Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 76 (2005) 121–133

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Background removal from spectra by designing and minimising a non-quadratic cost function

Vincent Mazet[a,*], Cédric Carteret[b], David Brie[a], Jérôme Idier[c], Bernard Humbert[b]

[a]CRAN-CNRS UMR 7039, Université Henri Poincaré, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France
[b]LCPME-CNRS UMR 7564, 405 rue de Vandoeuvre, 54600 Villers-lès-Nancy, France
[c]IRRCyN-CNRS UMR 6597, 1 rue de la Noë, BP 92101, 44321 Nantes Cedex 3, France

Received 15 July 2004; received in revised form 30 September 2004; accepted 7 October 2004
Available online 25 November 2004

Abstract

In this paper, the problem of estimating the background of a spectrum is addressed. We propose to fit this background to a low-order polynomial, but rather than determining the polynomial parameters that minimise a least-squares criterion (i.e. a quadratic cost function), non-quadratic cost functions well adapted to the problem are proposed. To minimise these cost functions, we use the half-quadratic minimisation. It yields a fast and simple method, which can be applied to a wide range of spectroscopic signal. Guidelines for the choice of the design parameters are given and illustrated on simulated spectra. Finally, the effectiveness of the method is shown by processing experimental infrared and Raman spectra.

Keywords: Background removal; Polynomial fitting; Half-quadratic minimisation; Huber function; Truncated quadratic; Infrared and Raman spectra

1. Introduction

Spectra (such as infrared or Raman spectra on which this method of background correction will be applied) generally consist in peaks and noise superimposed on a background. This background, or baseline, which can be either flat, linear with a positive or negative slope, curved or a combination of all three, is mainly characterised by the fact that it does not vary as quickly as the peaks do. In Raman spectra, this background is often due either to residual Rayleigh scattering at low Raman wavenumbers or to fluorescence of organic molecules intrinsic to the analysed sample or coming from contamination. In infrared spectra, the deviations in intensity can be due to a scattering of infrared beam caused by heterogeneities in the solid, extern light or a source of non-specific absorption. Subtracting the estimation of the background from the raw spectrum leads to a more interpretable signal, allowing to determine peak wavenumbers and to measure area and amplitude of peaks more accurately.

In most softwares (like Origin[1] or PeakFit[2]) as well as published papers [1,2], the background is estimated by a least-squares polynomial fitting performed on a user defined subset of points, which should belong to the background. Providing that the points are correctly selected, the fitting yields satisfactory results. This can be attributed to the ability of the polynomial model to represent a wide class of backgrounds. But selecting the right points is not always easy and could be a burdensome and time-consuming operation if one has to process many spectra since that should be done for each spectrum individually. In that respect, there is a real need to develop methods that perform an automatic point selection or that are insensitive to the occurrence of peaks [3–5]. This is the topic addressed in this paper; however, before going further, let us mention that

* Corresponding author. Tel.: +33 3 83 68 44 61; fax: +33 3 83 68 44 62.
E-mail address: vincent.mazet@cran.uhp-nancy.fr (V. Mazet).

[1] OriginLab.
[2] Systat Software.

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.10.003

many other approaches are available for the background estimation.

The wavelet transform has became a useful chemometric tool [8–11], in particular for background removal. Basically, the developed methods consist in applying a wavelet transform (Daubechies or Symlet wavelets in many applications) to the spectrum, from which the wavelet coefficients are computed, and then separating the background supposed to be in the low-frequencies part in the spectrum (approximation coefficients) from the peaks and the noise supposed to be in the high-frequencies (detail coefficients). The main shortcoming of such an approach is that it implicitly supposes that the background is well separated (in the transformed domain) from the rest of the signal. Direct orthogonal signal correction (DOSC) methods aim at removing, from a sequence of spectra, the variations which are, as much as possible, orthogonal to the concentration matrix [12,13]. This is carried out by projecting the spectra sequence matrix on the concentration matrix, thus decomposing the spectra sequence matrix into two orthogonal parts, one part lying in the concentration space (the estimated spectrum), and the other part being orthogonal to it (the estimated background). However, as this method is designed as a preprocessing step for spectroscopic calibration, it requires the recording of a sequence of spectral data as well as a known concentration matrix. Some authors have addressed the problem in a Bayesian framework [14,15]. They model the background as cubic splines and look at estimating both the knot positions and cubic spline coefficients while imposing some smoothness prior to the background. Because the posterior density is difficult to minimise, Monte Carlo Markov Chain (MCMC) methods are used, leading to a high computational burden. Note that, in Ref. [15], the background estimation is coupled with the peak deconvolution.

In this paper, we address the problem of background estimation as a polynomial fitting where the polynomial coefficients are estimating by minimising a non-quadratic criterion. The paper is organised as follows: in Section 2, the model is presented. Then, the cost functions are designed to avoid the peaks to be too influent on the estimation. In this respect, Huber and truncated quadratic cost functions are considered and extended to the case of only positive peaks resulting in asymmetric cost functions. Because the minimisation of these cost functions is not straightforward, we use half-quadratic minimisation [16,17]. The proposed methods are then linked with available approaches such as least trimmed squares [18] and the method proposed in [3] for which a mathematical foundation is provided. Section 3 presents the simulated data used to illustrate the performances of the methods. In this section, some experiments are also performed to discuss the respective advantages and disadvantages of the different cost functions (symmetric/asymmetric, Huber/truncated quadratic). The influence of the signal parameters (sample size N and contamination rate Q) is studied and some guidelines for the choice of the hyperparameters (the threshold s and the polynomial order p) are given. The effectiveness of the method is also demonstrated through applications on experimental infrared and Raman spectra in Section 4. Finally, some conclusions and perspectives are given in Section 5.

2. Theory

2.1. Problem modeling

Defining a model of the spectra will allow us to design the cost functions, but also to compare them using simulated spectra (Section 3.1). We note $y=(y_1 \cdots y_N)^T$ the N-point spectrum, such as $\boldsymbol{y}=\boldsymbol{b}+\boldsymbol{e}$, where:

- $\boldsymbol{b}$ denotes the background itself. It is modeled as a p-order polynomial. Actually, taking a polynomial for the background seems to be able to model most spectra [1–3];
- $\boldsymbol{e}$ denotes the residual, gathering peaks, physical noise and model uncertainties. The peaks, which can be either positive and negative or all with the same sign, have different shapes, amplitudes, positions and widths; the physical noise and model uncertainties are modeled (for algorithmic simplicity) as a white, Gaussian and additive noise $\boldsymbol{n}$ with variance $\sigma_{\boldsymbol{n}}^2$.

The background being modeled as a polynomial function, it can be written as $\boldsymbol{b}=\boldsymbol{T}\boldsymbol{a}$ where $\boldsymbol{T}$ and $\boldsymbol{a}$ are defined as

$$\boldsymbol{T}=\begin{pmatrix} t_1^0 & \cdots & t_1^p \\ \vdots & & \vdots \\ t_N^0 & \cdots & t_N^p \end{pmatrix},\quad \boldsymbol{a}=\begin{pmatrix} a_0 \\ \vdots \\ a_p \end{pmatrix}$$

and represent respectively the wavenumber Vandermonde matrix and the polynomial coefficients.

2.2. Design of the cost functions

From a general point of view, the proposed method consists in finding the polynomial coefficients $\boldsymbol{a}$ which will minimise a criterion of the form:

$$\mathcal{J}(\boldsymbol{a})=\sum_{k=1}^{N}\varphi\left(\boldsymbol{y}_k-(\boldsymbol{T}\boldsymbol{a})_k\right) \tag{1}$$

where $(\boldsymbol{T}\boldsymbol{a})_k$ represents the kth element of vector $\boldsymbol{b}=\boldsymbol{T}\boldsymbol{a}$. The criterion depends on the function φ, which, in the sequel, will be referred to as the "cost function". In this section, we consider the design of such a cost function well suited to the problem of background estimation.

First, consider the background estimation using the classical least squares approach. It consists in finding the

coefficients $\boldsymbol{a}$, which minimise the mean squares error between the estimated background and the signal. This approach gives a useless result because the cost function which is $\varphi(x)=x^2$ (Fig. 1(a)) gives a quadratic cost to each value $\boldsymbol{y}_k-(\boldsymbol{Ta})_k$. Thus, large values have a too high cost. The point is that the values, which are very far from the true background, will greatly affect the polynomial coefficient estimation. Another way to see the limitation of the least squares approach is to consider its probabilistic interpretation. Indeed, using a quadratic cost function is equivalent to assume that $\boldsymbol{e}$ is distributed as a zero mean Gaussian [19], which is obviously not the case since it is the sum of a Gaussian noise and peaks.

To handle this problem, we may use cost functions whose cost is lower for large values. These functions should be quadratic in the neighbourhood of zero, i.e. when the background and the spectrum are closed (this satisfies the probabilistic interpretation of a Gaussian noise around zero) but they should grow more slowly than a quadratic beyond a threshold s so that the peaks will be less influent on the background estimation (the choice of this threshold is addressed in Section 3.6). In Ref. [4], the two following cost functions, initially proposed in the context of outlier robust estimation [20,21], are considered:

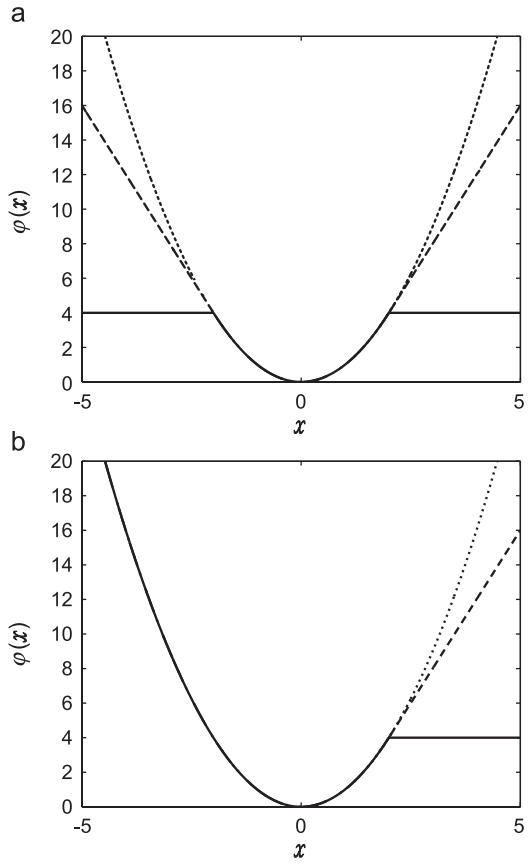

Fig. 1. Symmetrical (a) and asymmetrical (b) cost functions (···: quadratic, - -: Huber function, —: truncated quadratic). s=2.

- Huber function (Fig. 1(a), dashed line):

$$\forall x \in \mathbb{R}, \quad \varphi(x) = \begin{cases} x^2 & \text{if } |x|<s, \\ 2s|x| - s^2 & \text{otherwise;} \end{cases} \tag{2}$$

- Truncated quadratic (Fig. 1(a), plain line):

$$\forall x \in \mathbb{R}, \quad \varphi(x) = \begin{cases} x^2 & \text{if } |x|<s, \\ s^2 & \text{otherwise.} \end{cases} \tag{3}$$

The quadratic function used in the least squares method is also drawn on Fig. 1(a) (dotted line) for comparison with the previous functions. While the least squares estimator cost function is quadratic everywhere, Huber function and the truncated quadratic are respectively linear and constant beyond s. In the case of the truncated quadratic, the points whose distance from the estimation is higher than s have a constant cost. In other words, a great peak will affect the estimation as if it were a small one.

The previous cost functions are symmetrical, which means that they give a low cost for large positive values, but also for large negative values, which is adequate for spectra where peaks may either be positive or negative. In the particular case of optical spectroscopy where there are only positive peaks, we propose to use the corresponding asymmetric cost functions:

- Asymmetric Huber function (Fig. 1(b), dashed line):

$$\forall x \in \mathbb{R}, \quad \varphi(x) = \begin{cases} x^2 & \text{if } x<s, \\ 2sx - s^2 & \text{otherwise;} \end{cases} \tag{4}$$

- Asymmetric truncated quadratic (Fig. 1(b), plain line):

$$\forall x \in \mathbb{R}, \quad \varphi(x) = \begin{cases} x^2 & \text{if } x<s, \\ s^2 & \text{otherwise.} \end{cases} \tag{5}$$

These cost functions still give a low cost to positive peaks, but are quadratic in the negative part to account for the fact that, in this case, there is only Gaussian noise.

2.3. Half-quadratic minimisation of the cost functions

To sum up, we have to minimise the criterion (1). In the case of the least squares method, φ is quadratic and the minimisation is easy and yields the following explicit expression:

$$\hat{\boldsymbol{a}} = (\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{y}.$$

Contrarily to the least squares approach, the minimisation of the other cost functions is not straightforward. So, we propose to minimise these criteria using the half-quadratic (HQ) minimisation, which is an iterative technique simplifying the optimisation of a non-quadratic

criterion [16,17]. Provided that φ satisfies the following condition:

$$\exists \alpha_{\max} / \forall \alpha \in [0; \alpha_{\max}[, g_\alpha(x) = x^2/2 - \alpha\varphi(x) \text{ is strictly convex,}$$

the HQ minimisation consists in introducing an auxiliary variable $\boldsymbol{d}=(d_1 \cdots d_N)^T$ leading to an augmented criterion $\mathcal{K}(7)$ admitting the same minimum as $\mathcal{J}$:

$$\mathcal{K}(\boldsymbol{a},\boldsymbol{d}) = \frac{1}{\alpha}\sum_{k=1}^{N}\frac{1}{2}\left(\left(\boldsymbol{y}_k - (\boldsymbol{T}\boldsymbol{a})_k - \boldsymbol{d}_k\right)^2 + \zeta_\alpha(\boldsymbol{d}_k)\right),$$

where the function ζ_α is defined from φ as follows:

$$\zeta_\alpha(d) = \sup_x \left(\alpha\varphi(x) - (x-d)^2/2\right).$$

Note that the new criterion $\mathcal{K}$ is quadratic in $\boldsymbol{a}$ and convex in $\boldsymbol{d}$, justifying the name "HQ criterion". Huber function and the truncated quadratic verify the previous condition, both with $\alpha_{\max}$ equal to 1/2. Then, the algorithm LEGEND used in computed imaging for the minimisation of $\mathcal{J}$ can be applied [16]; it estimates alternately $\boldsymbol{a}$ and $\boldsymbol{d}$ as follows:

initialise $\hat{\boldsymbol{a}}^0$
repeat until convergence:

$$\hat{\boldsymbol{d}}^i = \arg\min_d \mathcal{K}(\hat{\boldsymbol{a}}^{i-1}, \boldsymbol{d})$$

$$\hat{\boldsymbol{a}}^i = \arg\min_a \mathcal{K}(\boldsymbol{a}, \hat{\boldsymbol{d}}^i)$$

The value of α should be chosen close to $\alpha_{\max}$ in order to speed up the convergence; in our application, we choose $\alpha=0.99\times\alpha_{\max}$ and $\boldsymbol{a}$ is initialised as the least squares estimation. The convergence is considered to be reached when the difference $\mathcal{K}^i - \mathcal{K}^{i-1}$ becomes lower than a predefined value. Then, the two steps of one iteration become (the superscripts are omitted for notational simplicity):

- the minimisation of $\mathcal{K}$ with respect to $\boldsymbol{a}$ when $\boldsymbol{d}$ is fixed yields:

$$(\boldsymbol{T}^T\boldsymbol{T})\hat{\boldsymbol{a}} = \boldsymbol{T}^T(\boldsymbol{y}+\boldsymbol{d}), \Rightarrow \hat{\alpha} = \left(\boldsymbol{T}^T\boldsymbol{T}\right)^{-1}\boldsymbol{T}^T(\boldsymbol{y}+\boldsymbol{d}). \quad (6)$$

This result can be interpreted as the least squares estimator on the signal $\boldsymbol{y}+\boldsymbol{d}$. To make the algorithm faster, the matrix $(\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}^T$ can be calculated and saved only once at the beginning.

- when $\boldsymbol{a}$ is fixed, the minimum of $\mathcal{K}$ is reached at:

$$\forall k, \hat{\boldsymbol{d}}_k = -\varepsilon_k + \alpha\varphi'(\varepsilon_k) \quad (7)$$

where $\varepsilon_k=\boldsymbol{y}_k-\boldsymbol{b}_k=\boldsymbol{y}_k-(\boldsymbol{T}\boldsymbol{a})_k$ and:

$$\varphi'(x) = \begin{cases} -2s & \text{if } x \le -s \\ 2x & \text{if } |x|<s \\ 2s & \text{if } x \ge s \end{cases} \text{ (symmetric Huber function),}$$

$$\varphi'(x) = \begin{cases} 2x & \text{if } |x|<s \\ 0 & \text{otherwise} \end{cases} \text{ (symmetric truncated quadratic),}$$

$$\varphi'(x) = \begin{cases} 2x & \text{if } x<s \\ 2s & \text{otherwise} \end{cases} \text{ (asymmetric Huber function),}$$

$$\varphi'(x) = \begin{cases} 2x & \text{if } x<s \\ 0 & \text{otherwise} \end{cases} \text{ (asymmetric truncated quadratic).}$$

Note that the Huber cost function being convex, the convergence of the algorithm to the unique global minimum is ensured [16]. On the contrary, the truncated quadratic being not convex, the criterion $\mathcal{K}$ may have local minima, even if it is convex with respect to $\boldsymbol{a}$ when $\boldsymbol{d}$ is fixed and vice versa. So, using LEGEND, we cannot guarantee the global minimum to be reached. However, in all the trials, we have performed on simulated spectra, the estimated backgrounds were very close to the actual ones, letting us to think that the global minimum is reached at each time.

2.4. Comparison with other asymmetric cost function based approaches

The idea of an asymmetric cost function has already been used in Refs. [6,7].[3] In particular, the method proposed in Ref. [6] is applied to the problem of background estimation, while the method of Ref. [7] could be easily adapted to that problem.

These two papers propose to minimise the following criterion:

$$\mathcal{K}(\boldsymbol{b}) = \sum_{k=1}^{N} f(\boldsymbol{y}_k - \boldsymbol{b}_k) + \lambda\sum_{k=1}^{N} g\left(\Delta^2\boldsymbol{b}_k\right)$$

where, in the first term, f corresponds to the asymmetric cost function, and the second (regularisation) term controls the smoothness of the solution through the minimisation of the second derivate of the estimation. The positive parameter λ sets the weight of the second term: the larger λ is, the smoother the estimation is. In Ref. [6],

$$f(x) = vx^2, \quad g(x) = x^2.$$

[3] The authors are very indebted to one of the reviewer for pointing out these references.

where $v=\tau$ if $x>0$ and $v=1-\tau$ otherwise, with $\tau\in[0,1]$. In Ref. [7],

$$f(x) = x(\tau - \mathbb{I}_{x<0}), \quad g(x) = |x|.$$

where $\tau\in[0,1]$ and the estimation is constrained to the class of cubic splines.

Because of their asymmetric shapes, the cost functions f can be used to perform the polynomial fitting proposed in this paper. They are expected to yield similar results than those provided by the asymmetric Huber function. Indeed, they give a higher cost to negative values than to positive values. But, as they do not have a constant part, the high value peaks still remain influent on the estimation.

In the approach proposed in this paper, the smoothness of the estimated background is controlled by fixing the polynomial order to a low enough value. Instead, in Refs. [6,7], the smoothness of the estimation is ensured by adding a regularisation term to the cost function, the smoothness being controlled by the parameter λ. Then, this estimation is not constrained to be a polynomial function, which can be interesting to fit very irregular backgrounds. The same can be done in the proposed polynomial fitting approach by increasing the polynomial order, but this may lead to numerical problems.

Future works could be directed at combining the proposed cost functions with the smoothness regularisation of Ref. [6], that is to use one of the proposed cost functions in the criterion of Ref. [6]. In that respect, the half-quadratic minimisation provides a very attractive approach to perform the minimisation of this new criterion.

2.5. Remarks about the truncated quadratic

2.5.1. Link with the least trimmed squares

It can be noted that the asymmetric truncated quadratic cost function is similar to that of a least trimmed squares (LTS) approach, as defined in Ref. [18]. Indeed, the LTS performs a least squares estimation on a subset of the data points, while ignoring the other points (peaks in our application): this is equivalent to assigning a constant cost to these peaks. The subset point number being set to a predefined value at the beginning, the algorithm aims at finding the subset which minimises the squared residual on the considered subset. To avoid an exhaustive search of this subset, a fast method is also proposed in Ref. [18]. Because the HQ minimisation approach gives a constant cost to peaks and a quadratic cost to the rest of the spectrum, it also defines two subsets of points. Assuming that the same points are affected to the two subsets, both methods would produce exactly the same result from which it can be concluded that the two methods are equivalent. In fact, the only difference between the two approaches comes from the way in which the subset search is performed. These methods give almost equivalent results but our method is faster [5]. As a conclusion, it appears also that the proposed method is equivalent to the one used in most commercial softwares, which computes the LS estimator on a predefined subset, but the point is that the subset selection is automatic.

2.5.2. Link with [3]

Lieber and Mahadevan-Jansen [3] have presented an iterative algorithm to estimate the background on a Raman spectrum (where peaks are positive) by a least-squares-based polynomial fitting in which peaks are eliminated. The estimation is computed from the signal $\boldsymbol{y}$, redefined at each iteration. Each point is set equal to the estimated background if the corresponding spectrum intensity is higher than the estimated background, otherwise it is set equal to the spectrum.

This method is equivalent to minimise an asymmetrical truncated quadratic with a threshold s set to zero. Indeed, LEGEND estimates the background as the least squares estimation on the signal $\boldsymbol{y}+\boldsymbol{d}$ (Eq. (6)):

$$\hat{\boldsymbol{a}} = \left(\boldsymbol{T}^T\boldsymbol{T}\right)^{-1}\boldsymbol{T}^T(\boldsymbol{y}+\boldsymbol{d}).$$

From Eq. (7), and considering a particular sample k,

$$\begin{aligned}\boldsymbol{y}_k + \hat{\boldsymbol{d}}_k &= \boldsymbol{y}_k - \varepsilon_k + \alpha\varphi'(\varepsilon_k)\\ &= \boldsymbol{b}_k + \alpha\varphi'(\boldsymbol{y}_k - \boldsymbol{b}_k)\end{aligned} \tag{8}$$

where, choosing the limit value $\alpha=\alpha_{\max}=1/2$, we have:

$$\alpha\varphi'(\boldsymbol{y}_k - \boldsymbol{b}_k) = \begin{cases}\boldsymbol{y}_k - \boldsymbol{b}_k & \text{if } \boldsymbol{y}_k - \boldsymbol{b}_k < s, \text{ i.e. } \boldsymbol{y}_k < \boldsymbol{b}_k + s,\\ 0 & \text{otherwise.}\end{cases}$$

So Eq. (8) becomes:

$$\boldsymbol{y}_k + \boldsymbol{d}_k = \begin{cases}\boldsymbol{y}_k & \text{if } \boldsymbol{y}_k < \boldsymbol{b}_k + s,\\ \boldsymbol{b}_k & \text{otherwise,}\end{cases}$$

meaning that LEGEND consists in computing the least squares estimator $\boldsymbol{b}$ on the signal $\boldsymbol{y}+\boldsymbol{d}$, then, at each sample k, the estimation is set to $\boldsymbol{y}_k$ if $\boldsymbol{y}_k$ is lower than the previous background estimation plus the threshold, otherwise it is equal to this estimation. Thus, the method presented in Ref. [3] is a particular case of the one presented in this article, in which the threshold of the asymmetric truncated quadratic is equal to zero. As mentioned in Section 3.6, a zero value threshold will result in an estimated background pushed down to the bottom of the spectrum. To overcome this problem, the authors propose to let the algorithm running until a maximum number of modified points is reached, which, in turn, is very similar to the LTS stopping rule (see Section 2.5.1).

2.6. Computation

All treatments were performed using in-house codes written in Matlab 6.5 (The Mathworks, MA), on a Pentium 4 1.8 GHz. The Matlab function `polyfit` is used to perform the polynomial fitting. To avoid numerical problems, the wavenumber vector $\boldsymbol{t}$ has to be centered and rescaled in [−1;1]. The sources can be freely downloaded for evaluation on the CRAN website. No particular toolboxes are needed to run the algorithm.

3. Influence and choice of the design parameters

3.1. Simulated data

Simulated data, in which the backgrounds are perfectly known, are used to evaluate and compare the accuracy of the different methods. The simulated spectra follow the model given in Section 2.1: they are the sum of a background, peaks and a white Gaussian noise. The background is randomly chosen as a 4- or 5-order polynomial (the two orders are equiprobable) whose coefficients are generated from a zero-mean Gaussian with variance 1. The polynomial function is then offset to be in the positive part. The peak signal consists is a sparse spike train signal (modeled as a Bernoulli–Gaussian signal whose Bernoulli parameter and peak variance are set by the user) convolved with a Gaussian pulse whose width is let to the user's choice. At last, the variance of the white Gaussian noise is fixed so as to obtain the needed signal-to-noise ratio.

The influence of the spectrum length N, noise variance, polynomial order, threshold, peak number and width will now be discussed.

The performance index used to assess the methods in terms of background estimation is the mean square error (MSE) defined as:

$$MSE = \frac{1}{N}\sum_{k=1}^{N}\left(\boldsymbol{b}_k - \hat{\boldsymbol{b}}_k\right)^2$$

where $\hat{\boldsymbol{b}}=\boldsymbol{T}\hat{\boldsymbol{a}}$.

3.2. Symmetric or asymmetric cost function?

To illustrate the influence of the shape of the cost function, two kinds of spectra are presented on Fig. 2 (a: positive and negative peaks; b: only positive peaks), on which the background is estimated with a symmetrical and an asymmetrical cost function (we choose the truncated quadratic because it yields the best estimation—see next section). On Fig. 2(a), the best estimation is given by the symmetrical form of the cost function and, on Fig. 2(b), the asymmetrical form yields the best result: this is in agreement with the discussion of Section 2.2.

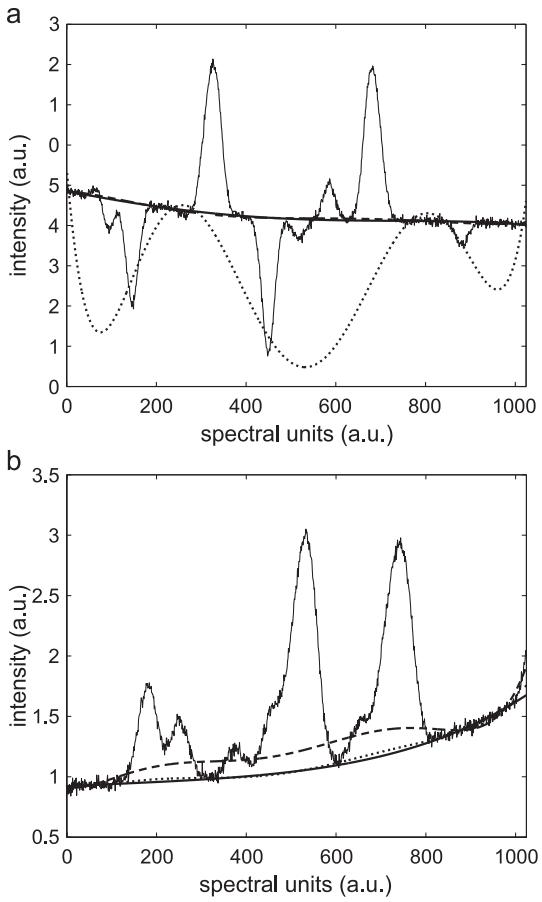

Fig. 2. Estimation with a symmetric (- -) and an asymmetric (· · ·) truncated quadratic, compared with the real background (—). (a) Positive and negative peak spectrum and (b) only positive peak spectrum (a.u. means arbitrary units).

In order to give some quantitative insights into the comparison of the symmetrical and asymmetrical forms of a cost function, we have computed 200 simulations on the two kinds of spectra (Table 1). For each simulation, the MSE between the real and estimated background is computed. It is clear that a symmetrical cost function is better suited for positive and negative peak spectra, while an asymmetrical one should be preferred for spectra with one kind of peak.

3.3. Comparison between Huber function and the truncated quadratic

We now study which cost function (Huber function or truncated quadratic) gives the best estimation. Fig. 3(a) and (b) shows two spectra (one with positive and negative peaks and the second with only positive peaks) whose backgrounds are estimated using the two cost functions, depicted on Fig. 3(c) and (d). For each cost function, the threshold (referred to as the optimal threshold) is chosen as the one giving the best estimation in terms of MSE.

Table 1
Comparison between the four cost functions

	Positive and negative peaks	Only positive peaks
Symmetric Huber function	$49.1 \cdot 10^{-5}$	$241.8 \cdot 10^{-5}$
Asymmetric Huber function	$1129.1 \cdot 10^{-5}$	$14.8 \cdot 10^{-5}$
Symmetric truncated quadratic	$2.7 \cdot 10^{-5}$	$21.4 \cdot 10^{-5}$
Asymmetric truncated quadratic	$6387.1 \cdot 10^{-5}$	$2.0 \cdot 10^{-5}$

The table shows the average MSE between the real and the estimated background for the two kind of peaks.

In all trials we have done (see also Table 1), the truncated quadratic yields the best estimation, providing that the shape (symmetric/asymmetric) is well-chosen with respect to the kind of spectrum. Indeed, with that cost function, all peaks have a constant cost, so that they do not affect the estimation, while, with the Huber cost function, the peaks still influence it. As a consequence, in general, to reduce peak influence, the optimal threshold of the Huber function is lower than that of the truncated quadratic (see Fig. 3(c) and (d)).

3.4. Influence of the signal length N

To assess the influence of the signal length on the estimation quality, we have created a 65,536-point spectrum, which was then decimated with different factors to obtain signals having different lengths but the same background, noise level and contamination rate (defined in Section 3.5). For each signal, the background was estimated and the corresponding MSE was computed (see Fig. 4). It appears that the spectrum length does not significantly affect the estimation quality, providing that the signal length is large enough.

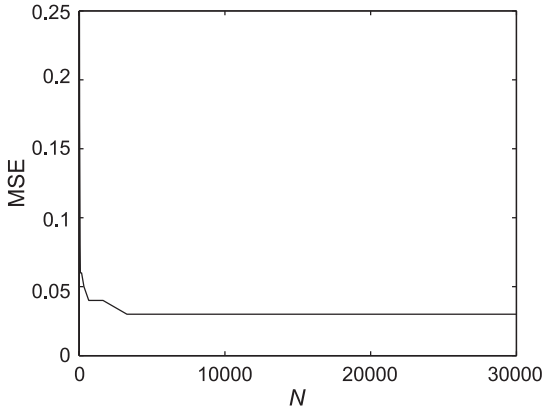

Fig. 4. Mean square error in function of the signal length.

Fig. 5 represents the computation time in function of the point number N. We observe that the computation time is almost linear with respect to the point number and that it remains reasonable even for large signals (less than 1 s for a 16,000-point signal).

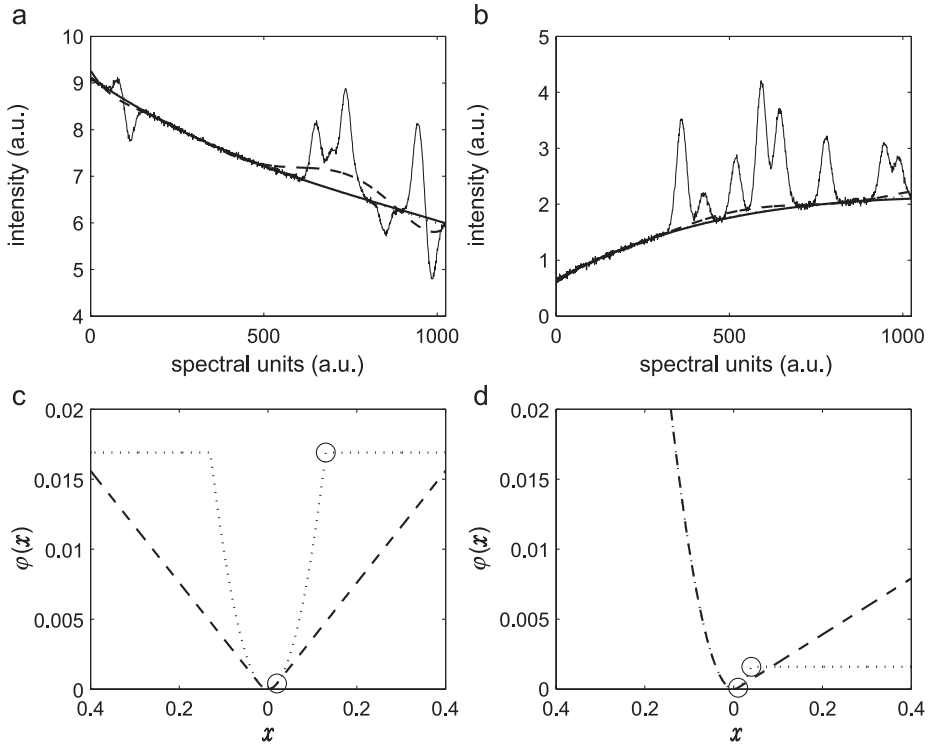

Fig. 3. Comparison between Huber function (- -) and the truncated quadratic (· · ·). The real background is represented as a plain line. (a) Spectrum with positive and negative peaks (s=0.02 for Huber function, s=0.13 for the truncated quadratic). (b) Spectrum with only positive peaks (s=0.01 for Huber function, s=0.04 for the truncated quadratic). (c) Symmetrical cost functions used in Fig. 3(a). (d) Asymmetrical cost functions used in Fig. 3(b). The circles in Fig. 3(c) and (d) correspond to the threshold.

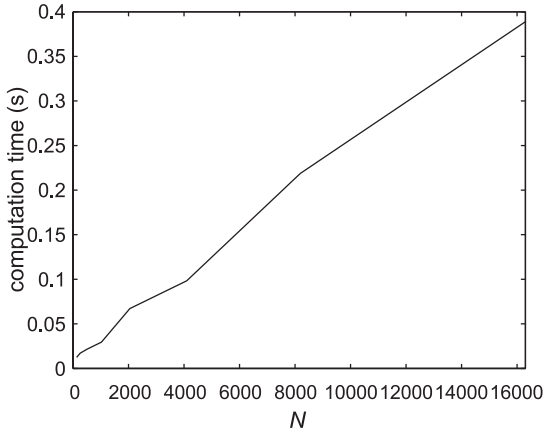

Fig. 5. Computation time in function of the signal length.

3.5. Influence of the contamination rate Q

The contamination rate Q is defined as the ratio between the points belonging to the peaks and those belonging to the background. Then, the higher the peak number or the peak width is, the higher the contamination is. Inevitably, the higher Q is, the more difficult it is to estimate the background, because there are fewer points only belonging to the background.

To assess the influence of the contamination rate, the following simulation has been performed. One hundred clean spectra (i.e. without background and noise) have been simulated. For each of them, 10 backgrounds and 10 noise signals have been generated, resulting in 10 different spectra with the same contamination rate. Then, the background was estimated and the average of the optimal thresholds and the MSE were saved. Fig. 6 shows the performances in terms of MSE (Fig. 6(a)) and the normalised threshold s/σ_n (Fig. 6(b)), both with respect to the contamination rate.

From this simulation, one can actually define three parts:

- in the first part (lower than 10%), the contamination rate is very low. In other words, there are few peaks. Then, the threshold may be higher than twice the noise standard deviation. Anyway, the estimation will be very satisfactory (see Fig. 6(a));
- in the second part (up to 70%, in fact the general case), the ratio s/σ_n decreases almost linearly from 2 to 1, while the MSE increases slowly. In that respect, Fig. 6(b) can be helpful to fix the threshold value;
- the third part (greater than 70%) corresponds to spectra containing almost only peaks. For these spectra, it is very difficult to fit the real background. The best way is to fit the bottom of the peaks by setting the threshold to a very low value. However, as seen in Fig. 6(a), the MSE is not satisfactory.

To conclude this section, the method can fit correctly the background if the contamination rate is not too high (typically lower than 70%). So, from a practical point of view, to have a low contamination rate (and, thus, to estimate more accurately the background), it would be better not to consider only the relevant spectral range, but to record spectra on an extended area including only background parts as well.

3.6. How to choose the threshold s?

As shown in Section 3.5, the threshold s has to be chosen according to the contamination rate Q and the noise level σ_n to yield a satisfactory background estimation.

This point is illustrated on Fig. 7 which shows two signals corresponding to the same spectrum (whose contamination rate is almost 35%) with two different noise levels. On Fig. 7(a), σ_n is set to 0.1 and the optimal threshold is 0.24: about 2.5 times the noise standard deviation. On Fig. 7(b), σ_n is set to 0.03 and the optimal threshold is found to be equal to 0.05, i.e. about 1.5 times the noise standard deviation. So, the optimal value is in accordance with the Fig. 6(b) where the optimal threshold is slightly lower than twice the noise standard deviation for a 35% contamination rate corresponding to that of the spectrum.

To assess the threshold influence on the estimation, Fig. 8 shows a spectrum with a 40% contamination rate and the

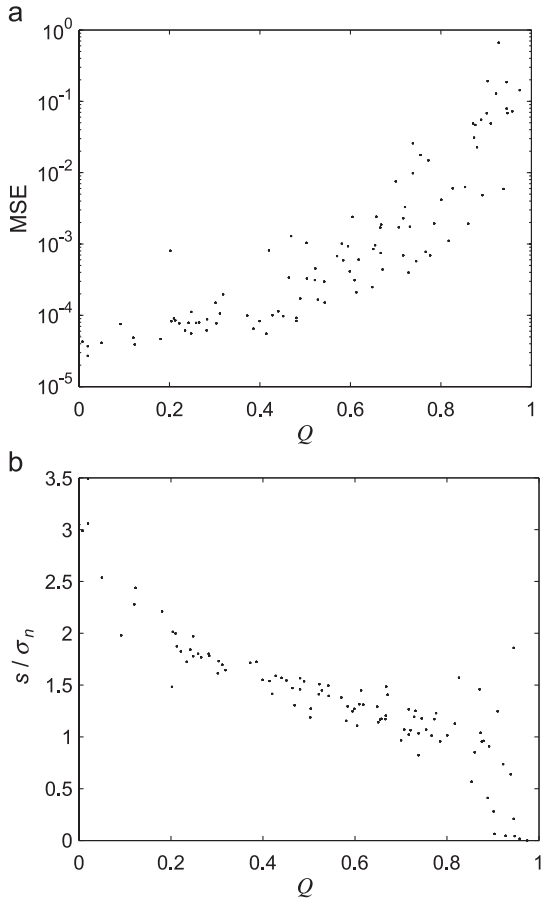

Fig. 6. Influence of the contamination rate (100 simulations). (a) Mean square error and (b) threshold.

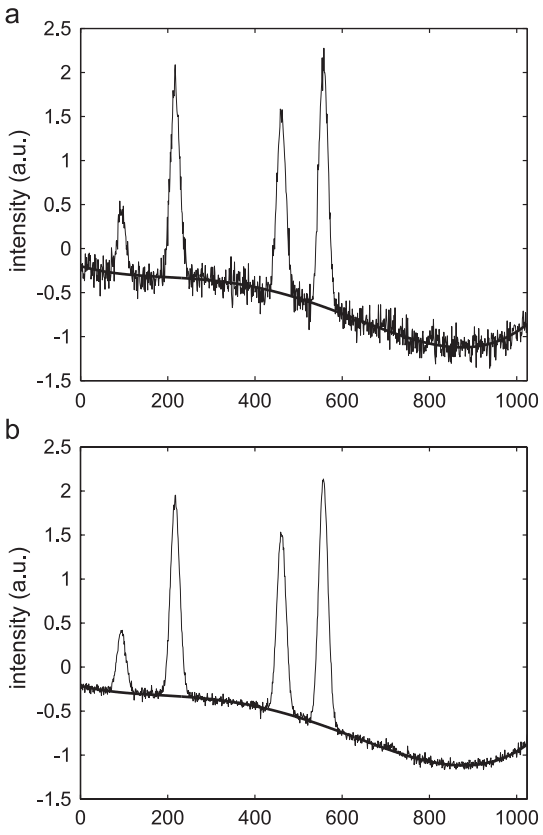

Fig. 7. Influence of the noise (asymmetric truncated quadratic cost function) the estimated background is represented as a plain line. (a) $\sigma_n \simeq 0.10$, s=0.24. (b) $\sigma_n \simeq 0.03$, s=0.05.

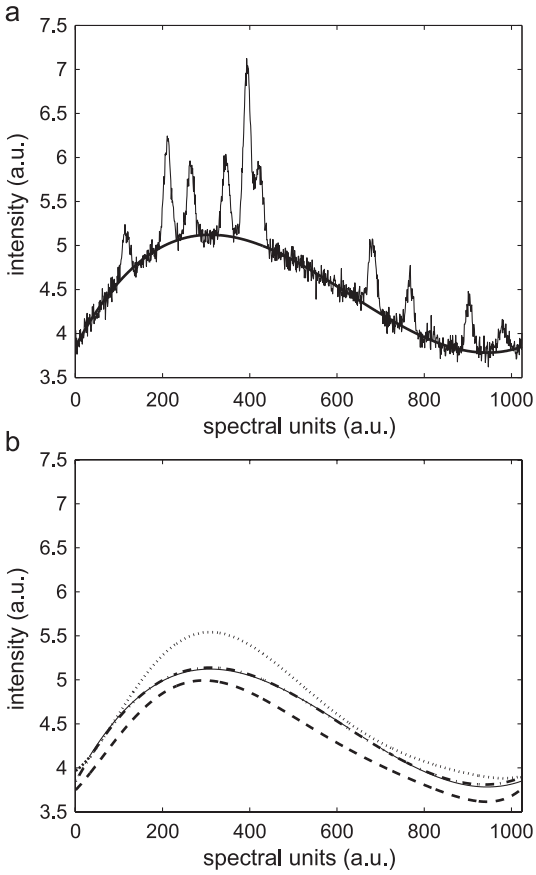

Fig. 8. Estimation of the background with the asymmetric truncated quadratic for three different thresholds. The real background is represented as a plain line ($\sigma_n \simeq 0.07$). (a) Simulated spectrum and its background. (b) Real background and its three estimations (dashed-dotted line: 0.14, dotted line: 10, dashed line: 0.1).

background estimation using the asymmetric truncated quadratic with three different thresholds.

- The threshold which gives the best estimation is twice greater than the noise standard deviation (that is to say, 0.14). This is in accordance with Fig. 6(b).
- If the threshold is set too high, say 10, the estimation tends to the least-squares estimation (dotted line): the cost function tends to a quadratic when the threshold tends toward the infinity;
- On the contrary, when the threshold is set too low, say 0.1, the estimation tends to fit the bottom of the spectrum (dashed line). This is due to the fact that at a low threshold, the negative costs are quadratic, while positive ones tend to zero, then all positive points have a zero cost and are privileged over negative points. Yet, the result would be different for a symmetric cost function. Indeed, when the threshold tends toward zero, the symmetric truncated quadratic tends to a constant function, so that the estimation does not change from one iteration to the next: it remains the least squares estimation, which is the initial background estimation (then, for the truncated quadratic, when the threshold tends either to zero or to infinity, the estimation tends toward the least-squares estimation anyway). In the case of the symmetric Huber function, the estimation tends to minimise the mean absolute error, which yields a non-satisfactory estimation too.

Fig. 9 shows the evolution of the MSE as a function of the normalised threshold for three different contamination rates (12%, 36% and 90%). This simulation yields some

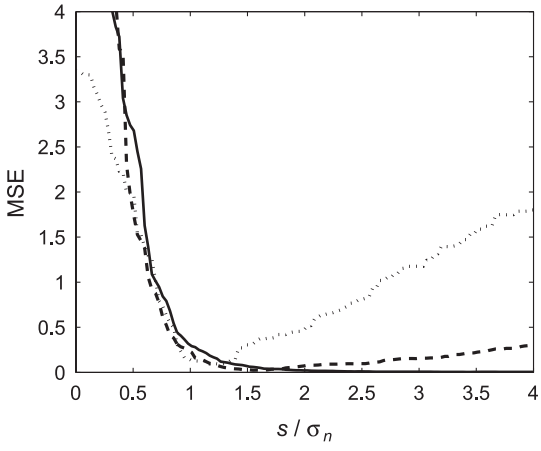

Fig. 9. Evolution of the MSE with respect to the normalised threshold for three different contamination rate (—: 12%, - -: 36%, ⋯: 90%).

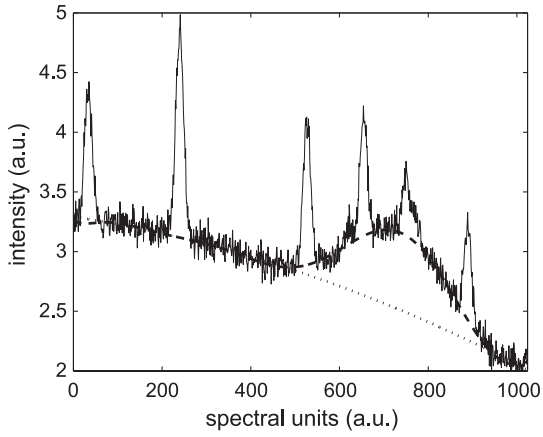

Fig. 10. Estimation of the background with the asymmetric truncated quadratic for two different polynomial orders (···: 2, - -: 10) and the same threshold.

insights into the robustness of the method with respect to the threshold choice from which the analysis of Section 3.5 can be completed.

- we have seen that, for a low contamination rate ($Q \leq 10\%$), the threshold can be set greater than twice the noise standard deviation. Note that, from Fig. 9, a greater threshold does not degrade significantly the estimation since the MSE keeps almost constant beyond the minimal value of s. In fact, when the threshold increases, the cost function tends toward a quadratic, so the estimation tends toward the least-squares estimation (see before). This is the best estimator when there is no peak, that is when all the signal points can be considered as background points;
- for a medium contamination rate ($10\% \leq Q \leq 70\%$), the estimation quality still remains acceptable even if the threshold is set around its optimal value. For example, for a 36% contamination rate, the threshold can be chosen in the interval $[\sigma_n; 4\sigma_n]$ without affecting significantly the MSE. Note that the range of acceptable values of s decreases as the contamination rate increases.
- for a high contamination rate ($Q \geq 70\%$), Fig. 9 shows that the threshold has to be precisely chosen, since the MSE significantly increases for a small variation of s. Anyway, in that case, the estimation quality is poor and the results become erratic, enlightening the limit of the proposed method.

3.7. How to choose the polynomial order p?

Some methods can be envisaged to estimate the polynomial order automatically such as the AIC criterion [24]. However, we believe that it should be let as a user defined parameter allowing to give some degree of freedom to the algorithm. Basically, the polynomial order allows to adjust the estimated background smoothness. So, it is clear that it has to be fixed in function of the background to fit but also of the user need. Indeed, some spectra could have a part of the signal, which can be considered as a background or not. For example, Fig. 10 shows such a simulated spectrum whose "bump" around 700 a.u. may be interpreted as belonging to the background or not. Adjusting the polynomial order allows to fit this part of the signal or not.

4. Application on real spectra

4.1. Experimental infrared and Raman spectra

The samples are gibbsite $Al(OH)_3$ particles prepared by oxidation of aluminum powder in sodium hydroxide solution (see Refs. [22,23] for details of gibbsite preparation).

4.1.1. Infrared spectra

Infrared spectra were obtained with a Fourier transform infrared spectrometer, Perkin Elmer system 2000, by transmission through self-supporting disks. Analysed disks were prepared by dilution of gibbsite in a non-absorbing medium as KBr (0.1% weight of gibbsite). Each 25-mm diameter

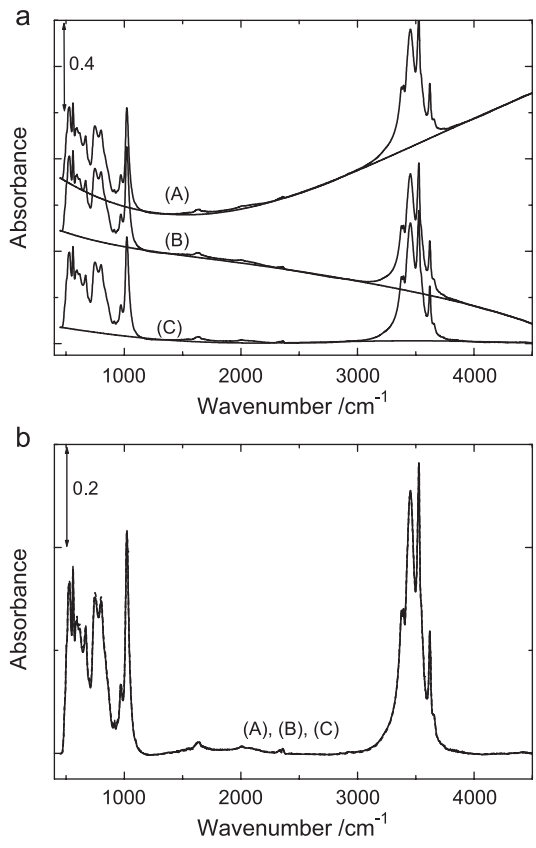

Fig. 11. Background correction with the asymmetric truncated quadratic cost function as it is applied on experimental infrared spectra of gibbsite $Al(OH)_3$. (a) Raw measured spectra (A), (B) and (C), and background estimations with five order polynomials and threshold of 0.005. (b) After removal of the backgrounds which correspond to real physical scattering process.

disk was obtained under a 10 MPa pressure. Spectra were recorded from 400 to 4500 cm^{-1} at 4 cm^{-1} resolution. Acquisition time was 1 min per spectrum. The spectra are presented in absorbance unit: $A=-\log_{10}(I/I_0)$ where I is the sample signal and I_0 is the reference signal of a pure KBr disk. For the same sample signal, we have used different reference signals obtained from different pure KBr pellets. By this procedure, we have obtained different infrared spectra composed of the same absorption spectrum combined with different scattering backgrounds. Noise amplitude is around 0.0005 absorbance unit.

4.1.2. Raman spectra

The Raman spectra were recorded with a triple-substractive-monochromator Jobin Yvon T64000 spectrometer equipped with a confocal microscope. The detector was a charged-coupled device (CCD) cooled by liquid nitrogen. The Raman spectra were excited by a laser beam at 514.3 nm emitted by an argon laser, focused on the samples with a diameter of about 1.5 μm and a power of about 50 mW. The Raman backscattering was collected through the microscope objective (×50) and dispersed by a 1800 groove/mm grating to obtain 2.7 cm^{-1} resolution and a data point each 0.6 cm^{-1}. We choose an integration time of 60 s per accumulation while accumulation number of each spectrum varied in order to obtain different noise amplitudes. We have recorded three spectra: a pure sample of gibbsite and two polluted samples of gibbsite. These polluted samples were obtained by leaving pure gibbsite powder in a cigarette smoking area. Both polluted samples have (almost) identical Raman signal but different fluorescence backgrounds.

4.2. Results

The previous sections showed that the asymmetric truncated quadratic is the cost function which gives the best results to estimate the simulated background on spectra with only positive peaks. It is now applied on real optical spectra. Since the three raw infrared spectra exposed in Fig. 11(a) contain the same quantity of infrared beam absorption combined with different scattering backgrounds, the performance of the background correction method could be evaluated by its ability to return only the infrared absorption component from raw spectra. The backgrounds are estimated by a 5-order polynomial with a threshold of 0.005. After removal of the scattering backgrounds (Fig. 11(b)), all three spectra become identical in agreement with what is expected.

Polluted samples of gibbsite exhibit fluorescence background in their Raman spectra (Fig. 12(a) and (c)). As

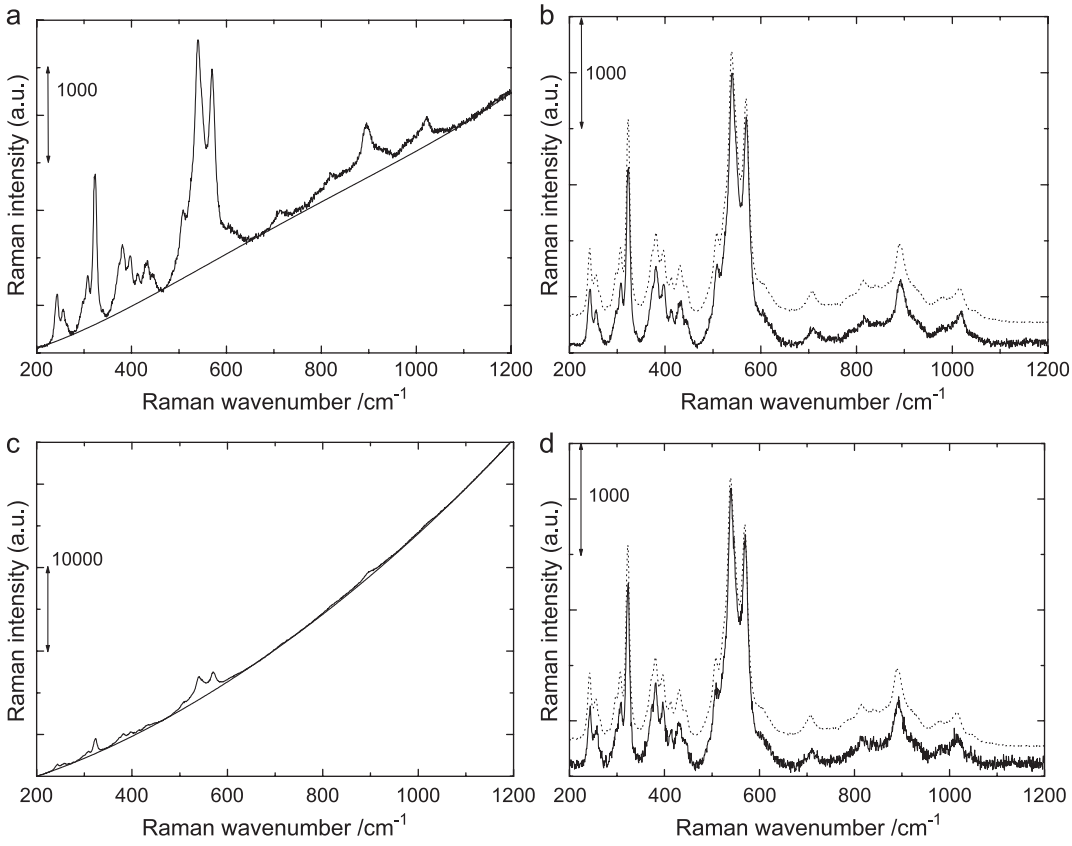

Fig. 12. Background correction with the asymmetric truncated quadratic cost function as it is applied on experimental Raman spectra of gibbsite. (a) Slightly polluted gibbsite sample and background estimation. (b) Background corrected spectrum of slightly polluted gibbsite sample. (c) Highly polluted gibbsite sample and background estimation. (d) Background corrected spectrum of highly polluted gibbsite sample.

Raman spectrum of pure gibbsite (in dotted line in Fig. 12(b) and (d) average SNR of 100) is free from fluorescence background, it can be used as a reference spectrum to test background correction of raw spectra recorded for polluted gibbsite samples. Fig. 12(a) shows the raw spectra of a slightly polluted gibbsite sample prior to fluorescence subtraction where maximum fluorescence intensity is of order of the maximum Raman intensity (intensity of 2500 at 540 cm^{-1}) and noise amplitude is around 25 (average SNR is 30). Fig. 12(c) shows the raw spectra of a highly polluted gibbsite sample prior to fluorescence subtraction where maximum fluorescence intensity is 20 times more than the maximum Raman intensity and noise amplitude is around 50 (average SNR is 15). The fluorescence backgrounds are estimated by a 4-order polynomial and thresholds of 13 and 30 for slightly and highly polluted spectra. After background correction, extracted Raman spectra are in good agreement with the reference spectrum of non-polluted gibbsite (solid line in Fig. 12(b) and (d)). The spectra were offset for clarity.

5. Conclusion

This paper presents an iterative method to estimate spectral backgrounds as the polynomial minimising a cost function. Providing that the cost function is correctly designed for the considered spectrum, the method is well adapted to a wide range of spectroscopy (infrared, Raman, UV–Vis, NMR, etc.) because it does not require any modeling of the peaks. Different cost functions have been designed to avoid the peaks to be too much influent on the estimation, as they are in classical least squares estimation whose cost function is a quadratic. Because the cost function is not quadratic, half-quadratic minimisation is used to optimise the criterion. Huber and the truncated quadratic functions have been proposed and evaluated: they are quadratic in the neighbourhood of zero (modeling a Gaussian noise) and respectively linear and constant beyond a threshold (modeling the peak distribution).

Experiments showed that the truncated quadratic cost function yields the best results. For optical spectroscopy, where the peaks are only positive, asymmetrical cost functions are proposed, actually providing better results while symmetrical cost functions are better suited for background estimation on data with both negative and positive peaks. To implement the method, two parameters need to be set up: the threshold of the cost function and the polynomial order. It turns out that the optimal value of the threshold depends on both the noise standard deviation and contamination rate. However, in many practical cases, the threshold can be set from once to twice the noise standard deviation, giving a satisfactory estimation. It is clear that there is no formal argument justifying this empirical rule. However, the robustness of the method to that choice has been verified experimentally providing that the contamination rate is low enough. Concerning the polynomial order, it should be chosen in dependence of the estimated background smoothness. Even if some automatic methods can be developed to estimate it, we believe it should be let as a user defined parameter allowing, for example, to fit (or not to fit that is the question) the background to "bumps" present in the spectra. Furthermore, the algorithm yields a low computational time, even on very large signals. Tests have been carried out for background correction on experimental optical Raman and infrared spectra. In all the considered cases, the method was very effective leading to an efficiency background removal and no alteration of the signal of interest. Finally, even if the background of most spectra can be correctly estimated, using splines rather than a polynomial could be a better way to fit irregular backgrounds; this will be the subject of future works.

Acknowledgments

This work was supported by the Région Lorraine, France and by the CNRS (Centre National de la Recherche Scientifique).

References

[1] R. Goehner, Background subtract subroutine for spectral data, Anal. Chem. 50 (1978) 1223–1225.
[2] T. Vickers, R. Wambles, C. Mann, Curve fitting and linearity: data processing in Raman spectroscopy, Appl. Spectrosc. 55 (2001) 389–393.
[3] C. Lieber, A. Mahadevan-Jansen, Automated method for subtraction of fluorescence from biological Raman spectra, Appl. Spectrosc. 57 (2003) 1363–1367.
[4] V. Mazet, J. Idier, D. Brie, B. Humbert, C. Carteret, Estimation de l'arriére-plan de spectres par différentes méthodes dérivées des moindres carrés, Chimiométrie 2003, Paris, France, 2003, pp. 173–176.
[5] V. Mazet, D. Brie, J. Idier, Baseline spectrum estimation using half-quadratic minimization, EUSIPCO 2004, Vienna, Austria, 2004.
[6] P.H.C. Eilers, Parametric time warping, Anal. Chem. 76 (2004) 404–411.
[7] R. Koenker, P. Ng, S. Portnoy, Quantile smoothing splines, Biometrika 81 (1994) 673–680.
[8] B. Liu, Y. Sera, N. Matsubara, K. Otsuka, S. Terabe, Signal denoising and baseline correction by discrete wavelet transform for microchip capillary electrophoresis, Electrophoresis 24 (2003) 3260–3265.
[9] H.-W. Tan, S. Brown, Wavelet analysis applied to removing non-constant, varying spectroscopic background in multivariate calibration, J. Chemom. 16 (2002) 228–240.
[10] U. Depczynski, K. Jetter, K. Molt, A. Niemöller, The fast wavelet transform on compact intervals as a tool in chemometrics: I. Mathematical background, Chemom. Intell. Lab. Syst. 39 (1997) 19–27.
[11] T. Tony Cai, D. Zhang, D. Ben-Amotz, Enhanced chemical classification of Raman images using multiresolution wavelet transformation, Appl. Spectrosc. 55 (9) (2001) 1124–1130.
[12] J. Luypaert, S. Heuerding, S. de Jong, D. Massart, An evaluation of direct orthogonal signal correction and other preprocessing methods

for the classification of clinical study lots of a dermatological cream, J. Pharm. Biomed. Anal. 30 (2002) 453–466.
[13] J. Westerhuis, S. deJong, A. Smilde, Direct orthogonal signal correction, Chemom. Intell. Lab. Syst. 56 (2001) 13–25.
[14] R. Fischer, K. Hanson, V. Dose, W. von der Linden, Background estimation in experimental spectra, Phys. Rev., E 61 (2000) 1152–1161.
[15] S. Gulam Razul, W. Fitzgerald, C. Andrieu, Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC, Nucl. Instrum. Methods, A 497 (2003) 492–510.
[16] J. Idier, Convex half-quadratic criteria and interacting auxiliary variables for image restoration, IEEE Trans. Image Process. 10 (2001) 1001–1009.
[17] D. Geman, C. Yang, Nonlinear image recovery with half-quadratic regularization, IEEE Trans. Image Process. 4 (1995) 932–946.
[18] P. Rousseeuw, K. Van Driessen, Computing LTS regression for large data sets, technical report, University of Antwerp (1999).
[19] J. Idier, Approche bayésienne pour les problèmes inverses, Traité, vol. IC2, Hermès, Paris, 2001.
[20] P. Huber, Robust Statistic, Wiley, New York, 1981.
[21] P. Rousseeuw, A. Leroy, Robust regression and outlier detection, Series in Applied Probability and Statistics, Wiley-Interscience, New York, 1987.
[22] M. Jodin, F. Gaboriaud, B. Humbert, Repercussions of size heterogeneity on the measurement of specific surface areas of colloidal minerals: combination of macroscopic and microscopic analyses, Am. Mineral 84 (2004 (October)) 1456–1462.
[23] N. Phambu, B. Humbert, A. Burneau, Relation between the infrared spectra and the lateral specific surface areas of gibbsite samples, Langmuir 16 (2000) 6200–6207.
[24] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Contr. 19 (1974) 716–723.