Analytica Chimica Acta 817 (2014) 9–16

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Probabilistic peak detection for first-order chromatographic data

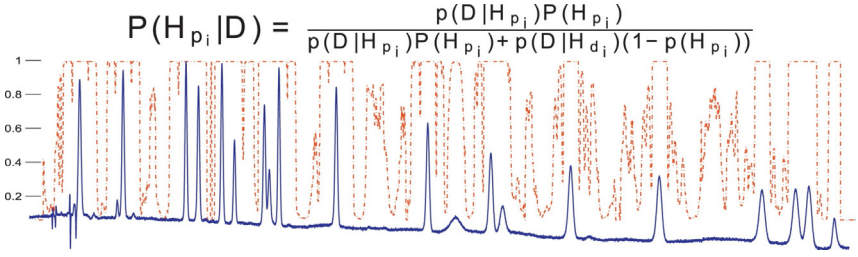
M. Lopatka[a,c,*], G. Vivó-Truyols[b], M.J. Sjerps[a,c]

[a] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Postbus 94248, 1090 GE Amsterdam, The Netherlands
[b] Van't Hoff Institute for Molecular Sciences, University of Amsterdam, Postbus 94248, 1090 GE Amsterdam, The Netherlands
[c] Netherlands Forensic Institute, Postbus 24044, 2490 AA Den Haag, The Netherlands

HIGHLIGHTS

- A novel algorithm for probabilistic peak detection in chromatography is proposed.
- The methodology follows a Bayesian inferential approach.
- Peak detection performance does not depend on the height of peaks.
- Probabilistic peak detection improves potential subsequent chemometric analysis.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 23 October 2013
Received in revised form 6 February 2014
Accepted 12 February 2014
Available online 14 February 2014

Keywords:
Peak detection
Bayesian statistics
Chromatography
Chemometrics

ABSTRACT

We present a novel algorithm for probabilistic peak detection in first-order chromatographic data. Unlike conventional methods that deliver a binary answer pertaining to the expected presence or absence of a chromatographic peak, our method calculates the probability of a point being affected by such a peak. The algorithm makes use of chromatographic information (i.e. the expected width of a single peak and the standard deviation of baseline noise). As prior information of the existence of a peak in a chromatographic run, we make use of the *statistical overlap theory*. We formulate an exhaustive set of mutually exclusive hypotheses concerning presence or absence of different peak configurations. These models are evaluated by fitting a segment of chromatographic data by least-squares. The evaluation of these competing hypotheses can be performed as a Bayesian inferential task. We outline the potential advantages of adopting this approach for peak detection and provide several examples of both improved performance and increased flexibility afforded by our approach.

1. Introduction

Accurate detection of compound-related signal peaks is critical to all subsequent analytical tasks and often defines the sensitivity of the chemical analysis. The detection of peaks constitutes a fundamental step whether one is calculating the concentration of a collection of compounds or using the peak areas for further inspection via multivariate methods. In this second case, multivariate analytical methods often operate over a set of peak areas associated with target compounds [1–4]. A requisite condition for these types of analyses is the preliminary detection, isolation, and integration of signal peaks associated with specific compounds of interest. In light of this paradigm, the importance of good peak detection is paramount to proper functioning of all multivariate analytical techniques performed over peak measurements.

Algorithmic approaches to peak detection have traditionally followed two strategies, detection by derivatives and by matched filter response [5]. Derivative-based peak detection methods make use of the fact that the first derivative of a peak will have a

https://github.com/mlopatka/getPeakConv

* Corresponding author at: Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Postbus 94248, 1090 GE Amsterdam, The Netherlands. Tel.: +31 205257038.
E-mail addresses: m.lopatka@uva.nl (M. Lopatka), g.vivotruyols@uva.nl (G. Vivó-Truyols), m.j.sjerps@uva.nl (M.J. Sjerps).

http://dx.doi.org/10.1016/j.aca.2014.02.015
0003-2670/

positive-to-negative zero-crossing at the local maxima of a peak [6]. To avoid false positives, a threshold on the slope is often imposed. By convention, this method smooths the first derivative of the signal prior to seeking zero-crossings with downward slope, after which only those zero-crossings whose slope exceeds a certain predetermined minimum are retained. Likewise the beginning and end point of a peak are often defined in terms of the zero-crossings in the second-order derivative relating to the same signal. Thus a central point, width, and fairly precise estimate of peak dimensions are accessible using derivative-based measurements [7]. Matched filtering is achieved by the application of a linear filter, which is designed to detect the presence of a particular pulse event with a known structure embedded in additive noise [8]. The filter response function typically exhibits high amplitude in locations corresponding to the presence of a pulse with a specified structure. When applied to chromatographic data [9,10], assuming a Gaussian peak shape, we may perform thresholding in the response function to determine the location of chromatographic peaks. Matched filter methods are becoming progressively more sophisticated [10–13] as data complexity increases, whereas derivative-based methods commonly require increasingly elaborate pre-processing [14,15] to prevent compounding noise effects [10].

Both of these approaches assert the imposition of a threshold to classify signal sections as either belonging to a peak or to the underlying noise present in the chromatographic signal. In other words, the algorithms deliver a binary decision indicating whether a given point is affected by the presence or absence of a chromatographic peak. In order to obtain satisfactory performance, users are required to tune a collection of parameters. The selection of an optimal threshold for different detection approaches has been thoroughly discussed [6,12,16–19], with no general consensus being reached. Threshold-based methods may yield sub-optimal results due to the general functional limitations inherent to a threshold operation. If a binary decision is implicit to the method, then errors in classification are unavoidable. For peak detection this means vital information may be lost. A probabilistic assertion does not necessarily discard low probability peaks, at least not in initial peak detection performed on a single chromatogram.

We propose an alternate data processing paradigm for probabilistic peak detection rather than a binary one. Unlike traditional methods, the peak presence probability is calculated over each point in a chromatographic signal using parameters that are directly related to the nature of the peaks expected in the chromatographic system. All functionality of threshold based methods can be achieved within a probabilistic context simply by imposing a threshold on the minimum acceptable posterior probability of observing a peak. Rather than employing an implicit binary decision, our method introduces a probability of peak influence for a specific point. This probability is calculated and potentially propagated through subsequent processing steps. Our method avoids prematurely eliminating candidate features in chromatographic data prior to multivariate analysis. Additionally, the application of Bayes' Rule for peak detection is complemented by the *statistical theory of component overlap* developed by Davis and Giddings [20]. We demonstrate that the method presented in this paper constitutes a novel way of updating the probability of whether a peak is occupying a chromatographic space. Starting with a collection of prior probabilities as derived by Davis and Giddings, we arrive at posterior probabilities via Bayes' Rule once the observation of chromatographic data is taken into account.

The idea of using Bayesian inference in analytical chemistry is not new [21]. The use of evidential reasoning has made an impact in domains such as genomics, proteomics, or metabolomics, which are largely dependent on analytical chemistry as an underlying source of data. These domains usually apply Bayesian methodology at the level of feature alignment or multivariate analysis

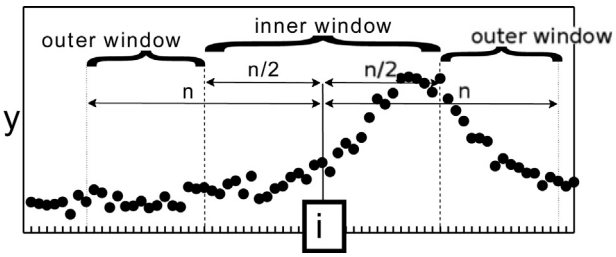


Fig. 1. Relevant window centered at point *i*.

[22,23,19]. The application of Bayesian statistics directly in the chromatographic field is more frequently limited to experimental calibration, signal alignment, compound concentration estimation, and deconvolution [24,25]. Existing research has also established the usefulness of a Bayesian inferential model for resolving overlapping peaks in a modulated signal [26]. The examination of post-fit residuals as a method of characterizing peak regions has also been used in multidimensional data [17] and for the estimation of the probability that a data point belongs to the baseline [27]. We present a direct implementation of Bayes' Rule to peak detection for a single first-order chromatogram.

2. Method

The fundamental task in peak detection is the identification of compound related peaks in chromatographic data, including their centers and tails, such that accurate qualitative and quantitative inferences may be drawn from the chromatographic analysis. To this end our method evaluates, for each point in the chromatogram, the possibility that such point is affected by a chromatographic peak. Operationally, the user is expected to provide two parameters with approximate accuracy. The first parameter, σ_{peak}, is the ideal width of a peak in the retention time domain for the specific chromatographic system. Using a static peak width assumes peak width stability over the course of the chromatographic run (as can be expected from gradient chromatography). This parameter is then converted to the dimensionless data point scale. It may be estimated from internal standard or calibration runs performed on the instrument. The second parameter, σ_ϵ, is the standard deviation of the baseline noise exhibited by a specific system. This parameter can easily be obtained from a blank chromatographic run.

An operational window is defined as described in Fig. 1. The width of the window is $8\sigma_{peak}+1$, and contains $2n+1$ points. With this width, a peak with a width defined by σ_{peak} can be contained with $n/2$ points on either side of a central point i, Hence n is approximately equal to $4\sigma_{peak}$.

We shall refer to this $n+1$ point window as our inner window, symmetrically centered around point i. An additional $n/2$ points before and after the inner window are referred to as the outer window. We include the outer window in order to include peak center positions outside the inner window. When a peak center occurs outside the outer window it can not affect points inside the inner window since the window size is determined by the peak width.

We aim to calculate the probability that the point i is affected by the presence of a chromatographic peak. Such a peak necessarily affects points in the inner window but may also influence points in the outer window. This operation is performed iteratively such that every point in the chromatogram occupies the i position before subsequently shifting the window one point in the retention time interval. We define the following competing hypotheses (in reference to point i) as per the conventional Bayesian inferential formulation [28,29].