Chemometrics and Intelligent Laboratory Systems 130 (2014) 1–5

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

# Modeling and correction of Raman and Rayleigh scatter in fluorescence landscapes

Paul H.C. Eilers [a,*], Pieter M. Kroonenberg [b]

[a] Erasmus University Medical Center, Rotterdam, The Netherlands
[b] Department of Child & Family Studies, Leiden University, Leiden, The Netherlands

ARTICLE INFO

*Article history:*
Received 9 January 2013
Received in revised form 3 September 2013
Accepted 5 September 2013
Available online 17 September 2013

*Keywords:*
Smoothing
Background subtraction
Rayleigh scatter
Raman scatter
Fluorescence excitation–emission spectroscopy

ABSTRACT

Rayleigh and Raman scatter in fluorescence (excitation–emission) landscapes are a nuisance in two-way and three-way data modeling. We provide a method to clean individual emission spectra. The scatter can be represented accurately by Gaussian peaks, characterized by location, width and height. The analytic signal of interest effectively acts as a background to the scatter peaks. Modeling it locally as a smooth curve, using penalized least squares, allows accurate estimation of the parameters of scatter peaks. Once the peaks are modeled, they can be subtracted from the spectrum, almost completely removing the artifacts. Apart from local smoothness, no assumptions are made about the fluorescence spectra.



## 1. Introduction

One of the success stories of multi-way models is their application to chemical spectra [1,2]. Here we consider fluorescence excitation–emission spectroscopy in which the components, representing chemical species, are known to be bilinear. Unfortunately, Raman and Rayleigh scatter usually are present. They can be quite strong, ruining the good fit of a model to the data.

Several techniques have been proposed for handling scatter in two-way and especially three-way fluorescence data when the intention is to model the signal. Most available methods [3–5] attempt to eliminate the offending artifacts. In Bahram et al. [4] the following methods are mentioned: downweighting of the scatter region [6–8], subtraction of a standard [9], inserting missing values for the scatter [10], inserting zeroes outside the data area [8], avoiding the part of the data matrix which includes the scatter, applying constraints to the model for the signal [11,12], and modeling the scatter via two-way component models [8]. Their own proposal is to eliminate the scatter where signal and scatter overlap via interpolation, using the model to be fitted, thus intimately linking the handling of scatter to the fitting of the signal. In addition, Engelen et al. [3] proposed to treat each point of the scatter as an outlying observation with respect to the model and to eliminate it if it is too far away from the main body of the data. Attempts have also been made to indirectly model the scatter and the signal more or less independently using component models, especially within the context of three-way data [8]. Each of the mentioned papers gives further comments and references about the background to the problems and ways in which attempts have been made to solve the fitting of the signal in the presence of scatter. Note that all but the papers by McKnight et al. [9] and Zepp et al. [5] deal with modeling three-way data, indicating the need for appropriately handling scatter ridges in that context.

The proposal in the present paper attacks the core of the problem, in the sense that it defines explicit and well-fitting models for both the Raman and Rayleigh scatter, making relatively mild assumptions. In particular, the assumption is that the ridges in the landscape can be modeled with separate normal distributions along the excitation wavelength. With this approach it will also be possible to assess in great detail the nature of the scatter itself. Due to the regularity in the scatter, estimates of the fluorescence intensity can be made at the same time for those locations in the landscape where it overlaps with the signal. The result of our fitting procedure is that the scatter in the excitation–emission matrix landscape can be eliminated in a very precise manner, without compromising the signal itself. Moreover, the estimation is independent of the type of signal and how it is being modeled. Thus, we use the characterization that Data = Scatter + Signal + Noise, and our concern here is exclusively with the Scatter part.

Our procedure cleans individual emission spectra, so it is also useful outside the realm of multi-way data analysis.

## 2. Models for scatter

Figs. 1 and 2 show a typical excitation–emission fluorescence landscape. Rayleigh scatter (both primary and secondary), as well as Raman scatter are visible as diagonal ridges. The primary and secondary Rayleigh

* Corresponding author.
*E-mail addresses:* p.eilers@erasmusmc.nl (P.H.C. Eilers), kroonenb@fsw.leidenuniv.nl (P.M. Kroonenberg).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.09.002

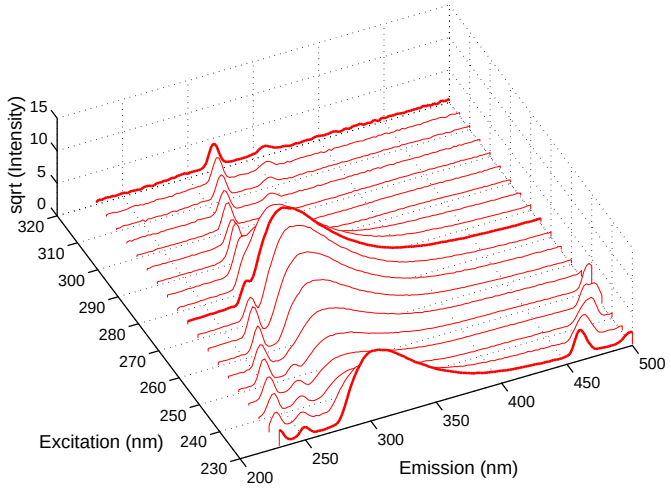


**Fig. 1.** Waterfall view of an excitation–emission landscape. The diagonal ridges, from left to right, represent artifacts caused by primary Rayleigh scatter, Raman scatter and secondary Rayleigh scatter. To improve visibility of the peaks, the vertical scale has been set proportional to the square root of fluorescence intensity. The three spectra drawn with thicker lines are shown individually in Fig. 3.

scatter ridges center quite closely to the lines $\lambda_{em} = \lambda_{ex}$ and $\lambda_{em} = 2\lambda_{ex}$, respectively. Raman scatter also forms a straight ridge, but its center line does not follow such a simple relationship: slope and offset vary between data sets.

The data were obtained from the Department of Food Science, in the Faculty of Life Sciences of the University of Copenhagen (http://www.models.kvl.dk/Fluorescence). See also Bro et al. [13].

Because the secondary Rayleigh scatter manifests itself in a region where there is essentially no analytic signal, we can get rid of it very simply, by limiting the excitation wavelength to values below 450 nm.

The perspective view already gives the impression that the observed scatter signals have a Gaussian shape, along the emission wavelength. This impression is amplified by Fig. 3, where we show cuts through the fluorescence landscape at three excitation wavelengths. Our strategy will be to model the scatter in individual emission spectra by

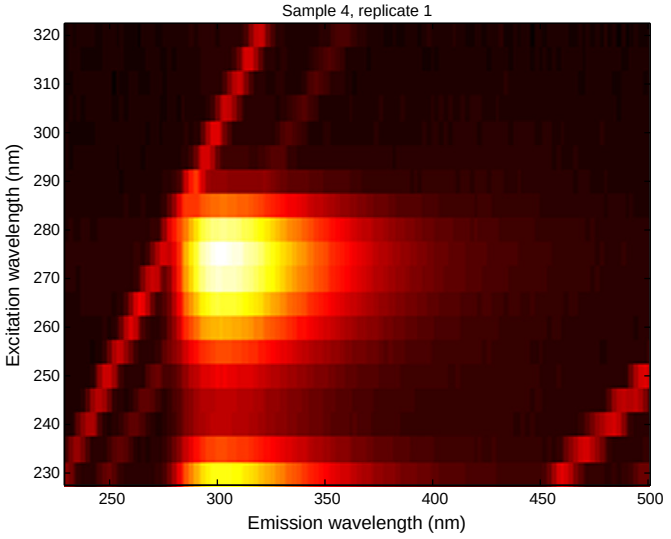


**Fig. 2.** Image view of the excitation–emission landscape. The diagonal ridges, from left to right, represent artifacts caused by primary Rayleigh scatter, Raman scatter and secondary Rayleigh scatter. To improve visibility, the color scale has been set proportional to the square root of fluorescence intensity.

Gaussian curves using non-linear regression. Three parameters describe such a curve completely: position, height and width.

Assuming that good starting values are provided, fitting of Gaussian curves to data is not difficult. The problem is that we do not only have the Gaussian curves but also the signal. In the upper and lower panels of Fig. 3 the scatter peaks are clearly separated from the signal, or the signal under the peak is weak. The middle panel shows a less favorable situation. Both primary Rayleigh and Raman scatter are superimposed on a strong signal. Trying to fit a Gaussian curve locally will lead to useless results.

To solve this problem, we locally model the signal as a smooth baseline. Locally means that for each excitation wavelength and each type of scatter we select an appropriate section of the emission wavelengths, which is known to contain the scatter peak. Because of the linear trend of the centers of the scatter peaks and their stable widths, it is not hard to define such windows. Details will be presented below.

The model for the fluorescence intensity is, at excitation wavelength $j$ and emission wavelength $i$:

$$y_{ij} = b_{ij} + g\left(x_i; \alpha_j, \mu_j, \sigma_j\right) + e_{ij}, \tag{1}$$

where $b_{ij}$ represents the signal as a smooth baseline, $x_i$ the emission wavelength, and $e_{ij}$ random noise. The Gaussian shape is described by the function $g$; $\exp \alpha_j$, gives its height, while $\mu_j$ determines its location and $\sigma_j$ its width, as can be seen from its definition:

$$g\left(x_i; \alpha_j, \mu_j, \sigma_j\right) = \exp\left(\alpha_j - \frac{\left(x_i - \mu_j\right)^2}{2\sigma_j^2}\right). \tag{2}$$

To simplify the equations, we drop the subscript $j$ when no confusion can occur. Note that we do not divide $g$ by $\sigma\sqrt{2\pi}$ as is the case for a normal density.

Our objective function is

$$S = \sum_i \left(y_i - b_i - g(x_i; \alpha, \mu, \sigma)\right)^2 + \lambda \sum_i \left(\Delta^2 b_i\right)^2. \tag{3}$$

The first term in Eq. (3) is the familiar sum of squares of differences between data and model. The second term is a penalty on roughness of $b$, based on second order differences of $b$; note that $\Delta^2 b_i = b_i - 2b_{i-1} + b_{i-2}$. When $b$ is smooth, second order differences will be relatively small and this term will not contribute much to the objective function. The parameter $\lambda$ balances fidelity to the data and smoothness of the baseline. The penalty term is inspired by the Whittaker smoother [14].

The objective function is non-linear in the parameters $\alpha$, $\mu$ and $\sigma$. If we consider small changes $\delta\alpha$, $\delta\mu$ and $\delta\sigma$, the first-order Taylor expansion gives:

$$g(x; \mu + \delta\mu, \sigma + \delta\sigma) \approx g(x; \mu, \sigma)\left(\delta\alpha + u\delta\mu/\sigma + u^2\delta\sigma/\sigma\right), \tag{4}$$

where $u = (x - \mu) / \sigma$. Using the partial derivatives, the non-linear regression problem becomes linear in $\delta\alpha$, $\delta\mu$ and $\delta\sigma$ and will be easy to solve. With proper starting values it converges in a handful of iterations. If no baseline were present, this is an effective approach to fit a Gaussian shape to scatter peaks.

A simple way to handle the baseline is to use back-fitting. Alternatingly one fits the Gaussian peak after correction for the baseline and one recomputes the baseline (by Whittaker smoothing) after removing the peak. Convergence is relatively slow: about 100 iterations can be needed to arrive at a precision of four significant figures. On the other hand, on a modern PC, one iteration takes about a millisecond, so it takes only a few seconds to correct one type of scatter for all excitation wavelengths.