Fluorescence rejection in Raman spectroscopy by local polynomial kernel regression

by

Anders Malmberg

1998:E16

Department of Mathematical Statistics
Lund University
Lund Institute of Technology
Box 118, S–221 00 Lund, SWEDEN

TYP AV DOKUMENT ☒ Examensarbete ☐ Kompendium ☐ Delrapport ☐ Rapport	**DOKUMENTBETECKNING** **ISRN: LUTFD2/TFMS—5066—SE**

INSTITUTION
Matematisk statistik, Lunds universitet, Box 118, 221 00 Lund

FÖRFATTARE
Anders Malmberg

DOKUMENTTITEL OCH UNDERTITEL
Fluorescence rejection in Raman spectroscopy by local polynomial kernel regression methods

SAMMANFATTNING

The nonparametric kernel regression method has proven to be a good choice when estimating regression functions and their derivatives. Here, this method is used in order to estimate Raman spectra. A local bandwidth selector is used when estimating the measured summarised spectrum and when estimating the fluorescence part of the spectrum the SiZer method might be a useful tool. The localisation of the peaks is decided studying pointwise confidence intervals for the derivative.

NYCKELORD

DOKUMENTTITEL OCH UNDERTITEL - SVENSK ÖVERSÄTTNING AV UTLÄNDSK ORIGINALTITEL
Fluorescenseliminering i Raman spektroskopi med hjälp av icke-parametriska kärnskattningsmetoder

UTGIVNINGSDATUM		**ANTAL SID**	**SPRÅK**
år 98	mån 12	**37**	☐ svenska ☒ engelska ☐ annat

ÖVRIGA BIBLIOGRAFISKA UPPGIFTER	**ISSN 0281—1944**
	ISBN
	PRIS

Signature ______________________ Date 19981208

Preface

This is a Master's thesis created within a collaboration between Lund University Medical Laser Centre and the Department of Mathematical Statistics at Lund University with Lund Institute of Technology.

At Lund University Medical Laser Centre I would like to thank my support team Stefan Andersson-Engels and Sara Pålsson. Stefan for introducing me to the physical problem and explaining the Raman theory and Sara for her quick response whenever needed and for helping me with the physical parts of the thesis.

At Mathematical Statistics I would like to thank my supervisor Ulla Holst for her absolutely fabulous guidance throughout the thesis.

Anders Malmberg

Lund, 16 days before Christmas 1998

Contents

1 Introduction

In this Master's thesis we investigate a new statistical method which can be applied when estimating Raman spectra. Many research groups have done a great deal of work in this area and many articles have appeared that discuss different approaches to this problem, cf. Moiser-Boss et al. [9] and Mahadevan-Jansen et al. [7]. Compared to the earlier techniques this method gives us an estimate of the magnitude of the bias and variance of the resulting Raman spectrum.

Modern laser technology might be useful in diagnosing patients without the need of taking biopsies. Raman spectroscopy is one method which applies the laser technique in order to generate spectra which can be used when diagnosing diseases. The suspected diseases can for instance be cardiovascular diseases or cancer, one of the most serious diseases we have to deal with today. An early detection of cancer might be of much help, and the Raman spectroscopy technique may provide a simple diagnosing procedure. A review of the use of Raman spectroscopy when diagnosing patients is given by Mahadevan-Jansen et al. [8].

At Lund University Medical Laser Centre the researchers are developing spectroscopical methods to diagnose tissue, Raman spectroscopy being one of them. Though, using the best possible techniques, the spectra which are obtained do not only contain Raman signals but also the stronger fluorescence signals. The fluorescence part can make it hard to evaluate the recorded spectrum. Therefore it is important to find statistical methods for fluorescence rejection. Techniques employing shifted-spectra, edge detection and FFT filtering are often used for this purpose.

In this thesis, however, the use of nonparametric kernel methods are proposed. These are fairly recently developed and have become popular in many applications, cf. the books by Wand and Jones [16] or Fan and Gijbles [5].

The data which we have worked with in this thesis are provided by Lund University Medical Laser Centre.

In Section 2, we give an introduction to the physical background and in Section 3 we describe the statistical model. The estimated Raman spectrum can be thought of as a subtraction of the fluorescence from the measured spectrum. The nonparametric kernel regression methods can be used not only to estimate the spectrum which is measured, but also to estimate the disturbing fluorescence. The selection of bandwidths in our method and its impact on the estimates are discussed in Section 4.

The estimators of the fluorescence spectrum and the summarised spectrum are discussed in Section 5. Having these estimates, we can estimate the Raman spectrum as the difference between the two previous. This is discussed in Section 6.

In Section 7 we discuss how to localise the peaks and in Section 8 we make a short summary of the thesis.

2 Physical background

A Raman spectrum can be thought of as a plot where the location of the peaks, the Raman peaks, corresponds to the vibrational frequencies in the tissue. These vibrational frequencies tell us what kind of substance we are dealing with. To be able to explain why our spectra look like they do, and to understand what we are looking for, an introduction to the Raman theory is given in this section. In Section 2.1 we discuss some elementary physics, in Section 2.2 the fluorescence phenomena and finally in Section 2.3 the Raman theory is studied.

A more thorough treatment of the physical background and the Raman theory is given by Svanberg [15].

2.1 Electronic states

When talking about vibrational frequencies, or energy, it is natural to start the physical analysis with a study of the behaviour of the electrons. The electrons in a molecule can have different energies and this is illustrated by an energy diagram. In the diagram we call the lowest state the ground electronic state and a state above this an excited electronic state, as illustrated in Figure 1. An electron with a higher position in the diagram has a larger energy. According to the nature of the electron, it always wants to minimise its potential energy. Therefore all the electrons in the molecule will tend to be in the ground electronic state.

This is a very schematic picture of the structure of the molecule and to be able to explain the Raman theory we are forced to use a slightly more complex picture.

Grouping two atoms into a molecule results in a splitting of the electronic states into vibrational and rotational states. The vibrational states are due to

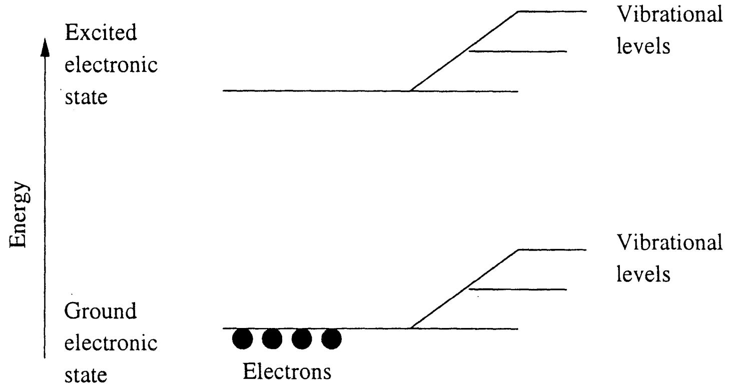

Figure 1: Every electronic state is divided into vibrational levels.

a movement which can be thought of as a result of combining the two atoms with a spring which vibrates as an oscillator and does not change the centre of gravity. The rotational levels are obtained by splitting the vibrational levels, but this is really of no interest to us, and we will not deal with this here.

The result is that we can regard the energy states in our molecule as a ladder where every step, the electronic states, can be divided into vibrational levels.

2.2 Light scattering and fluorescence

If we illuminate a tissue with a laser beam, the photons can interact with the tissue in a number of ways. The photons may be scattered one or several times inside the tissue leading to diffuse reflection and transmission or be absorbed by it; i.e. they will excite some of the electrons in the ground electronic state. If the outgoing light is of the same wavelength as the incoming we say that we have an elastic scattering process. When the scattered light has another wavelength than the incoming, we say that we have an inelastic scattering process and a larger wavelength implies that we have lost energy. This is according to the energy equation for a photon: $E = h \cdot c \cdot \lambda^{-1}$, which relates the energy E of the photon to Planck's constant h, the speed of light c, and the wavelength λ. A spectrum is a measure of the photon flux as a function of energy (or inverse wavelength).

To be absorbed the molecule, the energy that the photon carry must correspond to or be greater than the energy gap between the ground state and the first excited state. Then an electron can make a jump corresponding to the energy given by a photon. Hence it is for instance possible that an electron jumps to the fourth vibrational level in the first excited state.

Maybe, the electron did not come to the lowest vibrational state in the excited electronic state. If this is the case, the electron falls down to it through internal conversion. The internal conversion is a very fast process (10^{-12} s) as a consequence of interaction between different molecules. Striving to be in the ground electronic state, an excited molecule return to its ground state after a while. This can happen in a number of ways, and the way which is of interest to us is when fluorescence light is emitted from the molecule. The fluorescence process is illustrated in Figure 2.

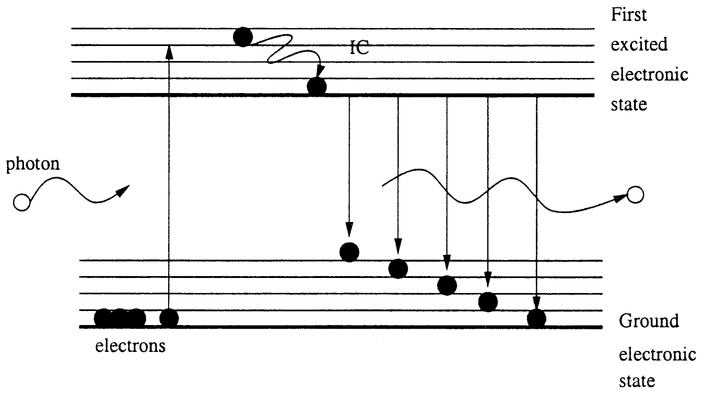

Figure 2: A photon excites a molecule. Due to internal conversion the molecule falls down to the lowest vibrational level in the first excited electron state. From there it falls down to the ground electronic state while emitting fluorescence light.

As illustrated in Figure 2, it is not certain that the electron falls back to the lowest level in the ground electronic state following excitation. If it ends up upon one of the higher positioned vibrational levels the emitted light will be of another wavelength than the incoming. Having many different substances and many different vibrational levels the radiated fluorescence light will be broadened.

All together, this implies that when we illuminate a tissue with a laser beam, it will emit both diffusely scattered light and fluorescence. Information about the tissue can be drawn from fluorescence spectra but a limitation of this method is that only a few substances are fluorescence active; i.e. they are fluorescencing. Since there are many more substances that are Raman active it would be a great advantage if it was possible to study the Raman spectra which is discussed in the next section.

2.3 Raman scattering

Information on the subject can be obtained by studying its Raman spectrum. The big drawback with Raman spectroscopy is that the Raman process is much weaker (10^6-10^8 times) than the fluorescence process. A typical example of a measured spectrum is given in Figure 3. The data has been provided by Lund University Medical Laser Centre and is a measured spectrum of a bone which comes from an amputated leg of a patient with diabetes. This will be our reference spectrum in this thesis.

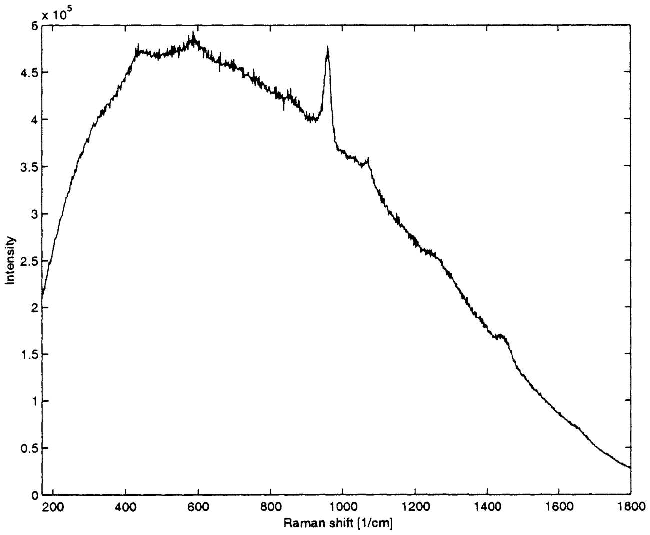

Figure 3: Example of a measured spectrum. (Integration time 100 s.)

The slowly varying curve is the fluorescence curve and the narrow peaks on top of this are the Raman peaks which carry important information about the tissue. For instance it is possible to tell that the narrow peak at 958 cm^{-1} comes from hydroxyapatite $Ca_{10}(OH)_2(PO_4)_6$, a member of the phosphate group.

1. We must calibrate the spectrograph so that the relation between wavelengths and pixels is known. This is done through illuminating the spectrograph with a neon lamp for which the emission lines are known.

2. We must calibrate the Raman scale. We illuminate a substance called Indene for which the spectral characteristic is known. Perhaps we must adjust our scale so that the known peak positions are at the correct wave numbers relative to the laser wave number. To finally arrive at the scale of Raman shift we subtract this wavelength form the others.

A more careful treatment of the practical aspects and how to optimise the set-up can be found in Pålsson [10].

3 Statistical treatment of Raman spectra

After the survey of the physical background in Section 2 we now turn to the statistical model. By working in the near-infrared wavelength area we minimise the number of fluorescencing subjects. Still there is much fluorescence left, and in order to make use of the Raman theory we must remove the fluorescence signal which is left. This can be done with different statistical techniques.

The general solution can be thought of as a subtraction of fluorescence from the received spectra and this can be done in a number of ways. Moiser-Boss et al. [9] showed that rejection of the fluorescence can be done both using shifted-spectra techniques and through filtering in the Fourier domain. Mahadevan-Jansen et al. [7] used a technique where they fit a fifth degree polynomial to the data with the Raman peaks excluded. The Raman spectrum is then taken as the difference between the polynomial fit and a smoothed version of the measured spectrum. However, these papers do not at all deal with estimates of bias and variance of the resulting Raman spectrum.

Another method, which takes variance and bias into consideration, is the nonparametric kernel regression method which will be used in this thesis. This technique will be used to estimate both the summarised spectra and the fluorescence part. Nonparametric kernel regression, cf. the books by Wand and Jones [16] or Fan and Gijbels [5], is a familiar smoothing method. We have chosen to work with this method since it is new and there exist well working programs which we can use. The method and the computer programs are thoroughly described in papers by Ruppert et al. [14, 13] and [11]. Nonparametric kernel regression has earlier been applied in physics in Lund concerning analysis of LIDAR (LIght Detection And Ranging) data, cf. Björklund [1] and Bratt [2].

3.1 The statistical model

In this thesis we model the received intensity signal $Y(x_i)$ at Raman shift x_i $[\frac{1}{cm}]$ as

$$Y(x_i) = m(x_i) + \sigma(x_i) \cdot \varepsilon(x_i), \tag{1}$$

where $\sigma(x_i)$ is the standard deviation of $Y(x_i)$, $\varepsilon(x_i)$ is a sequence of independent observations with expectation zero and variance one, and $m(x_i)$ is the expectation of $Y(x_i)$. The intensity $Y(x_i)$ will from here on be denoted by Y_i. Since our signal constitutes of both Raman and fluorescence signals we model $m(x_i)$ as

$$m(x_i) = m_R(x_i) + m_F(x_i), \tag{2}$$

where $m_R(x_i)$ is the Raman signal and $m_F(x_i)$ the fluorescence signal at Raman shift x_i. The solution to our problem will be the best possible estimation $\hat{m}_R(x_i)$ of $m_R(x_i)$. We estimate $\hat{m}_R(x_i)$ as

$$\hat{m}_R(x_i) = \hat{m}(x_i) - \hat{m}_F(x_i), \tag{3}$$

where $\hat{m}(x_i)$ is the estimation of the summarised spectrum (Raman + fluorescence), and $\hat{m}_F(x_i)$ is the estimation of the fluorescence spectrum.

Our problem can now be split into two parts:

1. Estimating $m(x_i)$: This is done using nonparametric kernel regression with local bandwidths.

2. Estimating $m_F(x_i)$: Also this is done with nonparametric kernel regression, but this time with a global and large bandwidth.

We start with a closer look at the nonparametric kernel regression method.

3.2 Nonparametric kernel regression

Decreasing noise in a set of data can be done through fitting a polynomial to it. This is a familiar technique but it can sometimes be too rigid. Fitting *local* polynomials to each point where we want to estimate the function might be more successful.

In a grid of points a polynomial

$$\beta_0 + \beta_1(x_i - x) + \cdots + \beta_p(x_i - x)^p \tag{4}$$

of degree p is fitted to the data (x_i, Y_i) through an optimisation of the weighted least squares criterion

$$\sum_{i=1}^{n}\{Y_i - \beta_0 - \beta_1(x_i - x) - \cdots - \beta_p(x_i - x)^p\}^2 K_h(x_i - x) \tag{5}$$

with respect to $(\beta_0, ..., \beta_p)$, where $K_h(x_i - x)$ is the kernel used (see below).

The procedure of minimising (5) is repeated over an arbitrary grid of x-values, not necessarily the measuring points x_i. Regarding (4) as a version of Taylors theorem, the first term β_0 will give us an estimate of the regression function and the second term β_1 the derivative in point x.

The weighting function is called a kernel and is denoted by $K(\cdot)$. The width h of the kernel affects the amount of data that will be considered when we estimate the function at point x. The kernel K is often chosen to be a unimodal density function. Two popular kernel choices are the Epanechnikov and the normal density functions, cf. Figure 6. They both satisfy $\int_{-\infty}^{\infty} K(u)du = 1$ and $\int_{-\infty}^{\infty} uK(u)du = 0$. In this thesis we use the Epanechnikov kernel. If $K(x)$ is our kernel then $K_h(x) = (1/h)K(x/h)$, where $h > 0$ is the bandwidth.

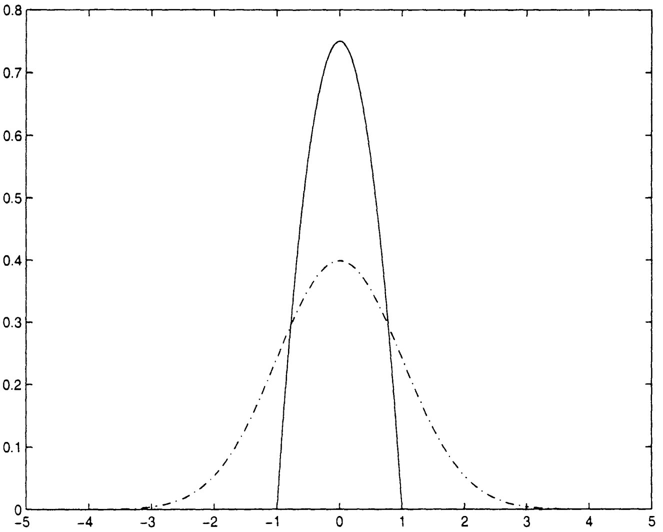

Figure 6: The Epanechnikov kernel, $K(x) = \frac{3}{4}(1 - x^2)$, $-1 \leq x \leq 1$ (solid) and the normal density kernel, $K(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$, (dashdotted).

The solution of (5) using matrix notation is the normal equations

$$[\hat{\beta}_0, ..., \hat{\beta}_p]^T = (X_{p,x}^T W_x X_{p,x})^{-1} X_{p,x}^T W_x Y, \tag{6}$$

where W_x is a matrix containing the weights, $X_{p,x}$ is the design matrix containing the measuring points and Y is a vector with the observations. More precisely $X_{p,x}$ is the matrix

$$X_{p,x} = \begin{bmatrix} 1 & (x_1 - x) & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - x) & \cdots & (x_n - x)^p \end{bmatrix},$$

$Y = (Y_1, \ldots, Y_n)^T$ and $W_x = \text{diag}\{K_h(x_1 - x), \ldots, K_h(x_n - x)\}$ is a diagonal matrix of weights; cf. Fan and Gijbels [5].

If we multiply (6) with the vector e_1, which is a $(p+1) \times 1$ vector having 1 in the first entry and all other entries zero, we get our estimate of the regression curve at point x as

$$\hat{m}(x, h, p) = \hat{\beta}_0 = e_1^T (X_{p,x}^T W_x X_{p,x})^{-1} X_{p,x}^T W_x Y. \qquad (7)$$

If we are interested in the derivative at point x we have to estimate β_1 as

$$\hat{m}^{(1)}(x, h, p) = \hat{\beta}_1 = e_2^T (X_{p,x}^T W_x X_{p,x})^{-1} X_{p,x}^T W_x Y. \qquad (8)$$

where e_2 is a $(p+1) \times 1$ vector having 1 in the second entry and all other entries zero.

Before proceeding further into the world of kernel regression we shall make two more statements about $\hat{\beta}_0$ and $\hat{\beta}_1$. Considering the estimation as a linear function of the observed data Y_i, we realise that the variance of $\hat{\beta}_0$ can be expressed as

$$V(\hat{\beta}_0) = e_1^T (X_{p,x}^T W_x X_{p,x})^{-1} X_{p,x}^T W_x V_y W_x X_{p,x} (X_{p,x}^T W_x X_{p,x})^{-1} e_1, \qquad (9)$$

where $V_y = \text{diag}\{\sigma^2(x_1), \ldots, \sigma^2(x_n)\}$ is a diagonal matrix which contains the variances of the measurements. In the case when we are estimating the first derivative the variance becomes

$$V(\hat{\beta}_1) = e_2^T (X_{p,x}^T W_x X_{p,x})^{-1} X_{p,x}^T W_x V_y W_x X_{p,x} (X_{p,x}^T W_x X_{p,x})^{-1} e_2. \qquad (10)$$

The discussion above of how to estimate the function $m(x)$ depends on the bandwidth h, so do the properties of the estimator $\hat{m}(x, h, p)$. But how shall we choose our bandwidth? How can we compare different estimators? These are very relevant questions which are treated in the next section.

4 Bandwidth selectors

When searching for suitable bandwidths to be used as described in Section 3.2, we must ask which features we are looking for. The question is:

- Are we searching for details, or rough structures?

The answer to this question will decide which of two possible approaches we should use. If we are searching for rough structures, then it is very important that we remove bumps from our function. For example, if we are looking for the fluorescence in a spectrum containing both Raman peaks and fluorescence, then we are essentially looking for the rough structures in the function. This can be achieved by using a large and global bandwidth; i.e. we use the same bandwidth over the whole data set. In Section 4.2 these ideas are discussed in more detail.

On the other hand, when we want to estimate the fluorescence plus Raman peaks in our spectra, it is important that we do not oversmooth the Raman peaks which contain the important information about the tissue. This estimate can be generated using local bandwidths. The method is developed by Ruppert [11, 12] and will be discussed in the next section.

4.1 The local bandwidth method

The local bandwidth selector employs a local bandwidth at each point of the x-grid, so that at each point the Mean Squared Error,

$$MSE(\hat{m}(x)) = E(\hat{m}(x) - m(x))^2 = V(\hat{m}(x)) + (E(\hat{m}(x)) - m(x))^2, \quad (11)$$

is minimised with respect to h. As the above equation shows the Mean Squared Error can be expressed as a sum of the variance of the estimator, $V(\hat{m}(x))$, and the squared bias, $(E(\hat{m}(x)) - m(x))^2$, where $E(\hat{m}(x))$ is the expected value of the estimator.

Both these terms contain unknown factors which have to be estimated. The variance part which is given in (9) contains the unknown variance function $\sigma^2(x)$ and the squared bias, for which we use an approximative expression, contains unknown polynomial coefficients. The estimation of the variance function will be dealt with in Section 4.1.1 and after that we discuss the estimation of the bias in Section 4.1.2.

4.1.1 Estimating the variance function

In order to calculate the variance of the estimator we have to estimate the unknown variance function $\sigma^2(x_i)$. We first observe that our model

$$Y(x_i) = m(x_i) + \sigma(x_i) \cdot \varepsilon(x_i),$$

contains the standard deviation $\sigma(x_i)$. By a normalisation of the squared residuals we can estimate $\sigma^2(x_i)$. This is a very intricate procedure and for a more careful description of the technical details, see Ruppert [11]. The squared residuals from our observations are formed as

$$e_i^2 = \{Y_i - \hat{m}(x_i, h_m, p)\}^2, \tag{12}$$

where $\hat{m}(x_i, h_m, p)$ is the estimate of the regression function using bandwidth h_m. This means that in order to calculate the squared residuals we have to use a pilot estimate of $\hat{m}$ which is calculated with a user supplied bandwidth h_m. The only restriction on h_m is that it is sufficiently small, or else it will render bias to the residuals.

Then we re-estimate $\hat{m}$ using local bandwidths where we use the residuals to estimate $\sigma(x_i)$. This procedure gives us a second estimate of $\hat{m}$ and using this estimate we can calculate new residuals. These new residuals are smoothed and they are then taken as the final estimate of the variance function $\sigma^2(x_i)$. The estimated variance function is then plugged into (9) and (10). This gives us an estimate of the variance of the estimator of the regression function and its derivative.

4.1.2 Estimating the bias

An asymptotic approximate expression for our estimate when estimating the k:th. derivative at point x using bandwidth h is equal to:

$$\hat{m}^{(k)}(x, h, p) \approx c_0 + c_{p+1-k}h^{p+1-k} + \ldots + c_{p+t-k}h^{p+t-k}, \tag{13}$$

as $h \to 0$ and $t \to \infty$ where t is the number of terms in the Taylor expansion for the bias; see Ruppert [11]. When estimating the function itself k is equal to zero and when estimating the first derivative k is equal to one.

The terms after the first one in (13) represent the bias, so

$$\widehat{BIAS}(x, h, p) \approx c_{p+1-k}h^{p+1-k} + \ldots + c_{p+t-k}h^{p+t-k}. \tag{14}$$

The c's are estimated through ordinary polynomial regression with the data set $\{(x, y) = (h_{0,j}, \hat{m}(x, h_{0,j}, p)) : j = 1, ..., J\}$. That is, in a neighbourhood of h_0, where we want to estimate the bias, we construct a sequence $\{h_{0,j} : j = 1, ..., J\}$ where J is chosen so that the number of terms is for example four or five. Then we evaluate $\hat{m}(x, h, p)$ with these bandwidths. We can now estimate our unknown bias coefficients $\{c_0, \ldots, c_{p+t-k}\}$ and also the approximate bias at each point. A thorough treatment of this procedure can be found in [11].

Having the estimates of the variance function and the bias we can express the estimated $MSE(x; h_0)$ as

$$\begin{aligned}\widehat{MSE}(x, h_0) = {} & [\hat{c}_{p+1-k}(x)h_0^{p+1-k} + \ldots + \hat{c}_{p+t-k}(x)h_0^{p+t-k}]^2 \\ & + e_k^T(X_{p,x}^T W_x X_{p,x})^{-1} X_{p,x}^T W_x \hat{V}_y W_x X_{p,x} (X_{p,x}^T W_x X_{p,x})^{-1} e_k,\end{aligned} \tag{15}$$

where k is the order of the derivative at point x using bandwidth h_0 and $\hat{V}_y = \text{diag}\{\hat{\sigma}^2(x_1), \ldots, \hat{\sigma}^2(x_n)\}$. This is a rough function of x and therefore we smooth this first version of the $\widehat{MSE}$ to get a $\widehat{SMSE}$; i.e. an estimation of the Smoothed MSE. The optimal bandwidth is then chosen as the one which minimises $\widehat{SMSE}$.

The result of this bandwidth selection is that we in some sense have got an adaptive bandwidth which adjusts itself from point to point in order to follow the unknown function as well as possible. Two real benefits with this method are that we have got an approximate estimate of the bias and estimates of the variances of the regression function and its derivatives.

4.2 The SiZer approach

SiZer (Significant Zero crossings of the derivative) is a method which contrary to the traditional regression approach does not focus on the true underlying curve. The main goal is now to study a wide range of global bandwidths and for each bandwidth ask the question "which features/peaks are really there?". The method and the MATLAB code is developed by Chaudhuri and Marron [3].

The SiZer approach departs from the classical approach in two ways. Firstly we avoid the need of choosing a particular optimal bandwidth. Instead we simultaneously study a wide range of bandwidths. And secondly, the bias problem is ignored through shifting focus from the true underlying curve to the curve studied at different bandwidths.

The idea is closely related to the scale space ideas from image analysis since different bandwidths bring out different features in the curve. In computer vision where we for instance apply the wavelet transform to a picture, different scale space levels enhance different features. In our case the different bandwidths reveal different features, so the two areas certainly coincide.

4.2.1 The SiZer map

Our main idea is to create a picture which reveals the properties of estimators generated by different bandwidths, all at the same time and in the same picture; i.e. we want to know which bandwidth suppresses or enhances different peaks. To do this we have to continue the analysis and consider the zero crossings of the first derivative of the regression function.

The peaks are characterised by the fact that the derivative is zero on top of the peak and significantly different from zero on both sides with opposite signs. This can be tested by constructing confidence intervals for the first derivative. By defining $m^{(1)}(x,h,p)$ to be $E\hat{m}^{(1)}(x,h,p)$ we have ignored the bias problem. At point x a confidence interval for $m^{(1)}(x,h,p)$ is constructed as

$$I_{m^{(1)}(x,h,p)} = \hat{m}^{(1)}(x,h,p) \pm q \cdot d(\hat{m}^{(1)}(x,h,p)), \qquad (16)$$

where $d(\cdot)$ is the estimated standard deviation of $\hat{m}^{(1)}(x,h,p)$ and q an approximate quantile which will be discussed later on.

There are two differences in the estimation procedures between this method and the method discussed in Section 4.1. One minor difference is that the normal density kernel is used in the SiZer approach and the Epanechnikov kernel in the local bandwidth selector. When estimating the regression function itself the relation between optimal global bandwidths for the normal density kernel and the Epanechnikov kernel is $1.719 \cdot h_{normal} = 0.776 \cdot h_{Epan}$. The bandwidths in the SiZer maps in this thesis are rescaled to those of the Epanechnikov kernel. Secondly, both methods use local smoothing of squared residuals when estimating $\sigma^2(x)$. The methods are slightly different though, for further details see [11] and [3].

If the interval $I_{m^{(1)}(x,h,p)}$ is above zero the derivative is significantly different from zero with a positive sign (+), vice versa for the opposite case when the interval lays below zero (−). The third case is when the interval contains zero, then the derivative is not significantly different from zero. The

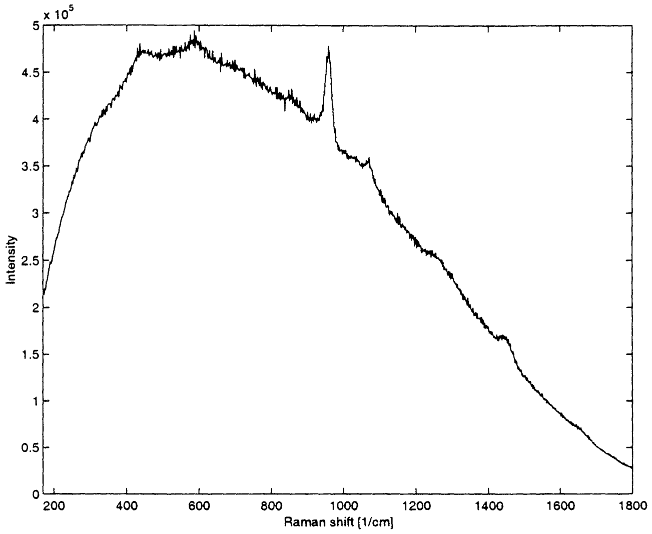

Figure 7: Example of a measured spectrum. (Integration time 100 s.)

three cases imply that the function is increasing (↑), decreasing (↓) or non-determined (→). Actually the SiZer method deals with a fourth case and that is when there is a lack of data in areas where the estimation takes place.

The SiZer map marks these four cases with colours, see Table 1. Figure 8 gives an example of a SiZer map when SiZer is applied on our reference data, shown in Figure 7. Note that there exists no area where there are too few measurements.

colour	dark	light	gray	darker gray
$\hat{m}(x,h,p)$	↑	↓	→	too few data
$\hat{m}^{(1)}(x,h,p)$	+	-	0	

Table 1: Description of the colours in a SiZer map.

The quantile q in (16) can be chosen according to four different approaches. The first one is based on pointwise Gaussian quantiles, the second one and the third one are two variants of a bootstrap method and the fourth, which is used in this thesis (and in practice) since it is faster to compute, is based on approximate simultaneous Gaussian quantiles. Though, according to [3], if there is any doubt about what really is there, the two bootstrap methods should be tested.

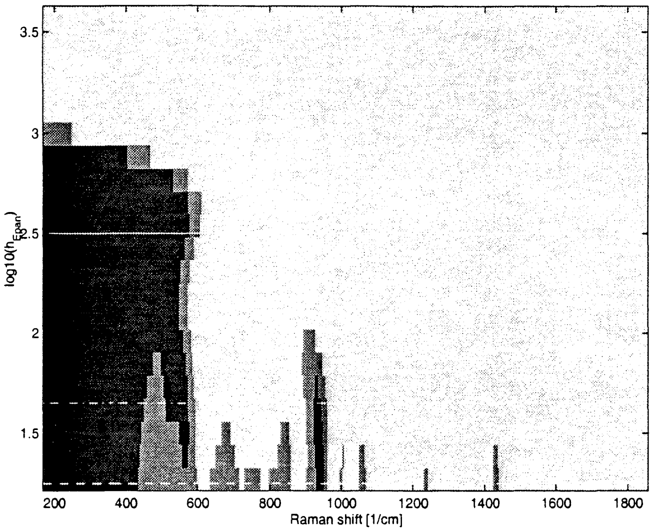

Figure 8: The SiZer method applied on measurements in Figure 7. The lines are $log(h_{Epan}) = 2.5$ (solid), $log(h_{Epan}) = 1.65$ (dashdotted) and $log(h_{Epan}) = 1.25$ (dashed).

4.2.2 How to use the SiZer map

The SiZer map is a powerful tool when we are searching for peaks; i.e. when we are bump hunting. When comparing Figure 8 to Figure 7 we realise that the large peak at 958 cm^{-1} is significant only for bandwidths up to about $h = 10^2 = 100$. For larger bandwidths the peak is smoothed away. When h is large enough $I_{m^{(1)}(x,h,p)}$ lays below zero for all x which implies that our function is decreasing; i.e. using large bandwidths will remove all features and bring out only the slowly varying parts. This is illustrated in Figure 9 where we have plotted the regression curve for three different global bandwidths. As can be seen there are no peaks when using the bandwidth $h_{Epan} = 10^{2.5} = 316$, but as we decrease the bandwidth more and more peaks become apparent.

The conclusion is that when we are using small bandwidths many details become visible but when increasing the bandwidth these details disappear and we will only see the slowly varying parts of the function. This fact will be used when we estimate the fluorescence curve in the next section.

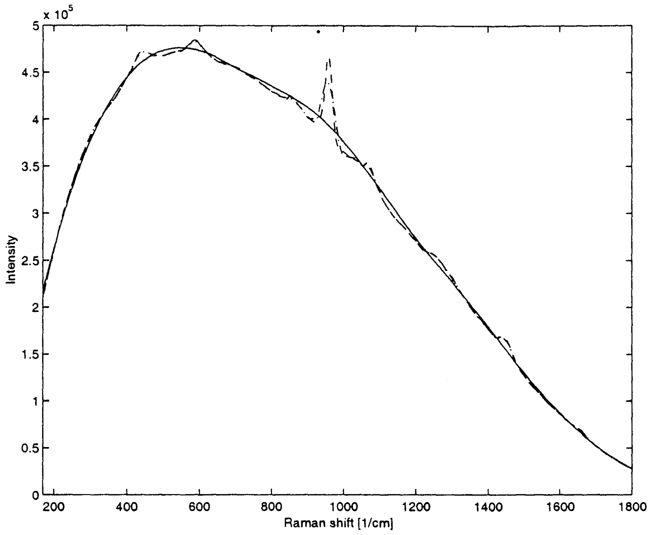

Figure 9: The regression curves when using the global bandwidths: $h_{Epan} = 10^{2.5} = 316$ (solid), $h_{Epan} = 10^{1.65} = 44.67$ (dashdotted) and $h_{Epan} = 10^{1.25} = 17.78$ (dashed).

5 Estimation of summarised and fluorescence spectra

In Section 3.1 we concluded that in order to estimate the Raman spectrum we need estimates of both the summarised spectrum and the fluorescence spectrum, that is

$$\hat{m}_R(x_i) = \hat{m}(x_i) - \hat{m}_F(x_i),$$

where $\hat{m}(x_i)$ is the estimate of the summarised spectrum and $\hat{m}_F(x_i)$ is the estimate of the fluorescence spectrum. These estimators will be treated in this section. From here on all our estimates will be calculated in the observations points x_i only.

In order to calculate a good estimate of the fluorescence spectrum we have to remove the Raman peaks, which otherwise will influence the fluorescence estimation. This can be done using robust statistical methods which point out areas where the spectrum is larger than expected and replace these observations with upper limits, cf. Lindström [6].

Here we have developed two other methods in order to remove the Raman peaks. They both require an estimate of the summarised spectrum $m(x)$

which is discussed in Section 5.1. When this is done we have two alternatives of how to select the fluorescence points and they are both discussed in Section 5.2. Having these points, we can estimate the fluorescence spectrum and this is discussed in Section 5.3.

5.1 Estimation of summarised spectra

In order to remove the Raman peaks we need an estimate of the summarised spectrum $m(x)$. This is calculated using the local bandwidth method described in Section 4.1. We have chosen this method since the estimate $\hat{m}(x)$ will follow the original spectrum closely.

According to (7) and (9) the estimator of the summarised spectrum and the variance of this estimator are

$$\hat{m}(x,h,p) = e_1^T(X_{p,x}^T W_x X_{p,x})^{-1} X_{p,x}^T W_x Y, \tag{17}$$

$$V(\hat{m}(x,h,p)) = e_1^T(X_{p,x}^T W_x X_{p,x})^{-1} X_{p,x}^T W_x V_y W_x X_{p,x}(X_{p,x}^T W_x X_{p,x})^{-1} e_1, \tag{18}$$

where $p = 2$. The kernel matrix W_x does not only depend on x but also on the local bandwidths. The local bandwidth method gives us an estimate of the variance function, $\sigma^2(x)$, as well as an estimate of the bias, cf. (14). The variance function and the bias of our estimated reference spectrum are shown in Figure 10.

We estimate a larger variance in the peak neighbourhoods. This is due to the fact that our estimators can not follow the curve closely enough at these points. This is also is revealed in the bias plot where the bias is large in the peak neighbourhoods.

Figure 11 (a) shows the estimated regression function when applying the local bandwidth method on our reference spectrum. The estimated variance function $\sigma^2(x)$ is directly used in (18) where V_y is replaced with $\hat{V}_y = \text{diag}\{\hat{\sigma}^2(x_1), \ldots, \hat{\sigma}^2(x_n)\}$, and this equation gives us an estimate of the variance of our regression function, cf. Figure 11 (b).

Figure 12 shows the sequence of bandwidths which were used when estimating the summarised spectrum. Evidently the bandwidth decreases when we enter a neighbourhood of a peak. This is most obvious around the largest peak at 958 cm^{-1}. The decrease in bandwidth is explained by the fact that we aim at enhancing the peak. The use of a large bandwidth implies that we consider data points which are not in the immediate neighbourhood of the

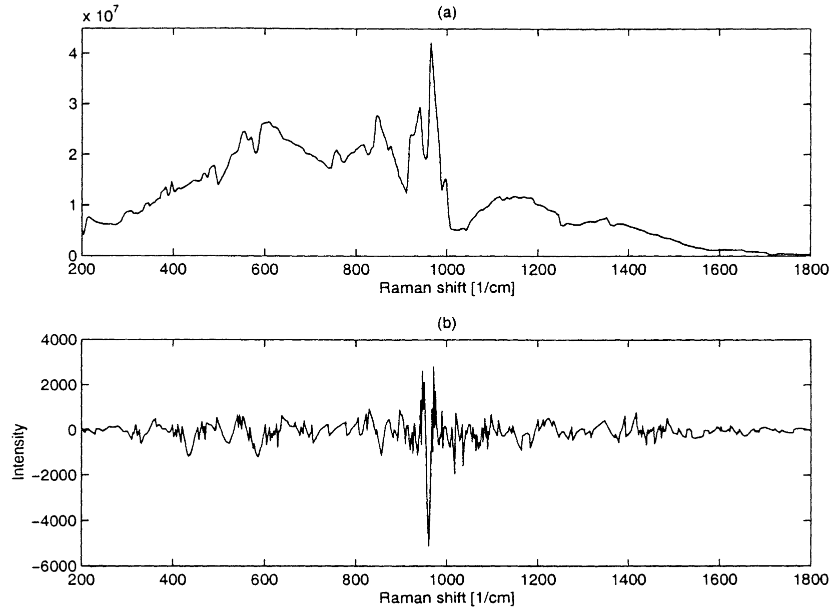

Figure 10: (a) Estimated variance function $\sigma^2(x)$ and (b) estimated bias in our reference spectra.

peak and these data points will contribute to the fading of the peak.

5.2 Selection of fluorescence points

We have developed two methods which will remove the Raman peaks and keep only those points where the fluorescence is undisturbed. The first method is the most obvious. We simply let the user herself decide which regions will be used when estimating the fluorescence spectrum. The idea behind this method is that the physicist often knows where the suspected Raman peaks are located, though she hardly can see them. The manual method is illustrated in Figure 13 (a).

The second method is an automatic data analysis approach. It starts with a pilot estimation of the fluorescence curve $\hat{m}_{P,F}(x)$ using a large global bandwidth and $p = 1$. The bandwidth might be decided from the SiZer map. As an alternative to this we can choose the largest bandwidth of the sequence of local bandwidths which were calculated when the summarised spectrum was estimated. In this thesis we make use of the latter choice. By choosing a large global bandwidth we will smooth away all Raman peaks as described

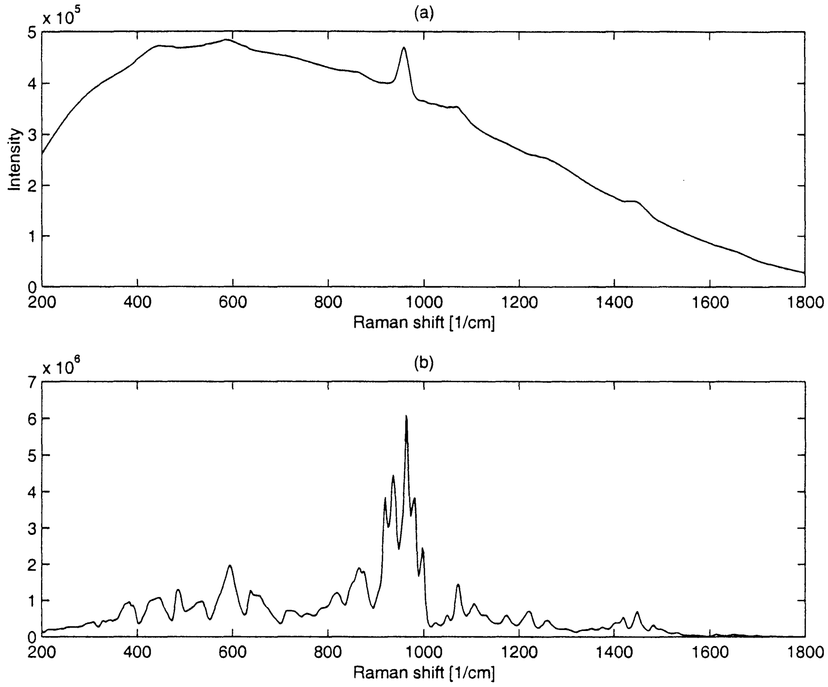

Figure 11: (a) The estimated reference spectrum, $\hat{m}(x, h, p)$, using the local bandwidth method and (b) the estimated variance.

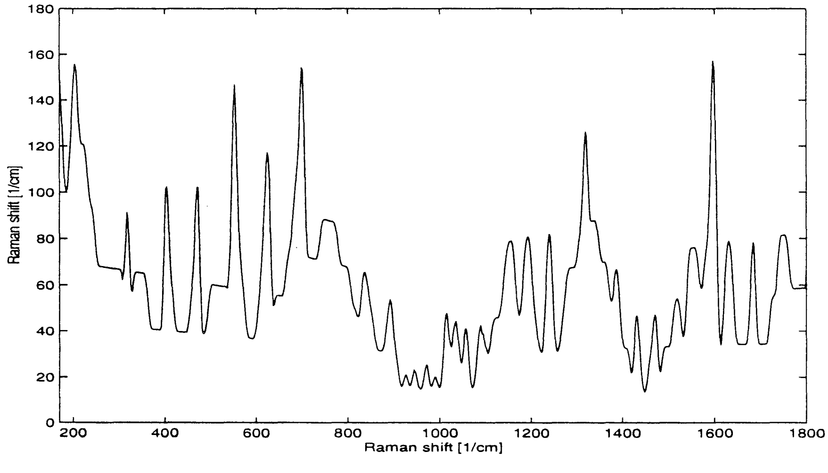

Figure 12: The local bandwidths which are used when estimating the summarised spectrum.

in Section 4.2.

Having this pilot estimation $\hat{m}_{P,F}(x)$ and the estimation of the summarised spectrum $\hat{m}(x)$ a pilot estimation of the Raman spectrum can be calculated as

$$\hat{m}_{P,R}(x) = \hat{m}(x) - \hat{m}_{P,F}(x), \tag{19}$$

where P denotes that this is a pilot estimation. We then check where the difference is less than some fraction (1/10 in Figure 13) of the greatest difference. The points where this is fulfilled is chosen to be our fluorescence points. This procedure, which is illustrated in Figure 13 (b), brings us to approximately the same state as the first method.

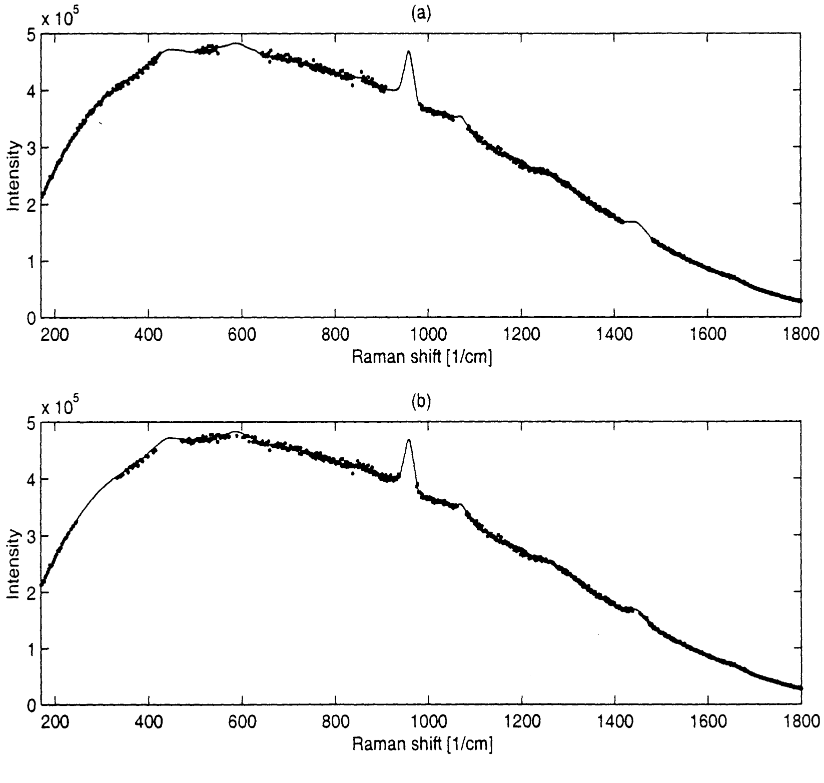

Figure 13: Fluorescence points chosen accordingly to the two methods. (a) manual procedure, (b) automatic procedure.

As Figure 13 shows, there is no big difference between the two procedures for this data set. Though, we have to remember that this may not be a typical spectrum, and in reality we have to decide from time to time which measuring points should be included.

5.3 Estimation of fluorescence spectra

With the chosen fluorescence data points we estimate the fluorescence spectrum. This is done with the aid of the nonparametric kernel regression method described in Section 3.2. We estimate the fluorescence as

$$\hat{m}_F(x,h,p) = e_1^T (X_{p,x}^T I W_x X_{p,x})^{-1} X_{p,x}^T I W_x Y, \tag{20}$$

where $p = 1$ and h is a large global bandwidth. It is crucial that the bandwidth is large enough so that we can bridge the gaps, or else the estimate at some of those points will be equal to zero. I is a diagonal matrix whose elements indicates which points should be included when doing the estimation. More precise $I = \text{diag}\{i_1, i_2, \ldots, i_n\}$ where

$$i_l = \begin{cases} 0 & \text{if the } l\text{:th point is not a fluorescence point,} \\ 1 & \text{if the } l\text{:th point is a fluorescence point.} \end{cases}$$

The estimated fluorescence curve is shown in Figure 14 (a). Though using $p = 1$ results in a somewhat piecewise linear estimate, we will use $\hat{m}_F(x,h,p)$ as our fluorescence estimate when we calculate the variance of the fluorescence. The variance of the fluorescence spectrum is calculated as

$$V(\hat{m}_F(x,h,p)) = e_1^T (X_{p,x}^T I W_x X_{p,x})^{-1} X_{p,x}^T I W_x V_y I W_x X_{p,x} (X_{p,x}^T I W_x X_{p,x})^{-1} e_1, \tag{21}$$

where $X_{p,x}$, W_x and V_y are the same as in (9). The estimated variance is shown in Figure 14 (b). In this figure we have chosen the fluorescence points with the manual procedure.

In this thesis we do not deal with an estimate of the bias for the fluorescence estimator. The fact that there are no peaks to be estimated suggests that the bias of the fluorescence estimate is approximately equal to zero and might be ignored. The reason for this is that the bias is approximately equal to

$$\frac{m_F^{(2)}(x)}{2} h^2 \int_{-\infty}^{\infty} u^2 K(u) du,$$

when $p = 1$. Since the fluorescence curve is slowly varying the second derivative function, $m_F^{(2)}(x)$, is almost zero. This implies that the bias of the fluorescence curve might be ignored.

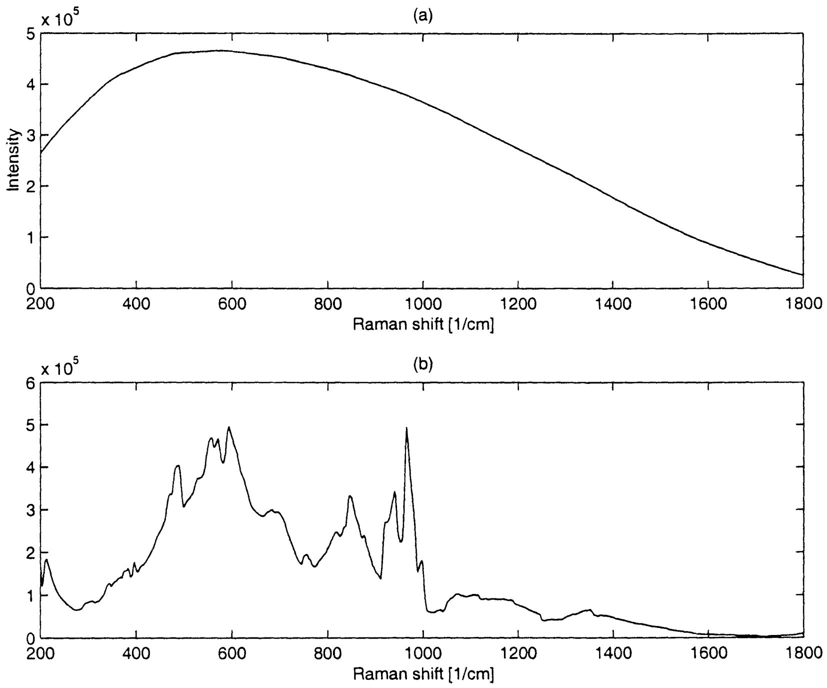

Figure 14: (a) The estimated fluorescence spectrum $\hat{m}_F(x, h, p)$ and (b) the estimated variance of the fluorescence estimator, $\hat{V}(\hat{m}_F(x, h, p))$.

5.4 A simulated fluorescence spectrum

If we want to, we can adjust the spectrum in such a way that it does not look too linear. This is done by smoothing a simulated fluorescence spectrum. The simulation is done by adding a noisy signal to $\hat{m}_F(x)$, that is:

$$\textit{simulated fluorescence} = \hat{m}_F(x) + \hat{\sigma}(x) \cdot \varepsilon(x)$$

where $\hat{\sigma}(x)$ is the square root of the estimated variance function, which was calculated when we estimated the summarised spectrum, and $\varepsilon(x)$ is a sequence of Gaussian independent observations with expectation zero and variance one.

Figure 15 shows us how a smoother fluorescence spectrum is developed by disturbing the piecewise linear fluorescence spectrum and then by smoothing this disturbed spectrum which is taken as our smoothed estimate of the fluorescence spectrum.

The simulated fluorescence spectrum can be used to get a smoother result when we subtract the fluorescence from the summarised spectrum as in (3).

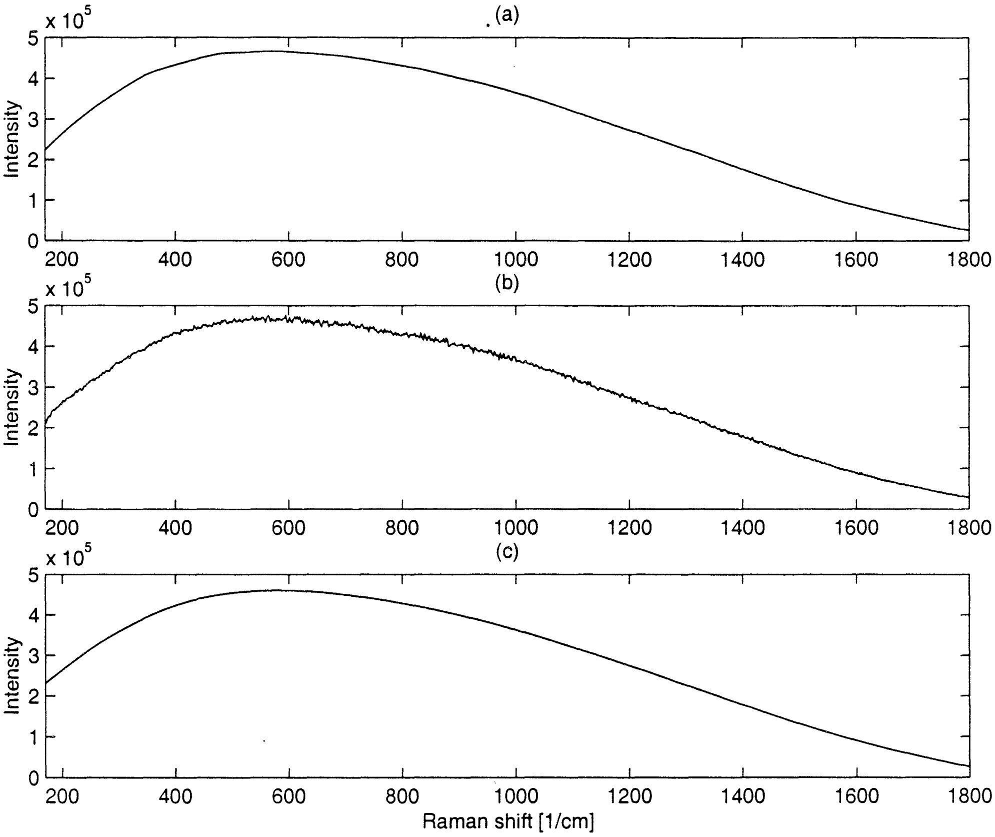

Figure 15: Example of how a smoother fluorescence curve comes up. (a) The piecewise linear fluorescence spectrum, (b) the simulated fluorescence spectrum and (c) the smoothed fluorescence spectrum. (The fluorescence points are chosen through the manual procedure.)

In this example there is no significant difference between the two fluorescence curves, this depends of course on the spectrum at hand.

6 Estimation of Raman spectra

This section will discuss the estimation of the Raman spectrum, which has been our main goal. In Section 3.1 we stated that the Raman spectrum is the difference between the summarised spectrum and the fluorescence. This imposed the following model:

$$\hat{m}_R(x) = \hat{m}(x) - \hat{m}_F(x),$$

where $\hat{m}(x)$ is the estimated summarised spectrum and $\hat{m}_F(x)$ is the estimated fluorescence spectrum. The difference between these two will give us the estimated Raman spectrum $\hat{m}_R(x)$. The estimated Raman spectrum in our reference spectrum is shown in Figure 16 (b).

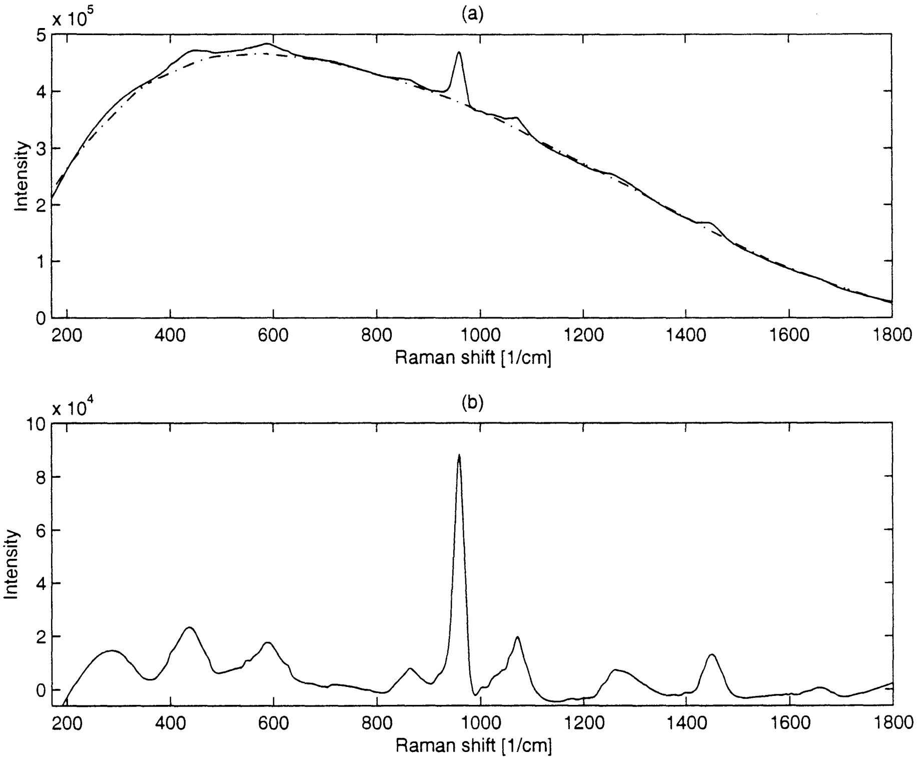

Figure 16: (a) The estimated summarised spectrum (solid) and the estimated fluorescence spectrum (dashdotted) and (b) the estimated Raman spectrum. Note the difference in intensity.

In the previous section we developed the tools needed to estimate the fluorescence spectrum. Having the estimators $\hat{m}(x)$ and $\hat{m}_F(x)$ and their estimated bias and variance, we can continue with the estimation of the Raman spectrum. The major effort will be spent on how to compute the variance of the Raman spectrum, since this consists of both the variances of the summarised and the fluorescence spectrum and the covariance between these two. This will be dealt with below.

by this fact. For example there may be weak peaks for which the derivative is zero between two areas where the derivative is negative. The weak peak area is defined as an area where the derivative changes sign from zero to negative.

According to these definitions the peak areas in Figure 21 are shown in Table 2. The peak at 958 cm^{-1}, which arises from hydroxyapatite $Ca_{10}(OH)_2(PO_4)_6$, should be located at 960 cm^{-1} accordingly to the literature. In the original data sequence the largest observation is at 958.58 cm^{-1} and the proposed peak interval is 954.5 - 958.6 cm^{-1}, cf. Table 2. It seems as if the peak area is biased. However, note that the definition of peak areas is based on simple pointwise confidence intervals for strongly correlated estimates. The statistical analysis might be improved using some form of simultaneous quantiles, cf. Marron [4].

Strong peak areas	Weak peak areas
419.8 - 440.6	284.3 - 286.7
561.3 - 565.8	445.5 -447.5
581.6 - 595.0	595.0 - 597.3
843.7 - 847.9	721.0 - 723.4
954.5 - 958.6	738.5 - 740.6
1064.0 - 1070.0	869.0 - 871.1
1272.4 - 1279.8	1849.8 - 1851.1
1445.0 - 1450.2	
1640.1 - 1652.6	
1771.2 - 1782.7	

Table 2: The different peak areas in Raman shift [1/cm].

8 Conclusions

This thesis indicates that the local polynomial regression method is useful when rejecting fluorescence in Raman spectroscopy and when estimating Raman spectra. Contrary to previous techniques, cf. [7] and [9], this method gives us an estimate of how large the variance and the bias are. They were both found to be larger in the peak neighbourhoods. Entering a peak area implies a larger variation in our spectrum and here it is more difficult to estimate the function.

Since we aimed at investigating the possibility of using nonparametrical kernel regression methods when estimating Raman spectra, and not at searching for the optimal parameters to be used, the results shown here might be improved. There are several parameter choices which have to be considered. Two of them are the polynomial degrees used when estimating the summarised and fluorescence spectra. Also the amount of smoothing used in the different steps of the algorithms should be further considered.

In this thesis we used $p = 2$ when estimating the summarised spectrum since there is a lot of curvature due to the Raman peaks. When we estimated the fluorescence we used $p = 1$ since this spectrum does not contain any peaks and is a slowly varying curve. Are these assumptions good, or do there exist even better choices?

The SiZer map proved to be a helpful tool when estimating the fluorescence spectrum. It can also be used when localising Raman peaks. In this thesis we developed a similar method in order to estimate the peak areas. Note that we do not estimate the derivative by taking the derivative of the estimator of the spectrum, but with β_1, cf. (8), which is a better estimator of the derivative.

Another way to improve the peak areas and the confidence intervals upon which the estimates are based, is to improve the statistical analysis using some form of simultaneous quantiles, cf. Marron [4].

A challenge for future research is to find appropriate confidence regions for simultaneous estimators of peak position and peak height.

References

[1] C. Björklund. Analysing LIDAR-measurements by the method of locally weighted least squares kernel regression (in Swedish). Master's thesis, Department of Mathematical Statistics, Lund University, Lund, Sweden, 1994.

[2] H. Bratt. Comparison of global and local bandwidth selectors for analysis of lidar curves. Master's thesis, Department of Mathematical Statistics, Lund University, Lund, Sweden, 1998.

[3] P. Chaudhuri and J. Marron. Sizer for exploration of structures in curves. October 1997.

[4] S. Chung and J. Marron. Presentation of smoothers: the family approach. unpublished manuscript, 1997.

[5] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications.* Chapman and Hall, 1996.

[6] T. Lindström. Robust local polynomial regression with application on the doas filtering process. Master's thesis, Department of Mathematical Statistics, Lund University, Lund, Sweden, 1996.

[7] A. Mahadevan-Jansen, M. Follen Mitchell, N. Ramanujam, A. Malpica, S. Thomsen, U. Utziger, and R. Richards-Kortum. Near-infrared raman spectroscopy for in vitro detection of cervical precancers. *Photochemistry and Photobiology*, 1998.

[8] A. Mahadevan-Jansen and R. Richards-Kortum. Raman spectroscopy for the detection of cancers and precancers. *Journal of Biomedical Optics*, 1(1):31–70, 1996.

[9] P. Moiser-Boss, S. Liebermann, and R. Newbery. Fluorescence rejection in raman spectroscopy by shifted-spectra, edge detecton, and fft filtering techniques. *Applied Spectroscopy*, 49(5):630–638, 1995.

[10] S. Pålsson. Development of a fiber-optical probe for near-infrared raman spectroscopy. Master's thesis, Department of Physics, Lund Institute of Technology, Sweden, 1997.

[11] D. Ruppert. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, 92:1049–1062, 1997.

[12] D. Ruppert. Local ploynomial regression and its applications in enviromental statistics. In V. Barnett and K. Feridun Turkman, editors, *Pollution Assessment and Control*, pages 155–173. John Wiley Sons, 1997.

[13] D. Ruppert, S. Sheather, and M. Wand. An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*, 90:1257–1290, 1995.

[14] D. Ruppert and M. Wand. Multivariate locally weighted least squares regression. *Ann. Statistics*, 22:1346–1370, 1994.

[15] S. Svanberg. *Atomic and Molecular Spectroscopy*, volume 6. Springer Series on Atoms and Plasmas, 2nd edition, 1992.

[16] M. Wand and M. Jones. *Kernel Smoothing*. Chapman Hall, first edition, 1995.

University of Lund with Lund Institute of Technology 1998:E16
Department of Mathematical Statistics ISSN 0281-1944
ISRN: LUTFD2/TFMS–5066–SE