A multi-scale approach to analysing High Performance Liquid Chromatography data

Jennifer Klapper* and Stuart Barber

Department of Statistics, University of Leeds

Abstract

HPLC is used to identify the individual chemical constituents present in a sample. Each individual constituent creates a ‘peak’ in a data trace and from this the constituent can be identified and subsequently quantified. Earlier work in this area by Walls et al (2007) introduced a wavelet based approach to the analysis of HPLC data. We extend this work to improve the accuracy of peak location methods and to accurately capture asymmetric and overlapping peaks. We also assess how large a peak has to be to ensure consistent and reliable detection.

1 Chromatography and HPLC

Chromatography is a process of separation 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 sample, which has been dissolved in a solvent, into a stream of solvent. The solvent is pumped into the 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 elutes 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 which is a collection of sequentially recorded numbers of ions which arrive at the detector coupled with the corresponding mass-to-charge ratio. These data sets display what 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. Hence the sample constituent corresponding to each of under 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.
2 The nature of HPLC data

There are many problems which confound the analysis of chromatograms, a full description of these problems and other related ones is given in Karpievitch et al (2010). Variations in conditions and instrumentation as well as the presence of background noise and compound peaks, amongst other factors cause problems in analysis. Hence, when modeling the raw spectra a commonly used model is composed of three components: the true signal, noise and baseline artifact (Chen et al, 2007). Generally, within HPLC experimentation noise is defined to be ‘instantaneously irreproducible signals’ and are usually 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 being 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 function (Chau et al, 2004).

3 Multiscale 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 with different frequencies. 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. We continue the work of Walls (2007,2008) who first implemented vaguelette-wavelet methods to improve the accuracy of peak location and subsequent quantification. We propose a new method for location of the peak end times and in doing so we include two user definable parameters. We experiment to find the optimal values of these parameters by comparing results from simulations to a known truth. We extend these simulations to different levels of noise to see whether this effects the optimal values of these parameters.

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.

As mentioned earlier, a particular problem with chromatographic data is the presence of compound peaks. A compound peak occurs when two analytes have not fully separated and therefore have run off into the detector very closely together. This effect is mainly due to the analytes under observation being too chemically similar for the detector to fully differentiate between them. This causes particular problems in baseline detection as the data trace does not return to baseline before once more spiking, producing two peak which overlap. The gradient change points approach originally adopted by Walls (2008) interpolated between each peak end point and thereby lost vital information about the sample constituents under study. We have adapted this original approach to allow for the presence of compound peaks and present simulations to demonstrate the effectiveness of these new methods.
References


