Chemometrics and Intelligent Laboratory Systems 139 (2014) 26–32

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Stochastic regression modeling of chemical spectra

Anthony J. Kearsley [a], Yutheeka Gadhyan [b], William E. Wallace [c,*]

[a] Applied and Computational Mathematics Division, National Institute of Standards and Technology, Gaithersburg, MD, USA
[b] Department of Mathematics, University of Houston, Houston, TX, USA
[c] Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD, USA

ARTICLE INFO

Article history:
Received 18 December 2013
Received in revised form 28 July 2014
Accepted 1 August 2014
Available online 14 August 2014

Keywords:
Signal to noise
Stochastic differential equation
Nonparametric estimation
Mass spectrometry
Spectroscopy

ABSTRACT

A stochastic regression model is presented that separates signal from noise in chemical spectra. Spectra are decomposed into additive contributions from signal and from estimated noise. Numerical results on sample spectra are presented and suggest that this strategy offers an effective and computationally efficient framework for comprehensive noise estimation and analysis. From this analysis more effective methods of feature extraction in chemical spectra can be created.

1. Introduction

In every study employing chemical spectroscopy a point is reached where the analyst must reduce the original volume of spectroscopic data into specific, unambiguous observations. These are typically the existence, position, and intensity of any and all peaks in the collected spectra where *peak* is defined commonly as a “significant” increase in the measured intensity above a baseline value. The questions the analyst must answer are: Are there any peaks in the spectrum? If so, what is the best estimate of the position and relative size of each purported peak? Peak position is required when compound identification is the primary goal; peak intensity is required when quantitative information is the goal. Often pre-processing of the collected spectra is performed to reduce the size and complexity of the data set. One of the most commonly employed pre-processing steps is the reduction of noise where *noise* is defined ambiguously as that part of the spectrum that is not signal. Smoothing (e.g., running averages) and frequency domain filtering (e.g. Fourier, wavelet) are the most common of the many available noise reduction methods [1]. In addition to peak picking, the challenges of spectrum abscissa calibration, baseline correction, alignment of multiple spectra, and intensity normalization between spectra all benefit from separating signal from noise [2–4].

To a limited extent functional modeling of spectra has been studied previously. The work most similar in spirit to that presented here was performed by Coombes et al. [5] in 2005. They represented the raw observed signal as a three-term sum of the signal plus a slowly-varying background plus random noise. They develop a method to denoise spectra using an *undecimated discrete wavelet transform*, UDWT. Following denoising the baseline is removed by fitting a monotone local minimum curve and peaks detected after normalization. In a similar vein Morris et al. [6] developed functional mixed models using a Bayesian wavelet approach for the simultaneous modeling of multiple spectra with non-parametric representation of the fixed and random effects. House et al. [7] used functional modeling of the spectrum where each peak is modeled as a probability density function such that the signal is modeled as a sum of such probability density functions and the background is modeled as an exponentially decaying function. The model is estimated using Bayesian approach based on Lévy adaptive regression kernels.

In this paper a different approach is taken. Mass spectra are modeled using stochastic differential equations (SDE) [8–10] where both the drift (signal) and the diffusion (noise) coefficients depend on time. In this way the frequency and the intensity of the signal and of the noise are allowed to vary independently across the spectrum. The SDE coefficients are estimated using a nonparametric technique based on kernel regression. The benefit of the stochastic regression model is that it seeks to decompose the spectrum into signal plus estimated noise. Having done this peaks that rise above the noise may be easily identified and features that do not rise above the level of noise can be subjected to further scrutiny. The stochastic regression model does away with the need for smoothing or filtering and the difficulties analysts face in choosing the right smoothing method.

Stochastic regression modeling is intended to yield enhanced peak picking information by providing a method to eliminate peaks that

[*] Corresponding author.
E-mail address: william.wallace@nist.gov (W.E. Wallace).

http://dx.doi.org/10.1016/j.chemolab.2014.08.002
0169-7439/© 2014 Elsevier B.V. All rights reserved.

have the same frequency and intensity as the approximated noise. Stochastic regression modeling can also provide a simulation tool to generate large amounts of statistically meaningful spectra in very little time. Having an estimate of the noise, acquired from an unbiased operator-independent regression, allows peak characteristics (height, relative area, distance to the nearest peak, etc.) to be compared and contrasted rapidly with local noise features. Peaks with, for example, heights that are less than the intensity of the noise estimated in exactly the same region would generate suspicion and perhaps elimination. Simulation experiments could be run to ascertain coarse grained spectral information only apparent in larger data sets. The noise estimate can also be useful when tuning instrument parameters, for example, instrument settings that minimize noise may not be the same as those settings that maximize signal to noise. Clearly having an unbiased estimate of noise in the spectrum would be a useful tool for the analyst.

2. Time-dependent stochastic differential equation model

The dynamic behavior of the chemical spectrum is assumed to be described by the following time-dependent SDE:

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t, \tag{1}$$

where $\mu(t, X_t)$ and $\sigma(t, X_t)$ are the drift (signal) and diffusion (noise) coefficients respectively of the process X_t. Here, W_t is the standard Brownian motion with independent and normally-distributed increments where $E[dW_t] = 0$ and $Var[dW_t] = dt$. It can be seen that this model captures the behavior of the spectrum which models with constant coefficients fail to capture. To be able to work with a model that is estimable given one trajectory of the spectrometer the following subclass of the SDE (1) is considered:

$$dX_t = [a_0(t) + a_1(t)X_t]dt + b_0(t)X_t dW_t, \tag{2}$$

with the coefficients assumed to be twice continuously differentiable. Given the data X_{t_i} at discrete points $t_1 < t_2 < \ldots < t_{N+1}$ equally spaced in time, the coefficients in Eq. (2) are estimated using a nonparametric technique as proposed in [11] for the estimation of financial term structure dynamics (for a review see Fan [12]). This estimation method is described in the next section.

3. Estimation of the SDE coefficients

A first order *Euler–Maruyama* discretization of Eq. (2) is given by:

$$\Delta X_i = [a_0(i) + a_1(i)X_i]\Delta + b_0(i)X_i\Delta W_i, \tag{3}$$

where $\Delta = t_{i+1} - t_i$, $\Delta X_i = X_{t_{i+1}} - X_{t_i}$, $a_j(i) = a_j(t_i)$, $b_0(i) = b_0(t_i)$, and $\Delta W_i = W_{t_{i+1}} - W_{t_i}$. Then $\Delta W_i \sim N(0, \Delta)$. For data observed at very closely spaced time steps the Euler discretization is a good approximation of the continuous time model (see [9]). The drift term is estimated using local weighted regression and the diffusion term is estimated using the maximum likelihood principle.

3.1. Estimation of the drift coefficients

The linear regression of $\frac{\Delta X_i}{\Delta}$ over X_i and a constant leads to an equation of the following form:

$$\frac{\Delta X_i}{\Delta} = \alpha X_i + \beta + \varepsilon$$

where the coefficients α and β are obtained as minimizers of the corresponding least squares problem and ε is Gaussian error. Following [11] we use local kernel regression over $(t_0 - h, t_0 + h)$ at each point t_0 where h is called the bandwidth parameter. The drift coefficients are then approximated by constants $a_j(t) \sim a_j(t_0)$, $j = 0, 1$ in the small neighborhood determined by h around t_0. As in [13], the *Epanechnikov kernel*

$$\begin{aligned} K_h &= \frac{3}{4h}\left(1 - u^2\right), \quad -1 \le u < 0 \\ &= 0, \quad u < -1, \quad u >= 0 \end{aligned}$$

is used to assign weights so that points closer to t_0 are given more weight than points farther away. It also uses only those points which lie within the window of size h. The local weighted linear regression leads to the following quadratic minimization problem at each point i:

$$\min_{a_0(i), a_1(i)} \sum_{j=1}^{N} \left(\frac{\Delta X_j}{\Delta} - a_0(i) - a_1(i)X_j\right)^2 K_h\left(\frac{t_j - t_i}{h}\right), \tag{4}$$

where the approximation $a_j(t) = a_j(i)$, $j = 0, 1$ for $t \in [t_i - h, t_i)$ over the window of regression is valid. Setting the first derivative with respect to $a_0(i)$ and $a_1(i)$ equal to zero the following first order conditions are obtained:

$$\begin{aligned} &\sum_{j=1}^{N} \left(\frac{\Delta X_j}{\Delta} - a_0(i) - a_1(i)X_j\right) K_h\left(\frac{t_j - t_i}{h}\right) = 0 \\ &\sum_{j=1}^{N} \left(\frac{\Delta X_j}{\Delta} - a_0(i) - a_1(i)X_j\right) K_h\left(\frac{t_j - t_i}{h}\right) X_j = 0. \end{aligned}$$

The weighted least squares estimators of $a_j(i)$, $j = 0, 1$ and $i = 1, 2, \ldots, N$ are obtained as solution of the above equations:

$$\hat{a}_0(i) = \frac{\sum_{j=1}^{N} \left(Y_j - a_1(i)X_j\Delta\right) K_h}{\Delta \sum_{j=1}^{N} K_h} \tag{5}$$

$$\hat{a}_1(i) = \frac{\sum_{j=1}^{N} K_h \sum_{j=1}^{N} Y_j X_j K_h - \sum_{j=1}^{N} Y_j K_h \sum_{j=1}^{N} X_j K_h}{\Delta\left(\sum_{j=1}^{N} K_h \sum_{j=1}^{N} K_h X_j^2 - \left(\sum_{j=1}^{N} K_h X_j\right)^2\right)}, \tag{6}$$

where $Y_j = X_{j+1} - X_j$ and $K_h = K_h\left(\frac{t_j - t_i}{h}\right)$.

3.2. Estimation of the diffusion coefficients

Under the stochastic regression model presented here, the residual

$$\Delta X_i - [\hat{a}_0(i) + \hat{a}_1(i)X_i]\Delta$$

is modeled as

$$\Delta X_i - [\hat{a}_0(i) + \hat{a}_1(i)X_i]\Delta \approx b_0(i)X_i\Delta W_i \tag{7}$$

for all $i = 1, 2, \ldots, N$ where $\Delta Wi \sim N(0, \Delta)$. Let

$$\frac{\Delta X_i - [\hat{a}_0(i) + \hat{a}_1(i)X_i]\Delta}{\sqrt{\Delta}} =: \hat{E}_i. \tag{8}$$

Given information up to time t_i the following holds from the normality and independent increments property of the Brownian motion:

$$\hat{E}_i \sim N\left(0, (b_0(i)X_i)^2\right).$$

Then the conditional density of $\hat{E}_i$ given information up to time t_i is:

$$\left(2\pi(b_0(i)X_i)^2\right)^{-1/2} \exp\left(-\frac{\hat{E}_i^2}{2(b_0(i)X_i)^2}\right)$$