Wavelet Analysis of High Performance Liquid Chromatography Data

Klapper, Jennifer H (1st author)
University of Leeds, Department of Statistics
Woodhouse Lane
Leeds (LS2 9JT), United Kingdom
E-mail: jennifer@maths.leeds.ac.uk

Barber, Stuart (2nd author)
University of Leeds, Department of Statistics
Woodhouse Lane
Leeds (LS2 9JT), United Kingdom
E-mail: stuart@maths.leeds.ac.uk

Chromatography and HPLC

Chromatography is a chemical process which is widely used as both a separative technique and an analytical tool. High performance liquid chromatography (HPLC) is a specific form of chromatography which is widely used in scientific fields including virology, pharmacology and clinical chemistry amongst many others. In its simplest form the separation is achieved by the injection of a solution of the chemical sample into a stream of solvent. The solvent is pumped into a chromatographic column which is packed with a solid separating material. Within the column a liquid-solid interaction takes place. Components with the highest affinity for the packing material stay in the column the longest and therefore running out, into the detector, last. The differential elution of the compounds is the underlying concept behind the separation which takes place in HPLC.

HPLC experimentation produces chromatograms, a type of spectral data. Chromatograms are a visual output showing the number of ions washing out into the detector as a function of time. The chromatograms display signals which are most commonly referred to as peaks. These peaks give qualitative and quantitative information about the sample under study. The qualitative information is the retention time of the component, which is constant under identical chromatographic conditions. Here we define the retention time to be the period between the injection of the sample and the recording of the signal maximum. Hence, sample constituents can be identified by the comparison of retention times. The quantitative information displayed relates to the area covered by a given peak, which when calculated is proportional to the amount of the corresponding substance present in the sample. One way in which this is done is by producing a calibration graph which is derived from peak areas obtained for various solutions whose concentration is precisely known. This is then used to compare the peak-areas and subsequently used to determine the concentration of an unknown sample (Meyer, 1988).

The nature of HPLC data

Generally, one of the main challenges in the analysis of HPLC data is that two experiments upon the same sample mixture can differ from each other. A full description of the problems which confound the analysis of chromatograms is given by Karpievitch et al (2010). Any observed differences are normally attributed to variations in experimental conditions and instrumentation, amongst other factors. One way in which to remove these variations is to carry out preprocessing of the data. These preprocessing steps are performed independently of any scientific information which one may wish to extract from the data and can therefore be problematic (Chen et al, 2007). Any inadequate or incorrect steps within the preprocessing of the data can result in data sets which display substantial bias, meaning that it is then difficult to reach any meaningful scientific conclusions (Coombes et al., 2005).

When attempting to analyse the raw spectrum, a commonly used model is composed of three components: the true signal, noise and baseline artifact. The model allows for signal reconstruction and consequentially biological or chemical interpretation in a mathematically principled manner (Chen et al., 2007). Within HPLC experimentation we usually define 'noise' to be instantaneously irreproducible signals which are typically caused by either physical or chemical interference in the experimental process, imperfections in the apparatus or many other irregularities. We aim to improve the signal to noise ratio of the raw data. The heuristic behind this denoising is that the removal of noise should facilitate the extraction of meaningful and accurate information from the data as background noise causes blurring of the analytical signal.

The baseline represents a type of artificial bias introduced by the machinery and keeps denoised data apart from the true distribution of the data. Baseline drift is mainly caused by continuous variation in the experimental conditions. Errors in the determination of the peak height and area of HPLC chromatograms are the main problem which baseline drift induces. Baselines are difficult to identify as they are usually represented by curves rather than linear functions (Chau et al, 2004). Hence, when we try to analyse chromatograms it is neccessary to not only denoise the data but to also correct for any baseline drift.

Wavelet analysis

Traditionally Fourier transforms played a major role in the analysis of this type of spectral data. However, recent developments in wavelet methods allow practitioners to decompose complex signals, including those produced by HPLC, into components at different scales. Fourier methods are unable to do this due to the lack of localization in the time domain. To counteract these problems we implement a wavelet based approach to HPLC data analysis; see, eg, Chau et al (2004).

As previously mentioned when dealing with HPLC data traces a certain amount of preprocessing must be undertaken. In additon to any denoising or baseline correction it is sometimes necessary to exclude the first part of the data trace as this area is typically dominated by noise. This leads to extreme variability and consequent unreliability. In addition during this interval saturation is typically observed, meaning that the number of ions arriving at the detector exceeds the number which are able to be detected (Coombes et al., 2005). Secondly, as we have elected to use a wavelet based approach to the analysis, we need the data to be of length 2^j for some integer j. Hence, it is usually necessary to pad-out the data by reflecting the ends of the data set so that it is of the correct length. This does not have any effect on our analysis. Additionally, when processing HPLC data is it to be expected that any information about the true peaks in the time domain will be represented by a small number of relatively large wavelet coefficients in the wavelet domain. Thus, in general, HPLC data satisfies the sparsity property of the wavelet transform. We therefore transform our data from the time domain into the wavelet domain. We threshold the resulting coefficients to remove any noise and then back transform into the time domain. We then adopt a gradient change points approach to baseline detection and removal. Figure 1 displays these preprocessing steps on an example data set.

Peak detection

Walls et al (2007) and Walls (2008) first implemented vaguelette-wavelet methods to estimate the derivate of the data traces. This derivative was then used to locate the start and end points of the peaks by first calculating a threshold $\lambda_1 = \bar{g} + k \cdot \sigma$. Here, $\bar{g}$ denotes the mean of the gradient vector, k denotes a user adjustable parameter and σ is the median absolute deviation (MAD) of the

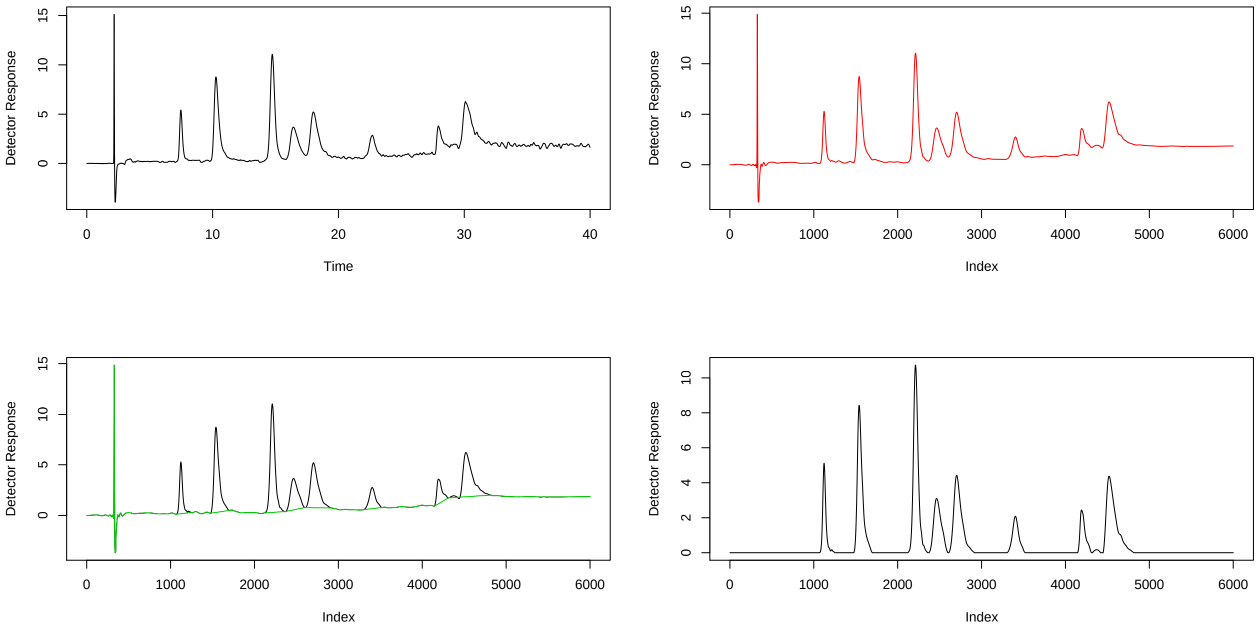

Figure 1: Top left panel: A example of a HPLC data trace. Top right panel: The denoised data trace. Bottom left panel: The denoised trace with the identified baseline highlighted. Bottom right panel: The final data trace with baseline removed.

gradient vector. When the derivative had exceeded λ_1 for a user definable window length a start time was defined. The window length parameter was originally arbitrarily set to be 30 clock ticks; further experimentation is required to find the optimal value of this parameter and it may be that the value of this parameter is truly dependent upon the data set under investigation. The end times were then found by using a similar method. Specifically once a start time had been identified, a new threshold of $\lambda_2 = \bar{g} - k \cdot \sigma$ was set and once the gradient had fallen below this threshold for another user definable window length an end point was defined as the first point in which the derivative once more rose above this threshold. This window length was originally set to be 20 clock ticks, again this parameter may be dependent on the data set at hand.

One of the findings of Walls (2008) was that whilst the accuracy and reliability of the method for detecting the start time was fair the same could not be said for the end time detection. Therefore we propose a new method for location of the peak end times and in doing so we include two user definable parameters. Specifically, we propose reversing the ordering of the data, so that the end points would become start points, to see whether this would remove the dicrepency between the start and end point estimation accuracy. We usually assume that the noise involved in HPLC experimentation is independent, which is why we use independent noise within our simulations. However Walls (2008) found evidence to suggest that the errors were not independent, instead concluding that they followed an autoregressive time series model. This was later attributed to problems within the experimental apparatus. It is therefore not thought of a problem commonplace to this type of experimentation. Generally if a process, $X(t)$, is a stationary Gaussian process then $X(t)$ is time reversible (Weiss, 1975).

As we are now reversing the time ordering of the data to detect the end points it would be possible to use the same threshold, λ_1, for the start and end time detection. However, as most peaks created by HPLC experimentation demonstrate a certain asymmetry, allowing the user to define the parameters individually for the start and end times is deemed preferable. Hence, we set a threshold of

$\lambda_S = \bar{g}+k_1 \cdot \sigma$ for start time detection and a threshold of $\lambda_E = \bar{g}+k_2 \cdot \sigma$ for end time detection. We then experiemented to find the optimal values of these parameters by comparing results from simulations to a known truth. We extended these simulations to different levels of noise to see whether this effects the optimal values of these parameters. These simulations showed that the values of k_1 and k_2 do in fact impact upon the accuracy of the start and stop time estimates, whilst also showing that the reversal of time vastly improved the accuracy of end time detection.

Peak area quantification

Generally, we aim to recover a signal $\boldsymbol{f}$ from a set of noisy observations $\boldsymbol{y}$. For VWD methods we suppose that we are unable to observe a vector $\boldsymbol{f}$ and instead wish to recover it from some function $g(\boldsymbol{f})$. In the specific case of integration we wish to recover the function

$$\boldsymbol{g} = \left[\int \mathrm{dt}\right] \boldsymbol{f} \approx K^{-1}\boldsymbol{f},$$

where K^{-1} represents an integration matrix. This type of problem is typically referred to as being 'ill-posed', as the somewhat naive estimate of $\boldsymbol{g}$ obtained from the inverse transform of K^{-1} applied to an estimate of $\boldsymbol{f}$ fails to produce reasonable results. Abramovich and Silverman (1998) detailed how two methods, first introduced by Donoho (1995), can be applied to these problems. They first describe a Wavelet-Vaguelette decomposition then, in response to limitations of this method, propose a Vaguelette-Wavelet decomposition.

Under the VWD approach we wish to find the wavelet coefficients of $K^{-1}\boldsymbol{f}$ as opposed to $\boldsymbol{f}$ by thresholding the wavelet coefficients of $\boldsymbol{g}$. Therefore the threshold levels usually applied need to be adjusted accordingly. Abramovich and Silverman (1998) proposed a threshold which is a factor of $\sqrt{1+2\alpha}$ higher than the standard universal threshold, $2\sqrt{\sigma \log n}$, giving the threshold $\lambda_{VWD} = \sigma\sqrt{2(1+2\alpha)\mathrm{log}n}$. Here n denotes the sample size and σ the standard deviation of the noise coefficients. We have experimented as to the optimality of the inflation factor in the case of integration and whether the further use of vaguelette-wavelet methods improve the accuracy of peak area estimation.

To further test the accuracy of our method we experiment to find the lower limit of detectability. This is a point of interest to practitioners who use HPLC as it is useful to know how small a peak can be and still be reliably detected. We simulate a peak and sequentially reduce the size each time adding the same amount of independent white noise. We set the noise level arbitrarily high as we are interested in testing the limits of the algorithm and lower noise levels, whilst being more realistic, would not allow for this.

REFERENCES (RÉFERENCES)

Abramovich, F. and Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems, *Biometrika* **85**(1): 115-129.

Chau, F.T., Liang, Y.Z., Gao, J. and Shao, X.G. (2004) *Chemometrics: From Basics to Wavelet Transform*, John Wiley & Sons Inc, New Jersey.

Chen, S., Hong, D. and Shyr, Y. (2007). Wavelet-based procedures for proteomic mass spectrometry data processing, *Computational Statistics and Data Analysis* **52**: 211-220.

Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M. C. and Kuerer, H. M. (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform, *Journal of Computational and Graphical Statistics* **5**: 4107-4117.

Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition, *Applied and Computational Harmonic Analysis* **2**(2): 101-126.

Karpievitch, Y.V., Polpitiya, A.D., Anderson, G.A., Smith, R.D. and Dabney, A.R. (2010). Liquid Chromatography Mass Spectrometry-Based Proteomics: Biological and Technological Aspects, *Annals of Applied*

Statistics **4**(4): 1797-1823.

Meyer, V. (1988). *Practical HPLC*, John Wiley and Sons Ltd, Chichester.

Walls, R.E., Barber, S., Gilthorpe, M.S. and Kent,J.T. (2007).Statistical analysis of high performance liquid chromatography data. *Systems Biology & Statistical Bioinformatics*, pp. 144. Edited by S. Barber, P.D. Baxter, & K.V. Mardia. Leeds, Leeds University Press.

Walls, R.E. (2008), *Statistical Analysis of Genomic and Proteomic Data*, PhD thesis, University of Leeds.

Weiss, G. (1975). Time-reversibility of linear stochastic processes, *Journal of Applied Probability* **12**(4): 831-836.