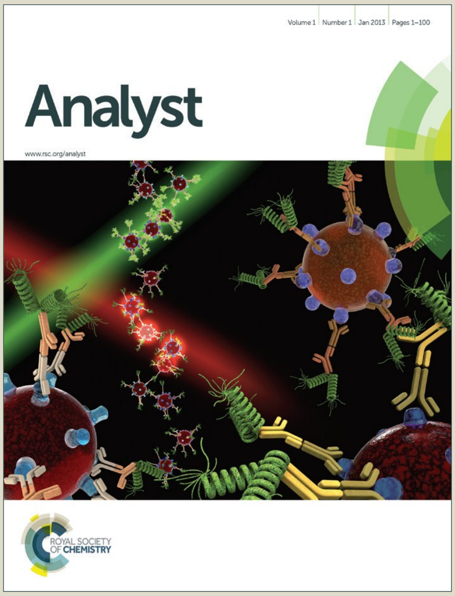
Analyst

Accepted Manuscript


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Analyst
www.rsc.org/analyst
ROYAL SOCIETY OF CHEMISTRY

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Please do not adjust margins

Analyst

ARTICLE

Multiscale Peak Detection in Wavelet Space

Zhi-Min Zhang[a,b] Xia Tong[a], Ying Peng[a], Pan Ma[a], Ming-Jin Zhang[a], Hong-Mei Lu[a,b, *], Xiao-Qing Chen[a, †], Yi-Zeng Liang[b]

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Accurate peak detection is essential for analyzing high-throughput dataset generated by analytical instruments. Derivative with noise reduction and matched filtration are frequently used, but they are sensitive to baseline variations, random noise and deviations in peak shape. Continuous wavelet transform (CWT)-based method is more practical and popular in this situation, which can increase accuracy and reliability by identifying peaks across scales in wavelet space and implicit removal of noise as well as baseline. However, its computational load is relatively high and the estimated features of peak may not be accurate in the case of peaks such as overlapping, dense and weak peaks. In this study, we present multi-scale peak detection (MSPD) for peak detection by taking full advantage of additional information in wavelet space including ridges, valleys, zero-crossings. It can achieve high accuracy by thresholding each detected peak with maximum of its ridge. It has been comprehensively evaluated with MALDI-TOF spectra in proteomics, CAMDA 2006 SELDI dataset as well as Romanian database of Raman spectra, which is particularly suitable for detecting peaks in high-throughput analytical signals. Receiver operating characteristic (ROC) curves show that MSPD can detect more true peaks while keeping false discovery rate lower than MassSpecWavelet and MALDIquant methods. Superior results in Raman spectra suggest that MSPD seems a more universal method for peak detection. MSPD has been designed and implemented efficiently in Python and cython. It is available as open source package at https://github.com/zmzhang/libPeak.

Introduction

Rapidly extracting quantitative information from high-throughput spectral profiles is crucial for providing insight of large, complex sample sets. Commonly, concentration of compound of scientific interest is often associated with height or area of peak(s) in profile of sample. Therefore, peak detection is a fundamental step in analysing dataset generated by various analytical instruments including chromatograph, mass spectrometry (MS), Raman spectrometer and Nuclear magnetic resonance spectroscopy (NMR). Therefore, it is a common requirement to detect peaks and calculate their positions, heights, widths and areas for signals of these analytical instruments. After obtaining large-scale organized data matrices such as peak area or height through peak detection across samples, further investigation for complex samples can be conducted with multivariate methods from statistics and chemometrics. However, there are random noises, alternating baselines, differing peak shapes, sample impurities, artefacts and overlapped peaks in real experimental signals. Due to adverse effect of these problems, it is complex and difficult to design an automatic and accurate peak detection method. Meanwhile, the acquired data become increasingly large with rapid advancements in analytical instruments. Therefore, automatic and accurate peak detection is still a significant challenge, particularly for high-throughput data processing in analytical chemistry.

Various peak detection methods have been developed, and most of them identify peaks by searching local maxima with SNR threshold to avoid false positives. However, simple derivative method (numerical differentiation) works poorly for signals with noise. In order to improve its poor result and recognize spectral features of Infrared (IR) spectra untouched by human hands, Savitzky and Golay provided sets of first- and second-derivative convoluting integers for noise reduction during derivative calculation. [1] Gaussian second derivative filtering has been presented by Danielsson, [2] which has the advantages of matched filters, derivatives and Gaussian function such as background subtracting, peak sharpening effects, SNR improvement and generating enhanced chromatograms for peak detection. The matched filtration with experimental noise determination (MEND) method can suppress chemical and random noise and baseline fluctuations, as well as filter out false peaks. [3] Recently, automatic chromatographic peak detection and background drift correction (ACPD-BDC) has been developed for chromatographic data analysis with robust noise estimation, first-order derivative and local curve-fitting. [4]

[a] College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, China.
[b] Institute of Chemometrics and Intelligent Instruments, Central South University, Changsha 410083, P.R. China.
* To whom correspondence should be addressed, Email: hongmeilu@csu.edu.cn.
† Additional correspondence author, Xiaoqin Chen, Email: xqchen@csu.edu.cn.

Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/x0xx00000x

Analyst Accepted Manuscript

However, height and width of peaks may vary a great deal in one signal. Therefore, fixed size matched filtering and derivative methods usually fail in this situation. Wavelet can exploit the multiscale nature of the measured signal. Lange applied CWT to matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectra using Marr wavelet with different dilation values. [5] By transforming the spectrum into wavelet space with Mexican hat wavelet, 2D CWT coefficients can provide additional information for identifying and separating the signal from noise and baseline. [6,7] Nguyen applied zero-crossing lines in multi-scale of Gaussian derivative wavelet for peak detection. [8,9] Haar wavelet has been introduced for both position identify and width estimation in alignDE [10] , MSPA [11] and CAMS [12] for chromatography dataset.

Peak detection methods based on above mentioned criteria have been applied in analysing various datasets, mainly including chromatography (particularly LC-MS), mass spectrometry (particularly MALDI-TOF and SELDI) and Raman spectroscopy. Second-derivative Gaussian filter has been used for de-noising and computing derivative of LC-MS dataset, and the filtered chromatogram will cross the x-axis roughly at the peak inflection points for feature extraction. [13] The centWave algorithm detects feature of LC-MS dataset with CWT and Gauss-fitting in the collected regions of interest, and it has been integrated into the XCMS package because of the higher recall and precision over the original matchedFilter of XCMS. [14] MZmine software has been developed by Orešič's group for mining LC-MS dataset, and recursive threshold peak detection method can reduce the false positives by avoiding detection of noise peaks. [15] Wavelet transform algorithm with Mexican hat wavelet has been integrated into MZmine 2 because of it is particularly suitable for noisy data. [16] Marc Sturm [17] presented the open source software framework OpenMS for rapid application development in mass spectrometry with an efficient peak picking algorithm [5]. eMZed [18] provides an flexible framework to process LC-MS data based XCMS, OpenMS and Python programming language. Peak detector based on CWT has also been applied to peak-to-peak matching for identifying unknown Raman spectra from reference spectral library. [19–21]

In summary, derivative with noise reduction, matched filtration and CWT are three major and popular peak detection methods for signals of analytical instruments. Meanwhile, one can observe that peak detection methods are of extremely important for analysing datasets of various analytical instruments. We can also see another significant trend that traditional peak detection methods have been substituted by CWT-based methods gradually because of its accuracy, performance and multiscale nature. However, peak detection by CWT is still less than satisfactory, especially for overlapping, dense and weak peaks. New approach is urgent in need to ensure high-quality peak detection, quantification, alignment across profiles for subsequent multivariate analysis.

In this study, we propose multiscale peak detection (MSPD) to achieve satisfactory results for signals of various analytical instruments. It is based upon CWT because of the multiscale, noise reduction and baseline removal properties. Therefore, it has all the advantages of CWT-based method. By taking full advantage of ridges, valleys and zero-crossings in wavelet space, MSPD can estimate features of peaks more accurate than MassSpecWavelet which is only using ridges in wavelet space. Furthermore, MassSpecWavelet is implemented in R programming language, which is not efficient enough. In order to process high-throughput and hyphenated datasets in metabolomics and proteomics efficiently, our approach has been designed and implemented in Python and cython, and it is significantly faster than MassSpecWavelet. It is particularly suitable for analysing high-throughput datasets.

This paper is organized as following. Firstly, relevant algorithm concepts of MSPD method are presented and investigated in theory section. Then MSPD method is applied to MALDI-TOF spectra in proteomics, CAMDA 2006 SELDI dataset as well as Romanian database of Raman spectra to demonstrate its accuracy and effectiveness. Results of above applications will be presented accompanying with discussions about the proposed algorithm. Finally, some conclusions and perspectives are given in conclusion section.

Theory

In this section, CWT will be introduced firstly as well as how to select the most suitable wavelet for peak detection. With CWT and wavelet, signal can be transformed into wavelet space. Then, strategy for locating ridges, valleys and zero-crossings in wavelet space will be elucidated as clear as possible. The position of peak can be detected accurately using its ridge, valley and zero-crossing information. The maximum of ridge has been used to threshold and eliminate false positive peak in peak-dense region. Finally, features of peak will be estimated from analytical signals for further statistical and chemometric analysis.

Wavelet

Wavelet used by CWT can be defined by analytical expressions, and dozens of wavelets have been invented for various applications in chemistry [22–24], including baseline correction [25–27], noise filtering [28–31], peak detection [6,7,10,12,32] derivative calculation [33–35] as well as compression [36–39]. Signal of analytical instrument has highly localized features; for instance, Gaussian peaks of chromatograms and Lorentzian peaks of Raman spectra. Therefore, wavelet should be purposefully chosen with specific properties. Therefore, it can extract information from analytical signal more effectively. Due to the features of Gaussian and Lorentzian peaks, the selected wavelet should have the basic features of these peak including approximate symmetry and one major positive peak. Mexican hat wavelet (Ricker wavelet) has been chosen in this study, which can be defined by following expression:

Analyst Accepted Manuscript

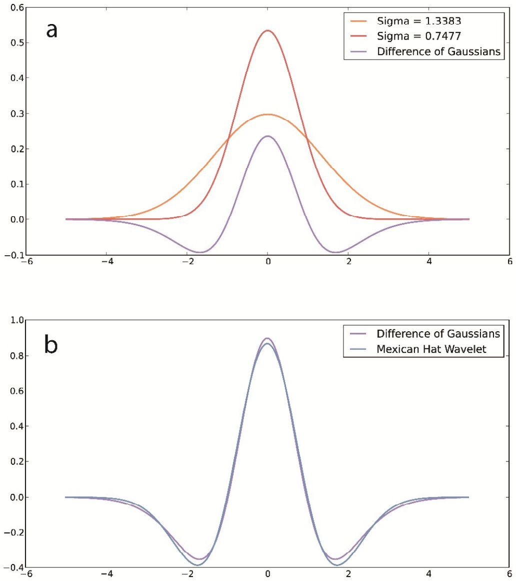


Figure. 1. Selection of Wavelet. (a) difference of Gaussians; (b) similarity between difference of Gaussians and Mexican hat wavelet.

$$\psi(t) = \frac{2}{\sqrt{3\delta}\pi^{\frac{1}{4}}}(1 - \frac{t^2}{\delta^2})e^{\frac{-t^2}{2\delta^2}} \qquad (1)$$

In practice, Mexican hat wavelet can be approximated by the difference of Gaussians (DoG). DoG is equivalent to a band-pass filter, which is believed to mimic how neural processing in the retina of the eye extracts details from images. [40,41] In figure 1, one can see the comparison of DoG with Mexican hat wavelet. In order to fit DoG best to Mexican hat wavelet, standard deviations of Gaussian functions and ratio constant between DoG and Mexican hat wavelet are optimized by least square fitting. One can observe from figure 1(a) that DoG is generated by Gaussian function with σ=0.7477 and σ=1.3383. Then, DoG is multiplied by 3.8128 and plotted with Mexican hat wavelet together in figure 1(b), and the root mean square error (RMSE) between them is only 0.0189. DoG for feature extraction has biological interpretation, and it has been applied extensively in computer vision. The scale-invariant feature transform (SIFT) [42] based on DoG is one of the most popular algorithms to detect and describe local features in images. The Mexican hat wavelet can be approximated by DoG. Therefore, Mexican hat wavelet is a reasonable choice as the mother wavelet for peak detection.

Ridge, Valley and Zero-Crossing in Wavelet Space

MassSpecWavelet by Du identifies ridges in 2D CWT coefficient matrix for peak detection. With the additional information from ridges, it can obtain more accurate and robust results than methods directly detecting peaks in raw signal. According to their results, this additional information can increase the accuracy of peak detection. However, only the ridges in wavelet space have been adopted in MassSpecWavelet package. Besides the ridges, there are also valleys and zero-crossings in the wavelet space, which are also critical for the accuracy of peak detection. Therefore, the ridges, valleys and zero-crossings should be fully used to extract important features (position and width) of peaks accurately and effectively. The concepts of ridge, valley and zero-crossing will be elucidated in the following paragraphs.

Mathematically, ridges of a function *f* of *N* variables are a set of curves whose points are local maxima in *N-1* dimensions. In this respect, the notion of ridge extends the concept of local maximum.

In this study, the 2D CWT coefficients can be regarded as a function of the dilation and translation parameters. According to its definition, ridge in wavelet space can be defined as curve of local maxima in the dilation dimension. Correspondingly, the notion of valley can be defined by replacing the condition of a local maximum with local minimum. The zero-crossing can be also defined similarly by replacing the condition of a local maximum with the condition of the sign changes (e.g. from positive to negative or vice versa).

Technically, these important features including ridges, valleys and zero-crossings can be located in wavelet space easily because of there are only two dimensions in the wavelet space. The rows and columns of CWT coefficients (denoted by **C**) are dilation and translation respectively. One can take columns *k, k+1, …, n+(k-1), n+k* and *–k, -(k-1),…, n-k* from matrix **C** as new matrix $^{k}\mathbf{C}$ and $^{-k}\mathbf{C}$ respectively. Indices that are too large are replaced by the last column of **C**, and too small by the first column. The ridges, valleys and zero-crossings can be located in the Boolean matrices generated by following Boolean algebra equations:

$$\begin{aligned}
\mathbf{R} &= (\mathbf{C} > {}^{-k}\mathbf{C}) \wedge (\mathbf{C} > {}^{-(k-1)}\mathbf{C}) \cdots (\mathbf{C} > {}^{k-1}\mathbf{C}) \wedge (\mathbf{C} > {}^{k}\mathbf{C}) \\
\mathbf{V} &= (\mathbf{C} < {}^{-k}\mathbf{C}) \wedge (\mathbf{C} < {}^{-(k-1)}\mathbf{C}) \cdots (\mathbf{C} < {}^{k-1}\mathbf{C}) \wedge (\mathbf{C} < {}^{k}\mathbf{C}) \\
\mathbf{Z} &= \left[\operatorname{sgn}(\mathbf{c}_1) \oplus \operatorname{sgn}(\mathbf{c}_0), \quad \cdots, \quad \operatorname{sgn}(\mathbf{c}_n) \oplus \operatorname{sgn}(\mathbf{c}_{n-1})\right]
\end{aligned} \qquad (2)$$

Where $\wedge$ and $\oplus$ are AND and XOR Boolean operations respectively. The *sgn* is a function, which can extract the sign of a real number. $\mathbf{c}_i$ is the ith column of matrix **C**. **R**, **V** and **Z** are matrices containing the ridges, valleys and zero-crossings information respectively.

Analyst Accepted Manuscript

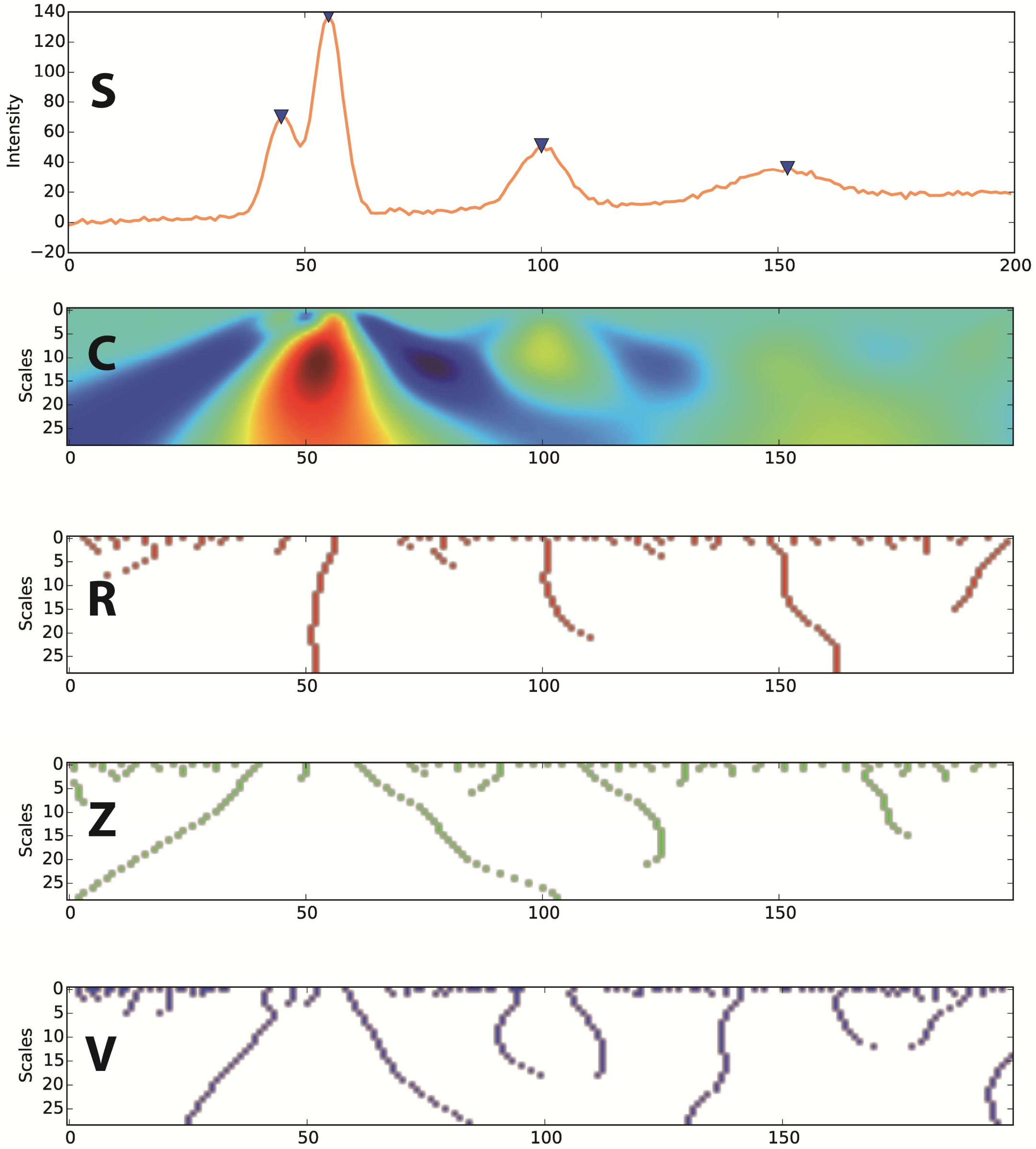


Figure. 2 CWT coefficients, ridges, valleys and zero-crossings of simulated signal with noise and baseline. (S) signal with noise and baseline; (C) CWT coefficients; (R) ridges; (Z) zero-crossings; (V) valleys.

Here is a simple example: consider a signal with four Gaussian peaks (one overlapped peak and two separated peaks) as well as noise and baseline (figure 2-S). By transforming the signal into wavelet space by CWT (figure 2-C), ridges (figure 2-R), valleys (figure 2-V) and zero-crossings (figure 2-Z) can be located in wavelet space with equation (2)~~(2)~~ and illustrated as images of different colours. One can observe from figure 2 that CWT procedure can generate wavelet coefficient **C**, which can provide additional information for peak detection in multiscale manner. And it also resists to noise and baseline of the signal. The ridges are accurate estimations of the peak positions, and the zero-crossing and valleys are extremely useful for locating the width, start and end points of the peaks. Please refer the *local_extreme*, *ridge_detection*, *peaks_position* and other functions of libPeak package at github for the details on computing and utilizing ridges, valleys and zero-crossings.

Analyst Accepted Manuscript

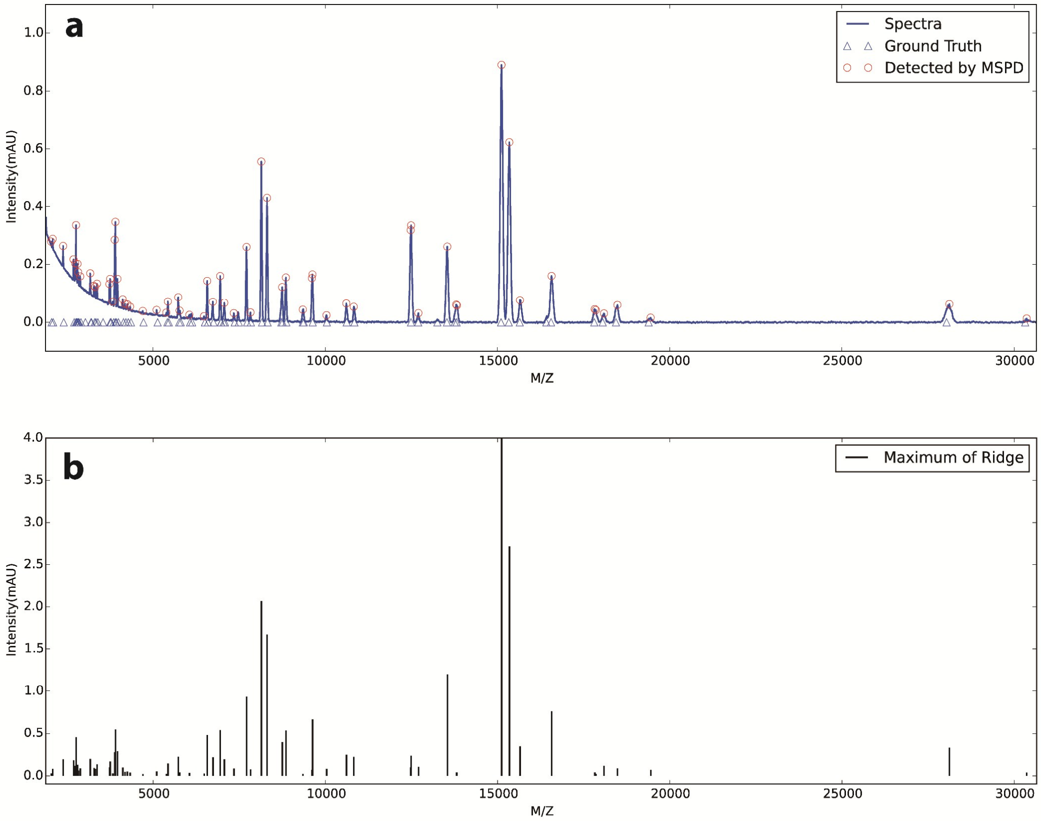


Figure. 3 Visualizing of the maximum of ridge and detection result by MSPD on MALDI-TOF proteomics spectra.

Peak Position Estimation

The ridges, valleys and zero-crossings are hidden in the matrices **R**, **V** and **Z** respectively. We should extract them from the matrices for further peak detection. However, the peak of analytical signals is not ideally Gaussian, and the other peaks affect wavelet coefficients when the dilation parameters are large. Consequently, the ridges of real signal are not a straight line in wavelet space. Therefore, an effective method should be developed to extract ridges from matrix **R**.

Firstly, several rows of **R** have been taken out to compute the sum of elements in each column. If the sum of one column is zero, there is no ridge at this column. If the sum of one column is larger than zero, there might be a ridge at this column. For each column with possible ridge, the first non-zero element at this column is chosen as the initial position. Then we scan matrix **R** from the initial position by increasing the row value from the small dilation to large dilation. Let's assume the current position of current element of ridge is i^{th} row and j^{th} column. For the next possible element of one ridge, we should consider the (i+1, j-1), (i+1, j) and (i+1, j+1) elements in matrix **R**. If any value of these elements is one, position of the element should be appended into the ridge. This procedure is repeated until reaching the end row of matrix **R** or the values of the next possible elements are all zero, and the ridge at this column has been extracted successfully. There may exist duplicated ridges in the ridges list, and the ridges will be merged if they have the same start element, end element and length.

The ridge line is not straight especially when the dilation parameter is large, so it can be only regarded as a rough estimation of position of peak. In order to estimate the accurate position of each peak, one should take full advantage of existing information including its ridge, valley, zero-crossing, wavelet coefficients as well as original signal. Firstly, peaks can be divided into two categories according to their wavelet coefficients. If some wavelet coefficients are larger than zero, this peak is a normal peak. If all wavelet coefficients are smaller than zero, this peak may be a small peak overlapped with a large peak. For a normal peak, the optimal ridge element are chosen, whose column appears most often in the ridge with wavelet coefficients larger than zero. Assume the optimal ridge element is (i,j), search the matrices **V** and **Z** with (i, j) as the starting point along the i^{th} row bi-directionally. When meeting nonzero value, save its column. Then we have two column m and n, and they are start and end point of one peak respectively. The column of the maximum value of the original signal between m^{th} and n^{th} columns is chosen as position of the peak. For overlapped small peak, the minimum and maximum

Analyst Accepted Manuscript

columns of the first half of its ridge are used for estimating position of the peak. The position is determined as the column of the maximum value of the original signal between minimum and maximum columns. The start and end points of this peak should be calculated and optimized by a deconvolution procedure.

Quantitative information including peak height and area should be obtained for further data analysis. In previous steps, position and width of the peak has been estimated. Therefore, both the height and area can be calculated with them easily. The intensity at peak position is a good estimation of peak height, and the trapezoidal numerical integration between start and end points of the peak is a good estimation of peak area. The accuracy of height and area are seriously influenced by baseline, and it should be fitted and corrected from the original signal with a proper baseline correction method [43–49].

Thresholding by Maximum of Ridge

Analytical dataset basically consist of signal of target compound(s), baseline and random noise. When removing false peaks, one needs to consider the issue of baseline and noise. Since the wavelet function is symmetric and satisfies zero mean, the slowly changing and locally monotonic baseline will be automatically removed during calculating CWT coefficients. CWT can be regarded as convolution of the signal with dilated and translated wavelets. Therefore, noises will be suppressed to a certain extent in wavelet space. Intensity of a peak in wavelet space can be defined as the maximum CWT coefficient on the ridge line within a certain scale range according to MassSpecWavelet [6]. In this study, we use the maximum CWT coefficient on the ridge line to remove false peaks, which can avoid the issue of baseline and noise. MALDI-TOF spectra have been selected to illustrate the advantages of thresholding by maximum of ridge. The circles in figure 3(a) are peak detection results from MSPD method, and the triangles are the ground-truth peaks. The good matching between detection result and ground-truth means that MSPD has good performance in peak detection. Each vertical line represents maximum of ridge of the peak. One can observe from figure 3(b) that maximum of ridge is proportional to height or area of the peak. It is also robust to noise and baseline. Therefore, it can be used as threshold to remove false peaks.

Experimental

Simulated MALDI–TOF Spectra

Mass spectrometry profiling combining with bioinformatic tools is a promising solution to identify novel disease-associated biomarkers, and peak detection algorithms are essential in the analysis pipeline.

Morris has developed a physics-based computer model of mass spectrometry to generate virtual MALDI–TOF spectra for method development and comparison, where the truth is known about what peaks are in each spectrum. [50] They provide a publicly available dataset, which can be downloaded at http://bioinformatics.mdanderson.org/. In this study, this dataset has been used to test the performance of MSPD method on MALDI-TOF spectra.

SELDI Spectra of CAMDA 2006

The international conference for the Critical Assessment of Massive Data Analysis (CAMDA) offer a forum for the researchers from computer science, statistics, molecular biology, and other areas to exchange ideas, and critical evaluation of various techniques of analysing massive dataset generated by instruments. In CAMDA 2006, there is a SELDI proteomics dataset, which is a real dataset of all–in-1 protein standard (Ciphergen Cat. # C100–007). There are seven polypeptides which create seven true peaks at 7034, 12230, 16951, 29023, 46671, 66433 and 147300 of the m/z values respectively. This dataset can be used to benchmark MSPD on real SELDI dataset by comparing the detected peaks with these true peaks.

Romanian Database of Raman Spectroscopy

Raman spectroscopy can be regarded as a “fingerprint” technique for compounds and mixtures identification non-invasively under ambient conditions without special sample preparation. This structural-rich information are often represented as peaks in Rama spectra, and peak detection is important for extracting information hidden in Raman spectra composed of thousands points. In this study, spectra from Romanian Database of Raman Spectra (RDRS) have been used to evaluate and validate MSPD method. It contains Raman spectra of mineral species with the description of crystal structure, sample image, origin of the sample, vibrations, which can be downloaded from http://rdrs.uaic.ro/. Peaks of each spectrum have been annotated and interpreted manually by experts in Raman spectroscopy. Therefore, it can be used to evaluate the peak detection method by comparing the manually annotated position and detected position by algorithms.

Results and Discussion

Comparison analyses were conducted to evaluate the performance of MSPD algorithm. The results mainly focus on the false positives, false negatives and reproducibility. The criteria for selection packages for comparison are widely used and freely available. The MassSpecWavelet is one of the most classical and popular peak detection method based on continuous wavelet transform. The MALDIquant is a recently proposed package, which was published at 2012 in bioinformatics. Therefore, MassSpecWavelet and MALDIquant were chosen finally for comparison in this study. Both of them were downloaded from their official site and ran locally.

Evaluation Criteria

The ground truth is known for MALDI-TOF, SELDI and Raman spectra. Therefore, a detected peak can be labelled as a false peak if its position is not within the given error range of the ground truth. With these detection results, false discovery rate (FDR) and sensitivity (true positive rate, TPR) can be calculated to measure the performance of algorithms:

Analyst Accepted Manuscript

Please do not adjust margins

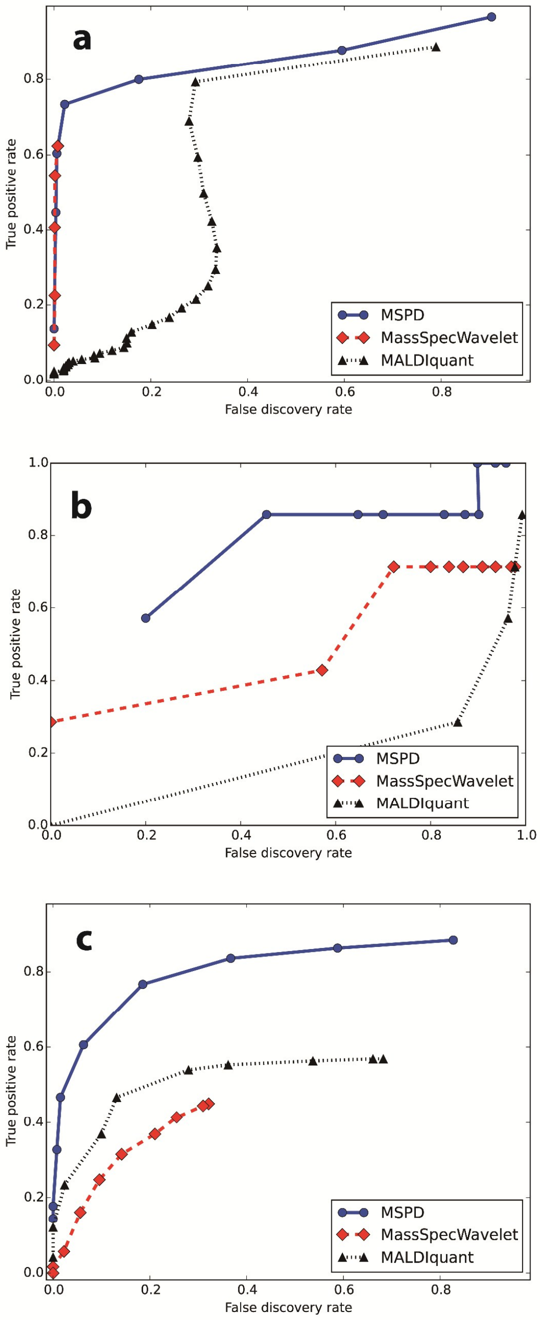


Figure. 4 ROC curves of three methods (MSPD, MassSpecWavelet and MALDIquant). (a) Average ROC curves of simulated proteomics dataset; (b) ROC curves of the 19th spectra of CAMDA 2006 SELDI dataset; (c) Average ROC curves of Romanian database of Raman spectroscopy.

$$TPR=\frac{TP}{TP+FN}=\frac{TP}{P} \qquad (3)$$
$$FDR=\frac{FP}{FP+TP}$$

TP is number of detected peaks within the ground truth. *FN* is number of peaks in the ground truth but not detected by algorithms. *P* is the total number of ground true peaks. *FP* is number of falsely detected peaks, which is not in the ground truth. For peak detection methods with the same *FDR*, one with larger *TPR* has better performance. The thresholding values of methods can be adjusted gradually to calculate a series of *TPR* and *FDR*. The ROC curves of different peak detection methods can be obtained by plotting the *TPR* against *FDR* at these thresholding settings, which is an informative measure for evaluation of different peak detection methods.

Peak detection results and comparisons

The MSPD is developed in Python language. But MassSpecWavelet and MALDIquant are developed in R language. In order to unify the code-base, rpy2 package is used as a communication interface between Python and R language. With the assistance from rpy2, peak detection functions of both MassSpecWavelet and MALDIquant can be accessed from Python easily.

The simulated MALDI-TOF spectra proposed by Morris are generated specifically for method development and comparison. There are hundreds of mean spectrum samples with hundreds of proteomics datasets in this data. In this study, 32nd dataset with 100 MALDI-TOF spectra have been chosen, and analysed by MSPD, MassSpecWavelet and MALDIquant. The performances of above methods have been evaluated by the ROC curve. Firstly, TPR and FDR of three methods are calculated for 100 simulated MALDI-TOF spectra. Then SNR parameters are adjusted gradually to create the ROC curves for MSPD, MassSpecWavelet and MALDIquant respectively. The SNR values are chosen from 0 to 12 for MassSpecWavelet method and 0 to 80 for MALDIquant method respectively. The threshold values of MSPD are chosen from 0.001 to 1, and spectra have been smoothed slightly by Whittaker smoother and normalized by its maximum before peak detection for MSPD method. The ROC curves have been plotted in figure 4(a). One can observe from figure 4(a) that FDR of MassSpecWavelet is limited to a small range because of its robust. But its TPR is significantly smaller than MSPD at the same FDR. TPR of MALDIquant varies greatly when FDR increasing, which means that peak detection of MALDIquant is not stable enough. TPR of MALDIquant is also significantly smaller than MSPD at the same FDR. It is clear that MSPD can achieve better performance than MassSpecWavelet and MALDIquant at all FDR. Utilizing ridge, zero-crossing and valley made significant contributions to accuracy and robust of MSPD method. Furthermore, our approach has been designed and implemented efficiently in cython, which is significantly faster than MassSpecWavelet.

Field Cod

Analyst Accepted Manuscript

Analyst Accepted Manuscript

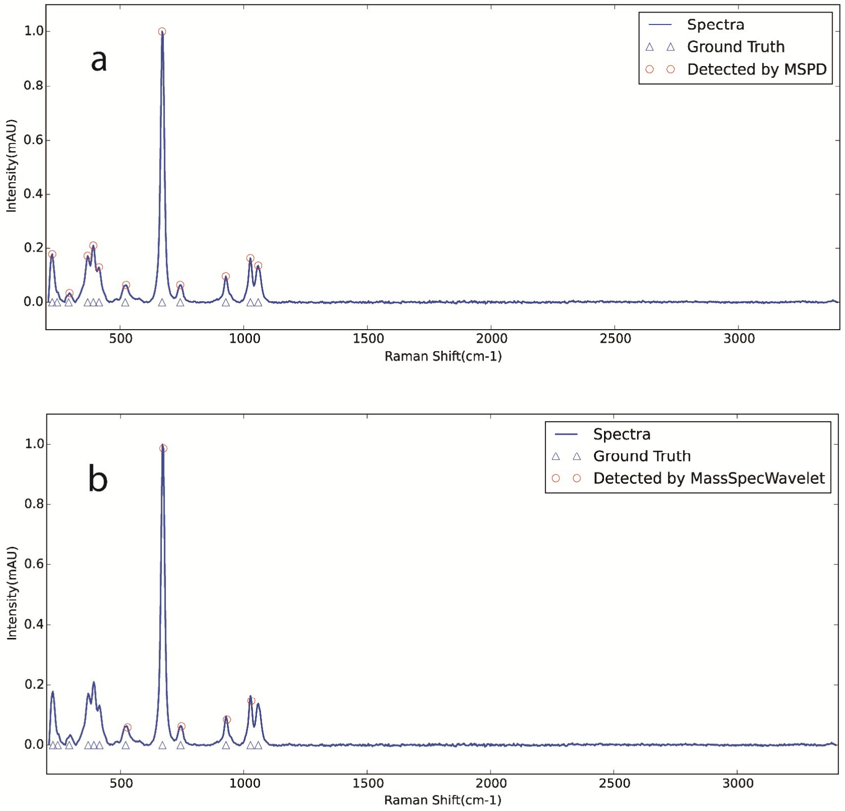


Figure 5. Advantage of MSPD on overlapped peaks when comparing to MassSpecWavelet

The 19th sample of CAMDA 2006 from GDWavelet package has been used to illustrate the performance of MSPD on SELDI spectra. The same performance test has been applied to this dataset. The SNR values have been varied from 0.5 to 80 for MassSpecWavelet method and from 0 to 80 for MALDIquant method respectively. For MSPD method, each spectrum has been normalized by its maximum and smoothed by Whittaker smoother. Then the threshold values are chosen from 0.001 to 0.5. The ROC curves of this dataset can be seen from figure 4(b). TPR of MSPD is significantly smaller than MassSpecWavelet and MALDIquant at every given FDR. It means that MSPD can also obtain better results than MassSpecWavelet and MALDIquant with the real SELDI dataset.

Raman spectra from RDRS are more challenging for peak detection algorithms because of random noise, fluorescent baseline, overlapped peaks and peaks dense regions. There are 66 Raman spectra, and peaks of each spectrum have been annotated by experts of Raman spectroscopy. The SNR values have been varied from 0 to 20 for MassSpecWavelet method and from 0 to 80 for MALDIquant method respectively. The threshold values are chosen from 0.001 to 0.5 for MSPD method. TPR and FDR of each peak detection methods can be calculated at these different SNR values, and ROC curves of RDRS dataset has been obtained by plotting TPR against FDR. TPR of MSPD is significantly larger than MassSpecWavelet and MALDIquant (figure 4(c)). MSPD is more stable than MassSpecWavelet and MALDIquant especially FDR is small, and this means that MSPD can identify more true peaks while keeping FDR low. The advantages of MSPD should owe to the ridge, valley and zero-crossing information in the wavelet space as well as thresholding by maximum of ridge. The superior results of MSPD in Raman spectra suggest that it is a more universal method for peak detection than its two competitors.

Performance on overlapped peaks

Peak detection in overlapped peak is more important than the separated peaks. In MALDI-TOF spectra, most the peaks are separated, and MassSpecWavelet can handle this kinds of dataset well. However, when there are overlapped peaks in dataset, the performance of MassSpecWavelet drops rapidly. MSPD can still achieve better performance on this kinds of dataset. For example, there are a lot overlapped peaks in the Raman spectra (RDRS dataset), and the performance (ROC curves) of MSPD is much better than the MassSpecWavelet and MALDIquant. This is one obvious advantage of MSPD, and we have illustrated this advantage in figure 5. One can see from figure 5 that there are two overlapped peaks at around 400 cm-1 and 1100 cm-1 respectively. Figure 5(a) is the peak detection result of MSPD, and it can detect each peak in the overlapped peaks. However, MassSpecWavelet can't find any peak in the overlapped peaks, and it can be observed from figure 5(b). By taking full advantages of ridges, zero-crossings and valleys, MSPD can locate the position of each peak in the overlapped peaks.

Multiscale advantage

In 2D matrix of wavelet coefficients, each row is the results of convolution between wavelet of different dilation parameter *a* and raw signal, and the dilation parameter *a* is increased gradually. The weak peaks can be identified in wavelet coefficients with small dilation parameter, and the strong peaks can be also identified in wavelet coefficients with large dilation parameter. Therefore, MSPD method can identify both strong and weak peaks at high sensitivity while keeping the FDR low.

Resistant to Baseline and Noise

Through transforming analytical signal into the wavelet space, baseline and noise can be suppressed effectively. Smoothing and baseline-correction steps aren't required to identify the peak position. It is the pivotal step of MSPD and MassSpecWavelet methods.

According to zero mean and square norm one requirements of wavelet function, convolution of the wavelet function ψ and the constant background *C* of above equation is zero. Mexican Hat wavelet is symmetric function, and the changing background is slow and monotonic. Therefore, convolution of ψ and the slow changing and monotonic background *B* of the equation is also approximately zero. In summary, both constant and slow changing baseline will be automatically removed during calculating procedure CWT.

When noises of signal are zero mean and independent identically distributed, it can be reduced by averaging nearby data points. Averaging and weighted averaging can be applied to signal by convolving signal with weighted function. CWT can be regarded as convolution of the signal with dilated and translated wavelets. Therefore, noises will be suppressed to a certain extent in wavelet space.

Conclusions

In this work, we present MSPD, an accurate and practical peak detection method for analytical signal by utilizing ridges, valleys and zero-crossings information in the wavelet space as well as thresholding by maximum of ridge. Features of each peak can be located and calculated precisely with its ridge, valley and zero-crossing in multi-scale wavelet space. Thresholding by maximum of ridge is an effective method to eliminate false peaks, which can avoid the issue of baseline and noise. MSPD is implemented in Python and cython, which is provided as an open source package. Performance tests on MALDI-TOF spectra in proteomics, CAMDA 2006 SELDI dataset as well as Romanian database of Raman spectra proved that MSPD method has much better performance in TPR and FDR than MassSpecWavelet and MALDIquant especially there are overlapped peaks in the dataset. Therefore, MSPD can achieve high TPR while keeping FDR low for a wide range of analytical signal, from MALDI-TOF spectra to Raman spectra, which makes MSPD suitable for extracting features of scientific interest from large, complex sample sets analysed by high-throughput analytical instrument. In future, MSPD will be applied to detect peaks in LC-MS dataset from metabolomics to proteomics.

Acknowledgements

This work is financially supported by the National Natural Science Foundation of China (Grants No. 21075138, Grants No. 21175157, Grants No. 21275164, Grants No. 21375151 and Grants No. 21305163), China Hunan Provincial science and technology department (Grants No. 2012FJ4139), Hunan Provincial Natural Science Foundation of China (Grants No. 14JJ3031), China Postdoctoral Science Foundation (No. 2014M552146) and National Instrumentation Program of China (No. 2011YQ03012407). The studies meet with the approval of the university's review board. We are grateful to all employees of this institute for their encouragement and support of this research.

Notes and references

1 A. Savitzky and M. J. E. Golay, *Anal. Chem.*, 1964, **36**, 1627–1639.
2 R. Danielsson, D. Bylund and K. E. Markides, *Analytica Chimica Acta*, 2002, **454**, 167–184.
3 V. P. Andreev, T. Rejtar, H.-S. Chen, E. V. Moskovets, A. R. Ivanov and B. L. Karger, *Anal. Chem.*, 2003, **75**, 6314–6326.
4 Y.-J. Yu, Q.-L. Xia, S. Wang, B. Wang, F.-W. Xie, X.-B. Zhang, Y.-M. Ma and H.-L. Wu, *Journal of Chromatography A*, 2014, **1359**, 262–270.
5 E. Lange, C. Gropl, K. Reinert, O. Kohlbacher and A. Hildebrandt, *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 2006, 243–54.
6 P. Du, W. A. Kibbe and S. M. Lin, *Bioinformatics*, 2006, **22**, 2059–2065.
7 P. Du, S. M. Lin, W. A. Kibbe and H. Wang, in *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, 2007. BIBE 2007*, 2007, pp. 680–686.
8 N. Nguyen, H. Huang, S. Oraintara and A. Vo, *Journal of Bioinformatics and Computational Biology*, 2009, **7**, 547–569.
9 N. Nguyen, H. Huang, S. Oraintara and A. Vo, *Bioinformatics*, 2010, **26**, i659–i665.
10 Z.-M. Zhang, S. Chen and Y.-Z. Liang, *Talanta*, 2011, **83**, 1108–1117.
11 Z.-M. Zhang, Y.-Z. Liang, H.-M. Lu, B.-B. Tan, X.-N. Xu and M. Ferro, *Journal of Chromatography A*, 2012, **1223**, 93–106.
12 Y.-B. Zheng, Z.-M. Zhang, Y.-Z. Liang, D.-J. Zhan, J.-H. Huang, Y.-H. Yun and H.-L. Xie, *Journal of Chromatography A*, 2013, **1286**, 175–182.
13 C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan and G. Siuzdak, *Anal. Chem.*, 2006, **78**, 779–787.
14 R. Tautenhahn, C. Böttcher and S. Neumann, *BMC Bioinformatics*, 2008, **9**, 504.

Analyst Accepted Manuscript

15 M. Katajamaa, J. Miettinen and M. Orešič, *Bioinformatics*, 2006, **22**, 634–636.
16 T. Pluskal, S. Castillo, A. Villar-Briones and M. Orešič, *BMC Bioinformatics*, 2010, **11**, 395.
17 M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert and O. Kohlbacher, *BMC Bioinformatics*, 2008, **9**, 163.
18 P. Kiefer, U. Schmitt and J. A. Vorholt, *Bioinformatics*, 2013, **29**, 963–964.
19 Z.-M. Zhang, X.-Q. Chen, H.-M. Lu, Y.-Z. Liang, W. Fan, D. Xu, J. Zhou, F. Ye and Z.-Y. Yang, *Chemometrics and Intelligent Laboratory Systems*, 2014, **137**, 10–20.
20 J. Cheng-zhi, S. Qiang, L. Ying, L. Jing-qiu, A. Yan and L. Bing, *Spectrosc. Spectr. Anal.*, 2014, **34**, 103–107.
21 G. Cooper, M. Kubik and K. Kubik, *Chemometrics and Intelligent Laboratory Systems*, 2011, **107**, 65–68.
22 B. Walczak and D. L. Massart, *TrAC Trends in Analytical Chemistry*, 1997, **16**, 451–463.
23 F. Ehrentreich, *Anal Bioanal Chem*, 2001, **372**, 115–121.
24 X.-G. Shao, A. K.-M. Leung and F.-T. Chau, *Acc. Chem. Res.*, 2003, **36**, 276–283.
25 Z. Pan, X.-G. Shao, H. Zhong, W. Liu, H. Wang and M. Zhang, *Chinese Journal of Analytical Chemistry*, 1996, 149–153.
26 H.-W. Tan and S. D. Brown, *J. Chemometrics*, 2002, **16**, 228–240.
27 L. Shao and P. R. Griffiths, *Environ. Sci. Technol.*, 2007, **41**, 7054–7059.
28 C. R. Mittermayr, S. G. Nikolov, H. Hutter and M. Grasserbauer, *Chemometrics and Intelligent Laboratory Systems*, 1996, **34**, 187–202.
29 B. K. Alsberg, A. M. Woodward, M. K. Winson, J. Rowland and D. B. Kell, *The Analyst*, 1997, **122**, 645–652.
30 B. Walczak and D. L. Massart, *Chemometrics and Intelligent Laboratory Systems*, 1997, **36**, 81–94.
31 X. Shao and W. Cai, *J. Chemometrics*, 1998, **12**, 85–93.
32 K. R. Coombes, S. Tsavachidis, J. S. Morris, K. A. Baggerly, M.-C. Hung and H. M. Kuerer, *Proteomics*, 2005, **5**, 4107–4117.
33 A. K. Leung, F. Chau and J. Gao, *Anal. Chem.*, 1998, **70**, 5222–5229.
34 X. Shao and C. Ma, *Chemometrics and Intelligent Laboratory Systems*, 2003, **69**, 157–165.
35 X. Zhang and J. Jin, *Electroanalysis*, 2004, **16**, 1514–1520.
36 A. Kai-man Leung, F. Chau, J. Gao and T. Shih, *Chemometrics and Intelligent Laboratory Systems*, 1998, **43**, 69–88.
37 J. Trygg and S. Wold, *Chemometrics and Intelligent Laboratory Systems*, 1998, **42**, 209–220.
38 U. Depczynski, K. Jetter, K. Molt and A. Niemöller, *Chemometrics and Intelligent Laboratory Systems*, 1999, **49**, 151–161.
39 L. Cao, P. de B. Harrington and C. Liu, *Anal. Chem.*, 2004, **76**, 2859–2868.
40 R. A. Young, *Spatial Vision*, 1987, **2**, 273–293.
41 M. J. McMahon, O. S. Packer and D. M. Dacey, *J. Neurosci.*, 2004, **24**, 3736–3745.
42 D. G. Lowe, *International Journal of Computer Vision*, 2004, **60**, 91–110.
43 Z.-M. Zhang, S. Chen, Y.-Z. Liang, Z.-X. Liu, Q.-M. Zhang, L.-X. Ding, F. Ye and H. Zhou, *Journal of Raman Spectroscopy*, 2010, **41**, 659–669.
44 S. Chen, X.-N. Li, Y.-Z. Liang, Z.-M. Zhang, Z.-X. Liu, Q.-M. Zhang, L.-X. Ding and F. Ye, *Spectroscopy and Spectral Analysis*, 2010, **30**, 2157–2160.
45 Z.-M. Zhang, S. Chen and Y.-Z. Liang, *Analyst*, 2010, **135**, 1138–1146.
46 Z.-M. Zhang and Y.-Z. Liang, *Chromatographia*, 2012, **75**, 313–314.
47 Z. Li, D.-J. Zhan, J.-J. Wang, J. Huang, Q.-S. Xu, Z.-M. Zhang, Y.-B. Zheng, Y.-Z. Liang and H. Wang, *Analyst*, 2013, **138**, 4483–4492.
48 X. Liu, Z. Zhang, Y. Liang, P. F. M. Sousa, Y. Yun and L. Yu, *Chemometrics and Intelligent Laboratory Systems*, 2014, **139**, 97–108.
49 X. Liu, Z. Zhang, P. F. M. Sousa, C. Chen, M. Ouyang, Y. Wei, Y. Liang, Y. Chen and C. Zhang, *Anal Bioanal Chem*, 2014, **406**, 1985–1998.
50 J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly and R. Kobayashi, *Bioinformatics*, 2005, **21**, 1764–1775.

Analyst Accepted Manuscript