFHT (*solid curve*) data. RMSEV $= \frac{1}{n}\sqrt{\sum_n (y-y_p)^2}$, where n is the number of external validation samples ($n=5$); y is the true concentration and y_p is the predicted concentration of the external validation samples. The *horizontal line* represents the mean RMSEV error for random Y data

and *position on the frequency axis*. For the *Arabidopsis* dataset, this peak list has 3×78,029 entries indicating that 78,029 peaks were detected in the 24 samples.

2. Decide on the range and resolution of the α_i values.
3. Create a zero-filled local HIT (**L1**, **L2**,…) per sample class, each spanning K dimensions plus one dimension for the frequency (ppm) axis. If memory problems occur in this step, go back to step 2 and reduce the parameter resolution (decrease length($\mathbf{a}_i$)) or analyze the dataset in sections by dividing the frequency axis into segments. Note that the number of classes can be one.
4. For each sample class, calculate **L** as described below. For each permutation of the parameters **α**, do (a–d):
   (a) Create one vector ($\mathbf{h}_{\text{local}}$) with k elements (one element per data point on the frequency axis).
   (b) For every peak in the peak list that belongs to the current sample class, calculate the peak location **δ** on the frequency axis corrected for the current set of parameters, **α** (Eq. 5).
   (c) Add a normalized Gaussian to $\mathbf{h}_{\text{local}}$ centered on this corrected maximum **δ** (Eq. 7).
   (d.) Update the slice of the local HIT corresponding to the current **α**:
   $\mathbf{L}(\cdot, l, m,\ldots)=\mathbf{h}_{\text{local}}$.
5. Normalize each HIT by dividing each element by the number of samples in the corresponding class.
6. Calculate the final HIT (**H**) by taking element-wise maxima of **L**:
   $\mathbf{H}(k, l, m,\ldots)=\max(\mathbf{L1}(k, l, m,\ldots), \mathbf{L2}(k, l, m,\ldots),...)$

Analogous to the image analysis application of the Hough transform, each local maximum in the HIT indicates a possible peak with parameterized correspondence over the sample dimension which is equivalent to the concept that each local maximum describes a parameterized shape in an image. Depending on the quality of the initial spectral peak detection and the MCSM, there can also be some

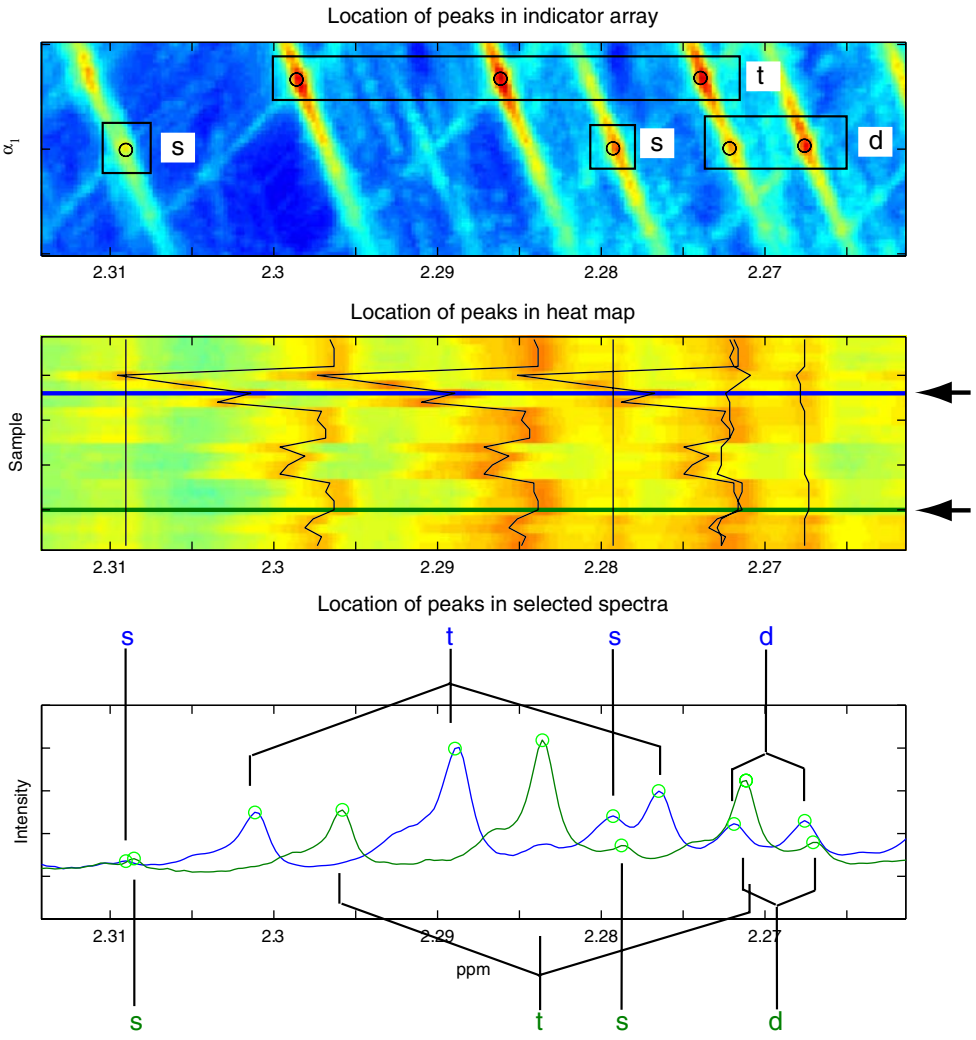

Fig. 4 Alignment results from a small segment of the *Arabidopsis* dataset. The *top panel* shows H (HIT) projected on α_1. Detected local maxima in H are marked (*white circles*). The *middle panel* shows a heat map of the raw spectra (*red* regions indicating high intensity and *blue* regions indicating low intensity), with overlaid predicted peak positions (*black lines*) equivalent to the Hough maxima for each of the 24 samples. The *bottom panel* shows two spectra (as indicated by *arrows*—the *blue* and *green horizontal lines* in the *middle panel*) where corresponding peaks are indicated (*s*—singlet, *d*—doublet, *t*—triplet)

false-positive maxima and some peaks that are missing a corresponding maximum. Since the HIT can have many dimensions, finding local maxima in the HIT is not a trivial task; in this work, we have manually given starting guesses for maxima locations of each peak by visual inspection of 2D projections of **H** (see Fig. 4, top panel) and then iteratively located the nearest local maximum. Other suggested methods can be found in, e.g., [18, 20].

Validation method

Since it is difficult to validate alignment of first-order data such as 1D-NMR data, we have opted for a data-driven approach: modeling capability and visual inspection. First, we acknowledge that the NMR data used is of (semi)quantitative nature, i.e., that the peak areas (or heights) are proportional to the concentration of analyte and that all peaks corresponding to one analyte (multiplicity) will covary linearly in a dataset where the concentration of analyte changes.

Equipped with a very controlled but real dataset such as the *Arabidopsis* set where all concentrations are known, all samples have a true internal standard and the samples are pH-controlled; we can test the following hypothesis: although we remove one or more of the peaks originating from one molecule, the remainder of the associated peaks should still reflect the concentration of that molecule. This hypothesis can be tested using a calibration model. A validation of the hypothesis that small peaks are consistently aligned can now be constructed as follows: (1) in the aligned (or bucketed) data, remove the largest peak (variable), (2) make a PLS model using the remaining data, (3) record the ability of the model to predict the concentration (RMSEV), (4) remove the second largest variable, etc. By examining the model error as a function of remaining variables, we can now draw conclusions about the quality of the remaining peak intensities and hence the alignment quality.

Results and discussion

The parameters used for the GFHT alignment of the ethionine and *Arabidopsis* datasets are provided in Table 1. A notable difference between the two datasets is that the ethionine dataset has a more complex shift pattern structure (K=3) than the *Arabidopsis* dataset (K=2); this is probably due to the samples in the latter dataset being titrated to constant pH.

Validation of the *Arabidopsis* alignment results

Using the variable removal where we consecutively remove variables from aligned and bucketed data while building PLS1 models, we can see, Fig. 3, that the bucketed data models starts to deteriorate when approximately 30% (75) of the largest variables are removed whereas for the aligned data the breakdown occurs when approximately 75% (260) of the variables are removed.

The difference between the breakdown rates between GFHT-aligned and bucketed data constitutes more than 350% difference in information retrieval between the methods. This does also indicate that the GFHT is capable of correctly assigning intersample peak correspondence for the *Arabidopsis* data.

Another useful feature from the GFHT is the possible support for peak annotation using the HIT. Figure 4, top panel, shows a window into 2D projected Hough scores in the HIT for different α_1 obtained from the alignment process. The location of one maximum reveals the value of the α_1 and when analyzing the second α-dimension (for the *Arabidopsis* data, K=2) we get α_2. By multiplying these alphas with their corresponding score vector in the MCSM, we can now predict the location of the peak in all samples.

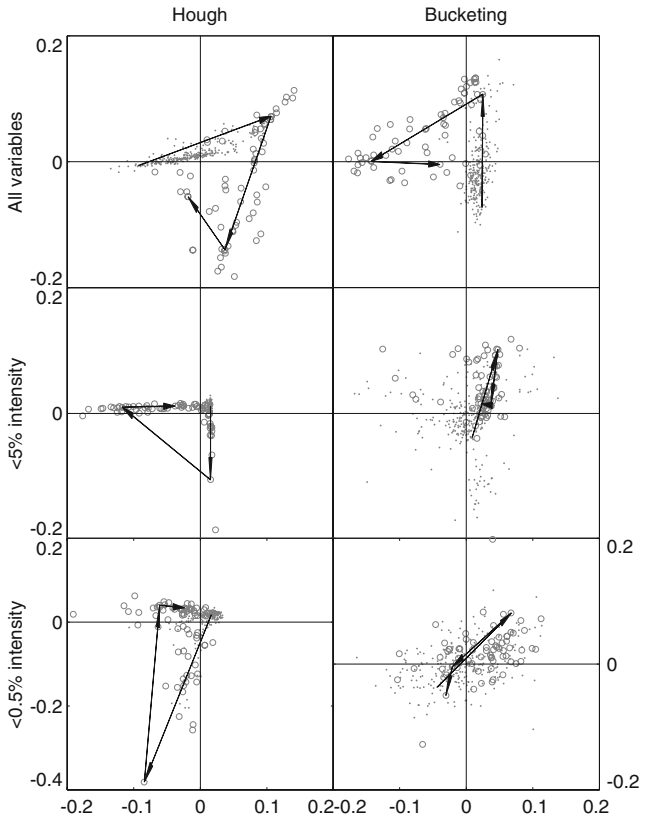

Fig. 5 PCA of the aligned ethionine data (PCs 1 and 2). Single high-dose samples (*circles*). All other samples (*dots*). The *arrows* indicate the same three samples in each pane. The samples with arrows are chosen to indicate the control samples (day −2), the effect of dosing (day 2), and the end of the experiment (day 7)

This is shown as the black lines in the middle panel in Fig. 4.

The top panel in Fig. 4 can also be chemically interpreted. Peaks originating from equivalent protons will have the same shift pattern and thus the same value of the α parameters. The similarity of the α parameter can be used to elucidate which peaks originate from equivalent protons, i.e., nearby Hough maxima with the same α are (likely to be) multiplets originating from one molecule. This is true even in cases where overlap makes manual assignment of multiplicity peaks difficult or impossible. This intersample proton peak correspondence feature can be viewed as pseudo-2D experiment data of proton coupling using mode support, i.e., by having Hough support over many samples, we can untangle the correspondence between multiplets in all the 1D spectra comprising the dataset. This feature of the GFHT can also improve quantification methods by using integrals (or intensities) from several corresponding peaks for quantification.

Results from the ethionine dataset alignment

Interpreting the results from the alignment of the ethionine dataset is not as straightforward as for the synthetic *Arabidopsis* dataset as the Y-block (class label) is not as well defined. Here, we have adopted a similar approach as used with the *Arabidopsis* data but settled for PCA models since PCA often is used for assessing metabolic trajectories. In this experiment, we have used the full set of GFHT-aligned and bucketed variables, two sets (GFHT and bucketed) where the variables are the ones with intensity of 5% of the maximum intensity (n=817, 183) and finally two sets where the intensity is less than 0.5% of maximum intensity (n=435, 142). In the score plots, Fig. 5, of these six models, we can see that indeed the scores patterns of the full models are similar. This is expected since both models are reflecting the most intense (varying) peaks—these should be represented in both the bucketed and GFHT-aligned data. At the 5%

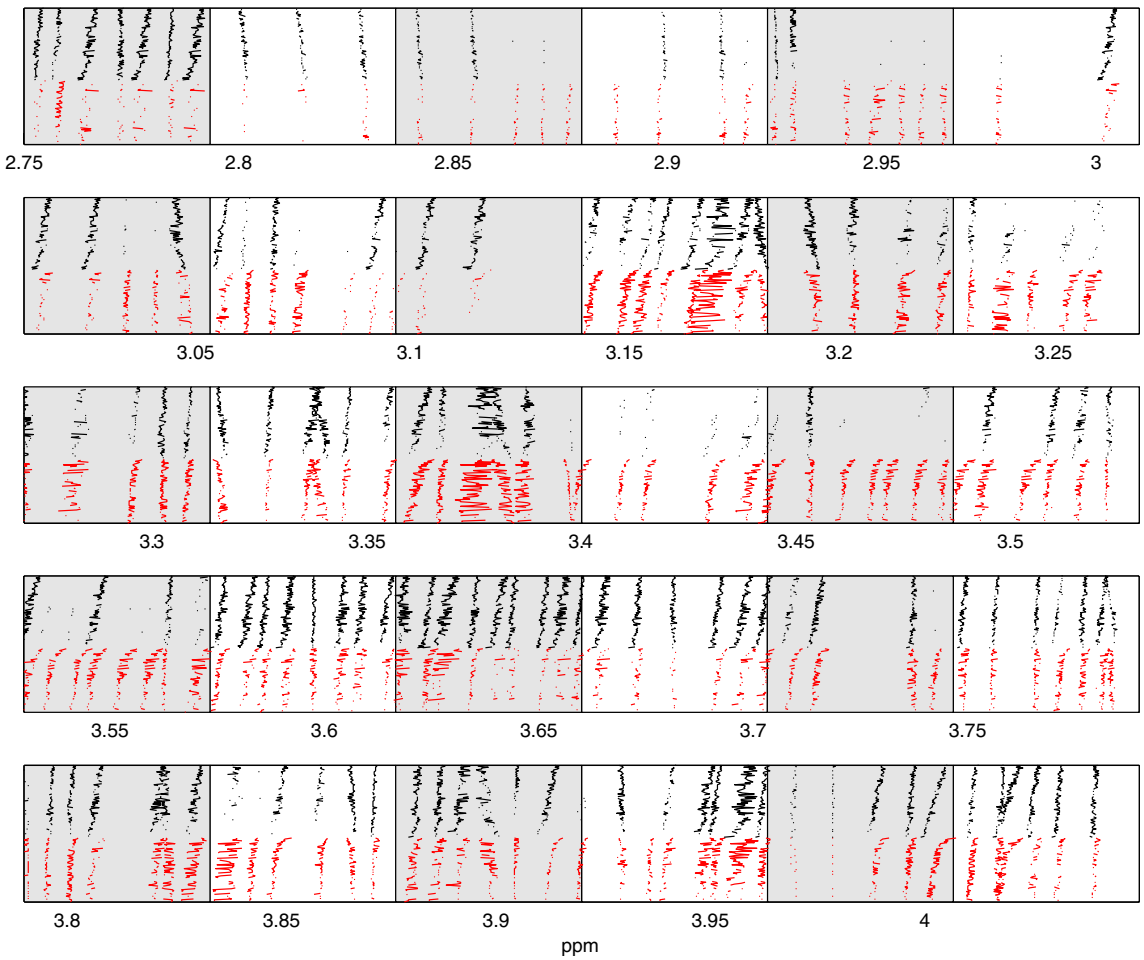

Fig. 6 A wider region of the ethionine data with GFHT-aligned peaks. Red peaks are assigned from the HIT partition assigned by the single high-dose samples. *Black peaks* are assigned from the rest of the samples. The alternating *gray*/*white fields* indicate typical buckets (0.04 ppm)

cutoff, we can see that the bucketed data still has some ability to separate the dosed time points whereas the GHFT-based model indicates *an even more refined model compared to the model of the full data*. The two phenomena seen in the full data are in the 5% GFHT model almost orthogonal and coinciding with the PC axes. Examining the 0.5% cutoff models, we can see that the bucketed data model has lost all separation power whereas the GHFT-based model still shows good separation. We interpret these models as that the GFHT is successfully aligning peaks in the more complicated ethionine data—even very small peaks. We can also see that the variability of the controls are increasing for the bucketed data as the magnitude of the variables is decreasing, indicating less interpretable PCA models for buckets with low intensities.

Using the information from the naive partitioning of the HIT, we also have an opportunity to label the peaks in the ethionine data according to which local HIT the maxima were found. In Fig. 6, the GFHT-aligned peaks are shown for a wider spectral segment. Here, the peaks which were class-labeled as the single high dose (and hence have a separate HIT) are plotted in red whereas peaks detected in the control set are plotted in black. The information this carries is that there are red peaks that are a consequence of the dosing event. This can also be seen in the figure—some of the red peaks are not present in control set. This information can be further used to either remove data from the dataset (obvious exogenous compounds) or to focus on peaks that are present in one partition but not the other (possible biomarkers). Figure 6 also indicates the extent of information loss when using bucketing—there are many peaks in each bucket for the ethionine data; the information each of these peaks carry is lost or confounded when bucketed.

Conclusions

The alignment results supported by the validation method demonstrate that the extended GFHT alignment method presented in this paper works for more complex samples such as urine.

The extended GFHT can effectively use the deterministic nature of peak shifts in ^{1}H-NMR data to construct a multicomponent shift model for the shifts of *all* detected peaks using mode support, i.e., using peak location and shift information from many samples. The Hough indicator tensor maxima location establishes the linear combinations of the MCSM (α_i) necessary to predict the location of *all* corresponding peaks in the analyzed dataset, hereby establishing intrasample peak correspondence.

The existence of a finite number of peak shift patterns holds true for the two datasets examined in this work and there is reason to believe it holds for any ^{1}H-NMR dataset which indicates that the peak shifts in ^{1}H-NMR data are deterministic (we have successfully analyzed several H-NMR datasets).

The extended GFHT hereby solves the correspondence problem for *any* dataset for which a multicomponent peak shift model with a finite number of parameters can be established.

We show that the HIT carries additional information to the intrasample peak locations, i.e., there is support for multiplet correspondence assignment in all 1D spectra analyzed.

We show that the partitioning of the HIT can be used to establish peak origin in time series or data with other known partitions.

We (implicitly) show that the postalignment information carried by low-intensity peaks is more readily available after alignment with GFHT, opening an opportunity for the field of metabolic profiling to establish more confidence about the generated data and to look further than the usual suspects when searching for biomarkers or biopatterns.

Lastly, we emphasize that the GFHT method presented, unlike many other methods, is symmetric, i.e., the order of which the spectra are analyzed does not influence the alignment results *and* that the GFHT is capable of aligning peaks which change order.

Acknowledgements We gratefully acknowledge AstraZeneca, Safety Assessment, Södertälje, Sweden, for cofunding the BioSysteMetrics Group at Stockholm University. We also gratefully acknowledge Ian Lewis, Department of Biochemistry, University of Wisconsin-Madison, for sharing the *Arabidopsis* dataset. The spectra of the synthetic *Arabidopsis* mixtures were collected at the National Magnetic Resonance Facility at Madison (NMRFAM; NIH grants P41 RR02301 and P41 GM GM66326). The authors would also like to thank one anonymous reviewer for numerous insightful and succinct suggestions for improving the paper.

References

1. Spraul M, Neidig P, Klauck U, Kessler P, Holmes E, Nicholson JK, Sweatman BC, Salman SR, Farrant RD, Rahr E, Beddel CR, Lindon JC (1994) J Pharm Biomed Anal 12:1215–1225
2. Holmes E, Nicholson JK, Nicholls AW, Lindon JC, Connor SC, Polley S, Connelly J (1998) Chemom Intell Lab Syst 44:245–255
3. Holmes E, Foxall PJD, Nicholson JK, Neild GH, Brown SM, Beddell CR, Sweatman BC, Rahr E, Lindon JC, Spraul M, Neidig P (1994) Anal Biochem 220:284–296
4. Bylund D, Danielsson R, Malmquist G, Markides KE (2002) J Chromatogr A 961:237–244
5. Pravdova V, Walczak B, Massart DL (2002) Anal Chim Acta 456:77–92
6. Wang CP, Isenhour TL (1987) Anal Chem 59:649–654

7. Wu W, Daszykowski M, Walczak B, Sweatman BC, Connor SC, Haselden JN, Crowther DJ, Gill RW, Lutz MW (2006) J Chem Inf Model 46:863–875
8. Tomasi G, van den Berg F, Andersson C (2004) J Chemom 18:231–241
9. Nielsen N-P V, Carstensen JM, Smedsgaard J (1998) J Chromatogr A 805:17–35
10. Kassidas A, MacGregor JF, Taylor PA (1998) Process Syst Eng 44:864–875
11. Tomasi G, van den Berg F (2006) DTW and COW, code for signal alignment by dynamic time warping and/or correlation optimized warping for mat lab. Dept. of Food Science, the Royal Veterinary And Agricultural University, Denmark. http://www.models.kvl.dk/source/
12. Torgrip RJO, Lindberg J, Linder M, Karlberg B, Jacobsson S, Kolmert J, Gustafsson I, Schuppe-Koistinen I (2006) Metabolomics 2:1–19
13. Åberg M, Torgrip RJO, Jacobsson SP (2005) J Chemom 19:1–9
14. Forshed J, Torgrip RJO, Åberg KM, Karlberg B, Lindberg J, Jacobsson SP (2005) J Pharm Biomed Anal 38:824–832
15. Csenki L, Alm E, Torgrip R, Åberg M, Nord L, Schuppe-Koistinen I, Lindberg J (2007) Anal Bioanal Chem 389:875–885
16. Duda RO, Hart PE (1972) J Commun ACM 15:11–15
17. Samal A, Edwards J (1997) Pattern Recogn Lett 18:473–480
18. Ballard DH (1980) Pattern Recogn 13:111–122
19. Pietrowcew A (2003) Opto-Electron Rev 11:247–251
20. Noriaki S, Eiji U, Kanae H (2006) Soft Comput 10:1161–1168
21. Han JH, Koczy L, Poston T (1994) Pattern Recogn Lett 15:649–658
22. Basak J, Pal SK (2005) Fuzzy Sets Syst 154:227–250
23. Jackson JE (1991) A users guide to principal components. Wiley, New York
24. Wold S, Esbensen K, Geladi P (1987) Chemom Intell Lab Syst 2:37–52
25. Lewis IA, Schommer SC, Hodis B, Robb KA, Tonelli M, Westler WM, Sussman MR, Markley JL (2007) Anal Chem 79:9385–9390
26. Savitzky A, Golay MJE (1964) Anal Chem 36:1627–1639