NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

DEPARTMENTT OF ECONOMICS

Master in Business Administration, Analytics and Information Systems

Applications of the best piecewise monotonic approximation to peak estimation of NMR data

Maria Rafaela I. Vogiatzi

Supervisor: Professor Ioannis C. Demetriou

ATHENS, 2019

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΠΜΣ ΔΙΟΙΚΗΣΗ, ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΕΠΙΧΕΙΡΗΣΕΩΝ

Εφαρμογή της κάτα τμήματα μονοτονικής προσέγγισης σε δεδομένα NMR για την εκτίμηση κορυφών

Μαρία Ραφαέλα Βογιατζή

Επιβλέπων Καθηγητής: Ιωάννης Δ. Δημητρίου

Αθήνα, 2019

ABSTRACT

Peak estimation problems appear inherently in spectroscopy. In nuclear magnetic resonance (NMR) spectroscopy, for example, magnetic resonance is based on matter property by analyzing the sample at a specific frequency of radio waves. An electrical signal is produced and is shown by the height of the peaks. NMR spectroscopy is the most common vibrational spectroscopy techniques for assessing molecular motion and fingerprinting species. For these spectra, the location of peaks and their intensities are the signature of a sample of an organic or an inorganic compound or a tissue.

Piecewise monotonic data approximation method makes the smallest change to the data such that the first differences of the smoothed values change sign a prescribed number of times. The algorithms that have been developed for this challenging combinatorial calculation are very efficient providing optimal solutions in quadratic complexity with respect to the number of data. We are going to investigate the efficiency and the efficacy of the piecewise monotonic data approximation method to peak estimation of NMR data that will be received from metabolomics data bases.

Our intention is to present results of the piecewise monotonic method on a muter of NMR spectrum. Therefore, our results are helpful in showing that piecewise monotonic method is particularly suitable for peak estimation in NMR data.

Subject Area: Data Analysis

Keywords: Algorithm L2WPMA, Piecewise monotonic data approximation method, Nuclear magnetic resonance (NMR)

ΠΕΡΙΛΗΨΗ

Το πρόβλημα εκτίμησης κορυφών εμφανίζεται εγγενώς στη φασματοσκοπία. Στη φασματοσκοπία πυρηνικού μαγνητικού συντονισμού (NMR), για παράδειγμα, ο μαγνητικός συντονισμός βασίζεται στην ιδιότητα της ύλης ανάλλογα μέσω της έκθεσης του δείγματος σε συγκεκριμένη συχνότητα ραδιοκυμάτων. Παράγεται ηλεκτρικό σήμα και φαίνεται από το ύψος των κορυφών. Φασματοσκοπία NMR είναι μια από τις πιο κοινές τεχνικές φασματοσκοπίας μοριακής δόμησης για την αξιολόγηση των μοριακών κινήσεων σε πολύ υψηλές στάθμες δόνησης και των ειδών δακτυλικών αποτυπωμάτων. Για αυτά τα φάσματα, η θέση των κορυφών και οι εντάσεις τους είναι η ένδειξη μιας οργανικής ή ανόργανης ένωσης ή ιστού. Η μέθοδος προσέγγισης τμηματικών μονοτονικών δεδομένων σε σειρά κάνει τη μικρότερη αλλαγή στα δεδομένα, έτσι ώστε οι πρώτες διαφορές των αλλαγών των εξομαλυνσμένων τιμών να είναι ανάλογες ενός προκαθορισμένου αριθμού. Οι αλγόριθμοι που έχουν αναπτυχθεί για αυτόν τον δύσκολο συνδυαστικό υπολογισμό είναι πολύ αποτελεσματικοί παρέχοντας βέλτιστες λύσεις σε τετραγωνική πολυπλοκότητα σε σχέση με τον αριθμό των δεδομένων. Πρόκειται να διερευνήσουμε την αποδοτικότητα και την αποτελεσματικότητα της μεθόδου κατά προσέγγιση μονοτονικών δεδομένων για την εκτίμηση κορυφών των δεδομένων NMR που θα ληφθούν από τις βάσεις δεδομένων που σχετίζονται με τον μεταβολισμό

Στόχος μας είναι να παρουσιάσουμε τα αποτελέσματα της τμηματικής μονοτονικής μεθόδου σε ένα φάσμα NMR. Επομένως, τα αποτελέσματά μας είναι χρήσιμα για να δείξουμε ότι η τμηματική μονοτονική μέθοδος είναι ιδιαίτερα κατάλληλη για εκτίμηση αιχμής σε δεδομένα NMR.

Θεματική Περιοχή: Ανάλυση Δεδομένων

Λέξεις κλειδιά: Αλγόριθμος L2WPMA,Κατά Τμήματα Μονότονη Προσέγγιση Δεδομένων, Πυρηνικός μαγνητικός συντονισμός (NMR).

Table of Contents

Image

Figure

Table

Chapter I

Introduction

This chapter consists of two parts. In the section 1.1 refers to spectroscopy, in particular how it appeared, what its applications, its advantages and its structure. Spectroscopy is a branch of Physics and especially of Optical or Wave Visual which deals with the study and study of the structure, composition and properties of the spectra of matter as well as the various radiations. We will analyze spectroscopy types such as Magnetic Resonance, NMR, IR, Raman, we will present some fundamental information about their function and we will discuss the reasons why we choose a spectroscopy. In the section 1.2 we will see how through IR, NMR and Proton magnetic resonance spectroscopy we can get useful information about the structure, and the function of the proteins depending on the peaks created in the spectra. In this way we will obtain information about the structure of protein that can then lead us to treat diseases.

1.1. Spectroscopy

Historical review -Spectroscopy is defined as a research of the interaction between the material and electromagnetic radiation. The representation can be done either on a diagram, on a computer monitor, or on a list of wavelengths and intensities. Spectroscopy began in the 1660s when Hook formulate in a book that light is a longitudinal wave. As early as 1666, Newton was the first one who had succeeded in experimenting with the spectrum of a beam of light with the insertion of a transparent prism and lead to the creation of a multicolor elongated multi-color bundle. Three optical phenomena, reflection, refraction, and light analysis are detected during the passage of a light beam from the prism. Newton concluded with this experiment that the light consists of particles traveling in a straight line (Nicholas C. Thomas, 1991).

Spectroscopy is the interaction of photons with material whose outcome gives us the energy of the photon. This is derived from the formula $E = h\nu$, whereas the constant of Planck is meant h=*6,626068* $\times 10^{-34}$ $m^2_{kg/s}$ and ν the frequency of the photon.

At that time it was a great deal of confusion if the light eventually was a wave or a particle. Light depending on the conditions would behave either as a wave or as a particle but not both. In most cases it acts like a wave with oscillating electric and magnetic fields. The principle of electromagnetism is based on the motion of charged particles with spin which generate electric fields, and electric fields in turn cause motion of the charged particles, thus generating electricity. The movement of the charged particles creates a magnetic field in a direction perpendicular to

plane of the load movement. In this way an electromagnetic field is created which propagates in the space in the form of electromagnetic radiation, like a wave that can be propagated in the vacuum at the speed of light without the interference of some material. (Ball D. W. 1962)

$$E=A_e\ sin(\frac{2\pi z}{\lambda}-2\pi vt), \quad (1)$$

$$B=A_b\ sin(\frac{2\pi z}{\lambda}-2\pi vt) \quad (2)$$

- As *E* is the electric field and *B* is the magnetic field.
- As A_e and A_b are meant the backs of the fields.
- As *L* is the wavelength of the light.
- As *v* is the frequency of light.
- As *T* is understood to mean time and *z* is the axis that we assume the light propagates

The reason $\frac{A_e}{A_b} = c$, which is the velocity of light that equals $c = \lambda v$ (3)

- As **λ** is the wavelength
- As **v** is meaning the frequency.

Spectroscopy has expanded to a great extent and now links light intensity to frequency because it is related to light analysis at individual frequencies and records the intensity in each of them. This phenomenon results from the displacement of electrons from an active state to another in atoms, molecules and ions, explaining the stability of material. With transitions of electrons to permissible levels, we can draw information on composition, structure and properties of material. Spectroscopy found application in scientific fields of astronomy and biomedicine. The creation of the MRI of a tool based on its origins on the principle of spectroscopy has contributed to the diagnosis of diseases. In particular, with the help of the biomedical spectra, we analyze patterns and dysfunctions for the diagnosis of brain tumors, depression and even changes in the structure of the muscles (Ball D. W. 1962).

1.1.1. Magnetic resonance spectrometers

Magnetic resonance based on the magnetic properties of some items, because most people in the periodic table have at least one isotope whose angular momentum is not zero and has a bipolar magnetic moment in the same direction. Moreover, when one core having a magnetic moment and a nonzero spin torsion C torque, which is responsible for varying the angular momentum q of, found external magnetic field B, the magnetic moment perform angular precession around the axis of the magnetic field, with a frequency proportional to the field .Initially we need to mention that individuals have the ability to rotate around their axis and this phenomenon is called self-turning. Of course the particles do not rotate on a single axis. (Ζαφειρόπουλος & Βασίλειος 2015). The total angular speed of rotation depends on the quantum

number S for the electrons, the quantum number I for the nucleus and the direction where it shows us the direction the electron is rotating (quantum number mS) and the kernel (quantum number mI).

For the cores in accordance with quantum mechanics, the quantum q m I number can take values *-I, -I + 1, ..., +I* giving a total *2L + 1* energy levels which are equal to each other. The electrons have a rotation S = ½ or S = - ½ .The first value indicates that the spin of the electron is homogeneous to B which has spin ½ and is called basic energy state (up). The second value shows that the rotation of the electron is anti-parallel to B which has a -½ spin and is called a down state. In the basic energy state they appear more than stimulated due to its lower energy.

$$m=\frac{1}{2} \rightarrow I_z = \frac{h}{2} \rightarrow E_{up} = -\frac{\gamma h B_o}{2}$$

$$m=-\frac{1}{2} \rightarrow I_z = -\frac{h}{2} \rightarrow E_{down} = \frac{\gamma h B_o}{2}$$

- As ΔE is the difference between the two energy states and equals:

$$\Delta E = E_{down} - E_{up} = \gamma h B_O = h\nu \qquad (4)$$

- As *ν* is the frequency of the proton

What determines the magnetic resonance spectroscopy is that when we want to absorb a photon must go from a rotating state to another rotational position. This is achieved by the ΔE which depends on the magnetic field and on the frequency wavelength of the photon (Μιχάλης Φαρδής, 2017).

1.1.2. NMR spectrum

A technique similar to Magnetic resonance is a NMR spectrum. Nuclear Magnetic Resonance Spectroscopy deals with the transition of permissible energy states to the core. This occurs when the nuclear spin which is the spin component protons and neutrons in the nucleus interacts with the external magnetic field. According to quantum mechanics, the maximum value of the spin of a particle is an integer or $^1/_2$ of the integer number of the Planck constant $h = \frac{h}{2pi}$.

Value of spin depends on the mass and an individual number of the atom. If the mass and the individual are even then spin it is zero. When mass (x) is unnecessary then spin it's the $\frac{1}{2} * x$ whereas if the spin is really massive has values of 1,2,3. When applying a static magnetic field, a total magnetic moment parallel to the outer magnetic field is called a longitudinal thermal equilibrium control (Mo) and is given by:

$$M_O = \frac{N\gamma^2 B_0 h^2}{4kT} \qquad (5)$$

- As *N* is the total number of cores
- As *h* is understood as the constant of Planck

- As *Bo* stands for the external static magnetic field
- As *K* is the constant of Boltzmann
- As *T* is the temperature of the material

Therefore, magnetization is influenced by particle type, temperature, and external magnetic field. To have a larger signal, we need to increase the outer field, and if we increase the heat we will have less spin hence, therefore, a smaller signal. Particular attention is paid to hydrogen since 65-85% of our body composition is water molecules resulting from the hydrogen compound with oxygen (H_2O).

For the representation of a particular part of the body, we need a pulse of radio frequency electromagnetic waves equal to the angular frequency of Larmor. The pulse frequency should be equal to the Larmor frequency to transfer the M-magnet from the longitudinal axis to the transverse plane because we cannot compute the total magnetization as long as it is parallel. At the time of transferring the magnetization from the parallel axis to the vertical (90°), a magnetic field B1 is produced which is perpendicular to the *Bo* aligned with the x- axis. However, as we change the frequency on the oscilloscope, we reach a different point. This happens because each issue contains different molecule compounds. H forms covalent bonds with other elements such as C, O for the construction of proteins, carbohydrates, water, etc. Electrons along with the electrons of an element make up a local magnetic field, this means that the transient frequency changes depending on the molecule, with the local magnetic field. After the pulse pause, we emit radio waves that are detected by the coil. The different magnetic field formed by the orientation of the spin gives us the depiction depending on the tissue of interest (Μιχάλης Φαρδής, 2017)

The magnetic field is applied for a very short time, so induced magnetization is lost and we return to equilibrium, that is, the return of the electron to the initial state can be done either by the emission of a photon or by a jump. This phenomenon is called longitudinal relaxation has a *T1* time. *T1* gets a different value depending on the web we're looking at. This occurs depending on the chemical bond of the proton because each chemical bond has a different bond strength. It is also affected by the time constant *T2* called transverse recovery time. *T2* is the time the system needs when we change the active level to reduce the value of transverse magnetization to *x* and *y* level. That is, for the particular tissue we want to look over the frequency we will give, we must also find the right time to have it greater clarity of the signal (Αγγελική Καλπαξή, 2012).

1.1.3. Raman spectrum

The Raman spectroscopy is a special spectroscopic technique that deals with non-elastic light scattering. Initially, it has to be mentioned that the energy of material has three components, the rotation of the material, the vibrations of its atoms, and the movement of its electrons.

Raman spectroscopy asserts that when a material is irradiated with monochrome radiation, a new spectrum is created. This is because the radiation when scattered can change direction and may have a different frequency from that of the incident-tracing radiation. We can interpret Raman spectroscopy either with the classical or the quantum description depending on the interaction of the radiation with the movement of the molecules. Raman spectroscopy is also used to identify a protein as well as to study its secondary and tertiary structure (Ball D. W.,1962).

Classical

In the band of elastic scattering it is true that the frequency scattering ($\omega_{scattered}$) is equal to the stimulating frequency(ω_{laser}). The band with scatter frequencies lower than the excitation frequencies ($\omega_{scattered}$-$\omega_{laser} < 0$) called the Stoke band while the band with scatter frequencies higher than the excitation frequencies ($\omega_{scattered}$-$\omega_{laser} > 0$) called Anti - Stoke belt.

That if a material is in the electric field E a bipolar momentum of a molecule is induced called induced polarization

$$P = aE \quad (6)$$

- As a is the coefficient polarizability
- As E is defined as the electric field of measures

Energy released from an alternate monochromatic electromagnetic wave is described by the relationship: $E= E_o\ sin(\omega t)$ (7)

So it results from (6), (7): $P=a\ E_o\ sin(\omega t)$ (8)

If the material performs an internal motion, motion will affect the polarization of the material and is described by the equation: $\alpha = \alpha_o + \beta\ sin(\omega_o t)$ (9)

- As a is meant the polarization of the material
- As b is meant the variation width of the polarizability

From (6), (7), (8), (9) arises:

$$P= \alpha E_o\ sin(\omega t)+\frac{\beta E_o}{2}[cos(\omega - \omega_o\ ) - cos(\omega - \omega_o)t] \quad (10)$$

- As α is the polarizability of the molecule

This type shows us that when there is an internal motion we have a change in amplitude and we get a modulated wave, which in addition to the frequency of elastic scattering also has the new frequencies corresponding to the lateral zones of non-elastic scattering. Raman diffusion is the result of a molecular vibration that can change the polarity given by the da / dQ type. The polar molecules through the intermolecular attractants are displaced and when they find themselves in

an appropriate orientation they go together and acquire less energy. So they have more stability. It is important to stress that when the width value is $\beta = 0$, the lateral zones of non-elastic scattering cease to exist. So the width is directly related to the internal motion of the molecule.

Quantum

As we already mentioned Raman scattering refers to a non-elastic impact where the photon either loses or wins an oscillating quantum.

In quantum mechanics, scattering results from a shift of electrons from a lower level of energy to a higher level and at the same time from the shift in electron vibration. The process lasts 10-14 seconds at most. The energy difference between scattered photons id represented in Image 1.1. Raman shift in wave number is given by the following type

$$\bar{\nu} = \frac{1}{\lambda_{scattered}} \qquad (11)$$

- As $\lambda_{scattered}$ is the wavelength of the Raman scattered photons

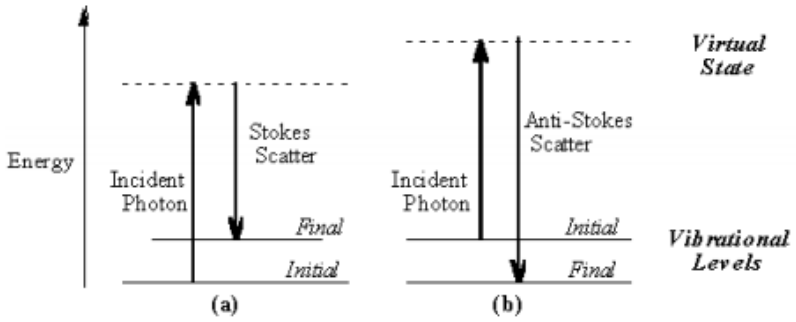


Image 1.1. (a) Stokes scattering (b) anti – Stokes scattering

The conservation of energy and momentum is given by the types:

$\omega_o = \omega_s + \omega_j(q)$	for zone Stokes	$\omega_o = \omega_s - \omega_j(q)$	for zone Anti-Stokes
$k_o = k_s + q$	for zone Stokes	$k_o = k_s - q$	for zone Anti-Stokes

With the quantum theory, we can present useful information on whether the scattering photon is transverse or what rules we should use to study crystalline systems. We achieve this because the quantum description gives us the Stokes / Anti-stokes tension and the laws of conservation of momentum and energy. (Λιαροκάπης Ε. & Τσιακλάγκανου Π., 2012). The Stokes zone in which we have a Stokes zone are at room temperature at a low energy level so a long wavelength. Scattered photon is located at higher energy levels as shown in Figure (b) with Zone Anti-Stokes. The Raman spectrum in the Stokes zone is weaker than the Anti-Stoker spectrum. From the Anti-Stoker spectrum we can get useful information about the structure of the molecule (Harris RK, 1983).

1.1.4. IR spectrum

IR spectroscopy is concerned with the absorption of infrared ray depending on the frequency. The molecule remains in the fundamental electron state, but the rotation and the donor movement of the molecule are affected. Infrared radiation does not cause electrons to migrate from one energy level to another such as ultraviolet radiation because it lacks enough energy. IR spectra are used to understand the structure of the molecules. Definitely, in order to get results, covalent bonds should have a bipolar momentum, because in order to absorb the molecule the radiation must undergo a change in bipolar momentum. There are two kinds of molecular vibrations: stress vibrations or extensive vibrations and bending or distortion vibrations. Trend vibrations are distinguished in symmetrical and asymmetric. They concern vibrations along the axis of the bond and the distance of the atoms changes. Bending vibrations are distinguished in scissors, rocking, vibrating and twisting. The vibrations change the angles between the molecular bonds.

As a general rule, the most important factors determining where to absorb infrared radiation a chemical bonds are the series of bonds and the types of atoms joining the bond. Thus, the same or similar functional groups in different molecules absorb at the same frequency levels and the more individuals are absorbing less radiation.

Absorption occurs at specific frequencies and must be coordinated, this means that the incident radiation has the same frequency as the dipole frequency.

To absorb infrared radiation a molecule should be subject to the following rules:

- The vibrations to which it applies $\frac{d\mu}{dx} \neq 0$ where $\frac{d\mu}{dx}$ tells us the change in bipolar moment in the change in distance between the atoms of a molecule.
- For molecules where the atomic vibrations transitions refer to adjacent molecules for which $\Delta v = \pm 1$ where vibrational quantum number (v = 0,1,2, ...)

  Based on the standard harmonic oscillator model, the potential equals:

$$V(x)=\frac{1}{2} k x^2 \qquad (12)$$

- As *k* is meaning of the restoring force constant is in the equilibrium position
- As *x* is the removal from the equilibrium position

According to the model of the simple harmonic oscillator and quantum mechanics, the vibrational energies of the levels are given by the formula:

$$E_v=(v+\frac{1}{2})\, hv \qquad (13)$$

- As *v* mean the vibrational quantum number
- As *h* is meant the Planck constant

- As V is meant the frequency of oscillation wherein for diatomic molecule equals:

$$v = \frac{1}{2\pi}\sqrt{\frac{\kappa}{\mu}} \text{ (14) and with a wavelength } \omega = \frac{\nu}{c} \text{ (15)}$$

So from (14), (15) we get the relation: $\omega = \frac{1}{2\pi c}\sqrt{\frac{\kappa}{\mu}}$ (16)

So we see that a simple harmonic oscillator transition between a calorific station located at equally spaced intervals and occurring at the same wavenumber *h*. Absolutely, in reality this is not just happening because some structures are rigid thereby presenting enharmonic oscillation. In this case, the energy levels converge and the vibrational quantum number increases and the dissociation energy D_e where the bond between people is broke up.

$$V_{(x)} = D_e(1 - e^{-\beta x})^2 \quad (17)$$

- As β is a k- like a parameter

In the non-harmonic oscillation, the transitions give $\Delta v \geq \pm 2$, so we are leading to the appearance of hypertonic absorption tapes.

The energy depends on the absorption frequency that is proportional to the wavelength. The absorption band of the first hypertonic (energy jump from $v = 0$ *to* $v = 2$) in the IR spectrum should occur at twice the wavelength of the fundamental absorption band. Absorption bands tapes are distinguished according to intensity in strong, moderate and patient. Generally, the absorption intensity depends on bipolar torque. The greater the change in the bipolar moment of the molecule, the higher the absorption of infrared electromagnetic radiation, so we have higher intensity absorption tapes. Bipolar torque depends on the severity of the bond, the greater the difference in electroreality rates, the more intense the absorption of radiation will be. (Ashenhurst J., 2016)

1.1.5. Mass Spectrometry, MS

Mass spectroscopy is a useful tool for the qualitative and quantitative determination of organic and inorganic substances. Its principle is based on the production of molecular ions or fragments by separating the masses of charged particles by bombarding an electron beam in our sample. This is achieved by the ionization of the molecule based on the ratio m / z where m is the mass of the ion and z the charge of the ion. Ionization can be performed by various techniques such as Electron impact ionization (EI), Photo ionization (PI), Chemical ionization (CI). Expulsion of the molecular mass can help in obtaining the cation of the chemical compound. Because of the excess energy of the ions, ions break up into ions of smaller size. All of this is done through fission and reshaping reactions, and in this way we get useful information about the structure of the union.

Then the ions are introduced to the mass analyzer after they have been influenced by an electric field and there we have the separation of the m / z ratio. In addition, the data is recorded in an electronic coil where it is possible to represent it.

Qualitative analysis can be done in gases, liquids, and under specific conditions to solids. Two important mass spectrometric methods are gas chromatography (GC) and liquid chromatography (LC) where they are used to measure metabolites in complex biological fluids. We examine the data from LC MS spectrum. in a mass spectrum through it (William Reusch,2013).

1.2. Behaviour of organic compounds

Identification of Organic Compounds using Nuclear Magnetic Resonance Spectroscopy, IR spectrum for the identification of organic molecules and mainly biomolecules. Proton spectra showing how many atoms of hydrogen contribute to the analysis.

1.2.1. Proton magnetic resonance spectrum

Proton spectra help to analyze organic compounds containing many hydrogen atoms because each hydrogen atom absorbs energy at different wavelengths. Each absorption energy is shown in the diagram as a peak. The vertical axis shows how much energy was absorbed. The horizontal axis shows the position of each peak depending on the energy absorbed using a catalyst in our union. Then the catalyst hydrogens are absorbed to give a single strong peak which is used as the starting point for the other peaks. Usually, the peaks of our compound appear on the left side of the top while the catalyst on the right side because they absorb lower energy radiation. The distance between the peaks and the start peak is expressed by a number called a chemical shift. This distance is measured in ppm. Individuals in a molecule have different chemical shifts because they approach different local magnetic fields depending on the electrons. Those atoms with high-density electrons are not so affected by the magnetic field that has been created. The chemical shift of hydrogen is important because we get information about the hydrogen-bonding links with other atoms. The area below our peaks shows how many hydrogen atoms contribute to each peak (John Wiley & Sons, 2004).

1.2.2. IR spectrum

The atoms in organic compounds are joined together by the bonds. The atoms, of course, do not stay at a fixed distance, they are free to vibrate in order to be separated. Molecular ions are thus actively unstable and some can be dissolved in smaller pieces. The simplest form is that the molecular ion to be disintegrated in two parts where one will be a positive ion and the other an

uncharged free radical. The free radical is either an atom or a group of atoms having an electron unbound, so it will not produce a line in the spectrum, while the positive ion will produce a line. Of course, it is important to emphasize that the pattern of lines in the spectrum of an organic compound shows something different from the spectrum of an element. Each line in the spectrum of one element shows a different isotope of the element in a compound shows a different fragment created when the molecule is broken down. Therefore, the infrared spectrometer is an instrument that takes the infrared light and through an organic compound produces a spectrum that shows us the amount of light on the vertical axis versus the wavelength of infrared radiation on the horizontal axis (William Reusch, 2013).

1.2.3. NMR spectrum

A spectroscopic technique that gives information on the number and types of atoms in a molecule. The number of signals shows how many different kinds of protons exist. The signal strength indicates the number of protons of the same type. Many atoms nuclei have a small magnetic field depending on their molecular and atomic numbers. When these atoms are in a strong magnetic field, we can have different energy states, the simplest is to have two energy states. At the bottom, the magnetic field is oriented parallel to the outer magnetic field, while at the higher it is oriented parallel to the magnetic field (B0) (Δρίτσα Αθηνά, 2002). The difference in energy states depends on the strength of the external magnetic field. NMR spectra using magnetic fields 1.4 to 18 Tesla. For example, in a magnetic field with a force of 1.4 T to shift from the lowest energy level to a higher one we need electro-magnetic radiation of about 60 MHZ in a H^1 core with 0.024 J / mol (-1). When we change the energy state of a core in our spectrum, a peak is created (Richard O.C. Norman, 2016).

Chapter II

Piecewise Monotonic Approximation Method

In section 2.1. we present the Piecewise Monotonic Approximation method applies to an NMR dataset and what output is printed. In section 2.2 we present the L2WPMA algorithm, the software of the method. This is helpful to present an experiment in Chapter III.

2.1. The method

Demetriou and Powell (Demetriou, I. & Koutoulidis, V. ,2013) have created a method for Piecewise Monotonic Data Approximation that aims to normalize data for better signal recognition. The algorithm is called L2WPMA, which is a Fortran 77 package and L2WPMA appears for Least Square Weighted Piecewise Monotonic Approximation. The method is used in many applications such as reducing noise for NMR data in the medical sector, business data and commercial transactions in the financial sector as well as restoring signals to a set of observations. The user can easily give the data $\{\varphi_i: i = 1, 2, \ldots, n\}$ which is a sequence of a set of numbers of a function $f(x)$ with truncated $x_1 < x_2 < \ldots < x_n$. The sequence has random errors (ε_i) so that $\varphi(x_i) = f(x_i) + \varepsilon_i$. The user also gives an integer k that symbolizes monotone changes in the sequence. The resulting change points will be at most *k-1* given that k is less than n. The algorithm instantly finds the optimal turning points by creating a piecewise linear monotonic changes. It calculates a vector y $\{y_i: i = 1, 2, \ldots, n\}$ to minimize the sum of squares of residues.

$$\Phi(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{n} w_i(\varphi_i - y_i)^2 \qquad (2.1)$$

Subject to:

$$y_{t_{i-1}} \leq y_{t_{i-1}+1} \leq \ldots \leq y_{t_i}, \text{ if j is odd,}$$

$$y_{t_{i-1}} \geq y_{t_{i-1}+1} \geq \ldots \geq y_{t_i}, \text{ if j is even} \qquad (2.2)$$

Where t_j is an integer number $\{tj : j = 0, 1, \ldots, k-1\}$ showing us the location of change points when the condition is satisfied.

$$1 = t_0 \leq t_2 \leq \cdots \leq t_k = n. \qquad (2.3)$$

These integers are also unknown to minimize, which reduces their possible combinations of choice to $O(n^k)$ and makes exhaustive control impossible. In addition, the problem can display so many local minima that classical optimization algorithms are excluded because they will end at a local minimum that will not necessarily be total. (Δημητρίου Κ. Ιωάννης, 1997). Calculate such problems, it is sufficient to use dynamic programming. If the data does not satisfy the conditions then $\{t_i: i = 1,2,...., k-1\}$ acts differently. The efficiency of the method depends on two characteristics of optimal fit. The first characteristic is if $1 \leq j \leq k-1$ then the turning points are the optimal value of y_{t_i} and satisfy the interpolation conditions.

$$y_{t_i}=\varphi_{t_i} \ , j =1, 2, ..., k-1 \tag{2.4}$$

The second characteristic is that each single monotonic part can be found separately from the best piecewise monotonic fit because it is the best value to the corresponding data.

This components $\{y_i : i = t_{j-1}, t_{j-1} +1, ... , t_j\}$ on $[x_{j-1}, x_{tj}]$ minimize the sum

$$\sum_{i=t_{j-1}}^{t_j}(y_i - \varphi_i)^2 \tag{2.5}$$

Subject to

$$y_i \leq y_{i+1} \ , \quad \text{i= } t_{j-1}, t_{j-1} +1, ... , t_j \ , \qquad \text{if j is odd} \tag{2.6}$$

$$y_i \geq y_{i+1} \ , \quad \text{i= } t_{j-1}, t_{j-1} +1, ... , t_j \ , \qquad \text{if j is even.} \tag{2.7}$$

The required minimum is given by the expression:

$$\alpha(t_0, t_1)+ \beta(t_1, t_2)+ \alpha(t_2, t_3) + ... + \gamma(t_{k-1}, t_k) \tag{2.8}$$

if k is an odd number then y is equal to α while when k is even then γ is equal to β where

$$\alpha(p,q) = min \{ \sum_{i=q}^{q}(y_i - \varphi_i)^2 : y_p \leq y_{p+1} \leq ... \leq y_q\} \tag{2.9}$$

and

$$\beta(p,q) = min \{ \sum_{i=q}^{q}(y_i - \varphi_i)^2 : y_p \geq y_{p+1} \geq ... \geq y_q \ \} \tag{2.10}$$

In this way, the property of separating the monochromatic segments is accomplished with a sort of dynamic programming.

We have $Y(k, n)$ with n data and the relation 2 is satisfied. It follows that We divide the data into m segments where $m \in [1, k]$ and the first given is a monotonous sequence and we define $t \in [1, n]$.

We define

$$\delta(m,t)= min_{z \in Y(m,t)} \sum_{i=1}^{t}(z_i - \varphi_i)^2 \quad (2.11)$$

where *Y (m, t)* consists of t data with m monotonic segments. So t_{k-1} satisfy relations

$$\left.\begin{array}{l} \gamma(k\text{-}1,\ t_{k-1}) + \alpha\ (t_{k-1},n)= min_{1\le s \le n}[(m\text{-}1,s)+\alpha(s,n)],\ \text{if } k \text{ is odd} \\ \\ \gamma(k\text{-}1,\ t_{k-1}) + \beta\ (t_{k-1},n)= min_{1\le s \le n}[(m\text{-}1,s)+\beta(s,n)],\ \text{if } k \text{ is even} \end{array}\right\} \quad (2.12)$$

Whenever dynamic programming results in the minimization with the following objective function:

$$\gamma\ (m,t) = \begin{cases} min_{1\le s \le n}[(m\text{-}1,s)+\alpha(s,t)],\ m \text{ odd} \\ \\ min_{1\le s \le n}[(m\text{-}1,s)+\beta(s,t)],\ m \text{ even,} \end{cases} \quad (2.13)$$

For *t = 1,2, ..., n* for any value of *m Є [2, k]*. Condition (2.13) is minimized when *γ (m, t)* is equal to s for each value of m, t. When *m = k* the value of γ (m, t) is equal to the value t_{k-1}. So t_0 *=1* and t_k= n and we get the best values of the sequence by the backward formula.

$$t_{m-1} = \tau\ (m, t_m),\ for\ m=k,k\text{-}1,\ldots,2. \quad (2.14)$$

Our goal is to find more efficient ways to circumscribe operations to search for the best values so we divide the values $\{t_i$: j = 1,2, ..., k-1$\}$ into subsets of data. Let *t Є [1, n]* be an index of data and φ_i the sequence $\{\varphi_i$: i = 1,2, ..., n$\}$. If $\varphi_i > \varphi_t$ is true, then t is a local minimum index and all indices of the local minima are summed up to one L set. Correspondingly, if $\varphi_i < \varphi_t$ then t is a local maximal marker and accumulate it in a U set. This reduces the complexity of *O(n)* in both subsets with the data being in ascending order and each of the subsets having fewer than $\frac{n}{2}$ elements.

In complexity of the method $O\ (kn)^2$ for numeric operations and *O(kn)* for memory locations. This package is effective because it uses properties of the Piecewise Monotonic Approximation that reduces computational cost. (Demetriou I. C. 2015) (Powell, M.J.D, 1970).

2.2. Use of L2WPMA software

L2WPMA is an algorithm that has been implemented in Fortran. The baseline is the creation of spectrum. Baseline value of x is the value of k that tells us how many peak we will have in the

spectrum. The algorithm supports its function in creating sections according to monotony. Then it separates the parts into smaller parts and checks the most is best. The algorithm deals with distinct signals. As an output the user gets the smoothed data according to the piecewise monotonic approximation method as well as the positions of the turning points (I. C. Demetriou, "Algorithm 863", 2007).

The dataset contains of a 81 observations that was downloaded from Human Metabolisms Database. It is also given in the CD that accompanies this thesis with the name Hmdb0000633.xls Figure 2.1, 2.2 and 2.3 show the smaller fit on the dataset for k= 1,2,3 and 4 respectively. Table 2.1. the data is displayed, and the abscissa (m_z) represents the molecular breakup and the ordinate (RI) presents the intensity that is needed to disintegration the D-Biopterin protein (Andrews, Kenneth J. M.; Barber, W. E.; Tong, B. P., 1969).

Table 2.1. Data of D- Biopterin protein

No.	mz	RI(%)	No.	mz	RI(%)	No.	mz	RI(%)
1	43.323	11.087	28	114.85	0.503	55	159.323	6.005
2	45.314	0.625	29	116.857	0.38	56	160.346	7.065
3	53.284	0.584	30	117.786	0.543	57	161.143	7.283
4	56.36	0.341	31	118.785	1.005	58	162.618	0.435
5	60.24	0.442	32	120.003	3.859	59	163.665	2.12
6	67.118	0.951	33	120.643	0.324	60	164.602	11.304
7	68.148	2.799	34	122.025	11.196	61	165.289	4.755
8	69.374	0.435	35	123.946	20.109	62	166.281	0.611
9	78.149	1.196	36	125.429	0.306	63	174.94	24.783
10	80.147	12.283	37	129.989	4.538	64	177.134	100
11	81.412	1.617	38	131.652	1.046	65	177.657	0.87
12	82.443	0.659	39	132.269	4.946	66	178.204	30.87
13	83.317	0.462	40	132.948	16.739	67	179.672	0.324
14	86.057	1.114	41	133.503	1.556	68	184.919	12.717
15	92.209	1.06	42	134.393	6.332	69	189.62	0.768
16	93.302	1.25	43	135.291	6.087	70	190.378	2.228
17	94.489	1.596	44	136.282	1.291	71	191.26	9.13
18	95.394	1.664	45	139.038	0.462	72	192.056	35.652
19	96.37	0.951	46	146.995	28.696	73	192.595	1.433
20	97.377	0.496	47	148.369	0.387	74	193.478	26.196
21	105.075	4.375	48	148.947	16.739	75	194.251	8.37
22	105.716	0.917	49	149.563	1.461	76	202.013	7.609
23	106.77	4.103	50	150.321	2.717	77	202.997	2.826
24	107.839	12.065	51	151.414	1.155	78	219.927	11.087
25	108.44	2.092	52	152.265	0.7	79	236.318	0.367
26	109.385	1.719	53	156.958	3.288	80	237.552	5.217
27	110.345	0.761	54	158.652	1.399	81	238.279	4.81

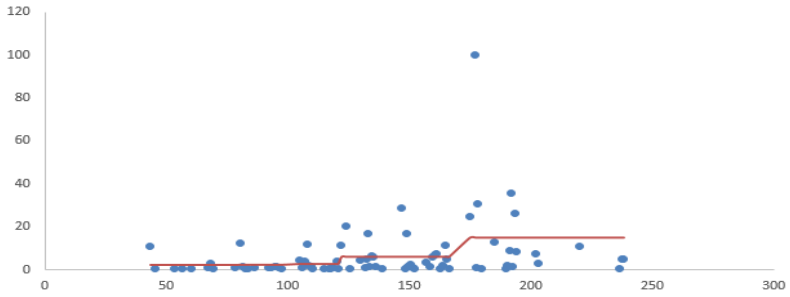
We can see that the method is a piecewise constant function that gives the best piecewise monotonic approximation method. This gives a lot of flexibility to fit with local trends an global trends. L2WPMA has recognized an optimal fit between 81^{10} combinations.

Table 2.2. D- Biopterin Turning points of k = 1, k= 2, k= 3

	xi	tj
k=1	0	1
	1	81
k=2	0	1
	1	64
	2	81
k=3	0	1
	1	64
	2	79
	3	81
k=4	0	1
	1	64
	2	69
	3	72
	4	81

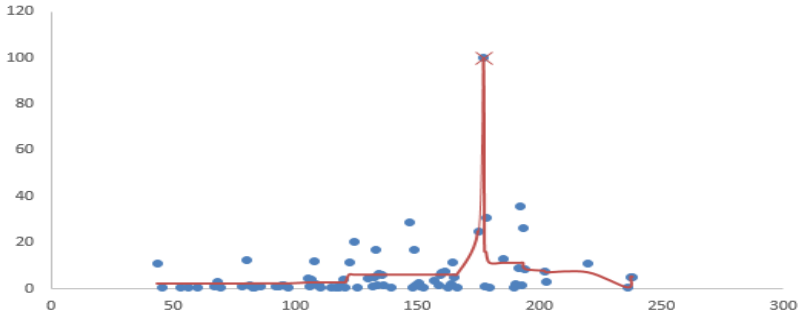
For k = 1 we do not get any turning point. For k = 2 we get a turning point in the position t_j = 64. For k = 3 we get the positions 64,79 and for k = 4 we get the positions 64,69,72. From Table 2.2 and from Figure 2.1 and Figure 2.2, we find that the point created for each increasing k is used as the starting point for moving the positions of the next k. This is logical since we follow the piecewise monotonic approximation method where it looks before and after that next local locality.

Figure 2.1. D-Biopterin Best fit with k=1



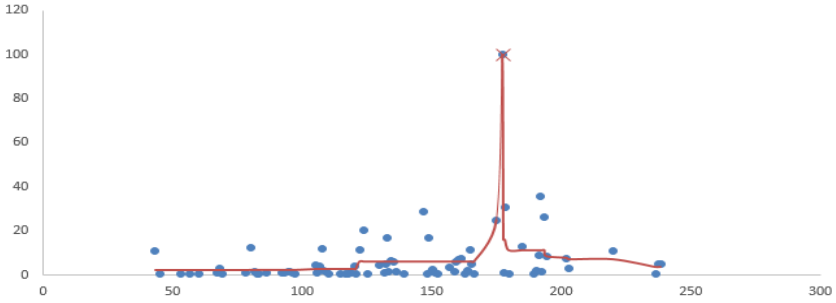
For k = 1 no turning point gives us a peak, because the case is simply the best monotonic fit to the data.

Figure 2.2. D-Biopterin Best fit with k=2



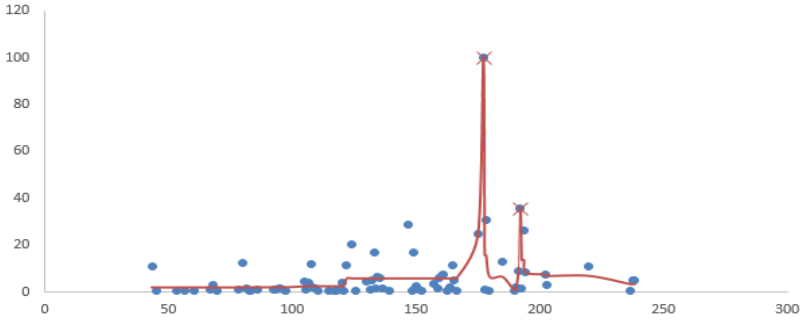
The case k = 2 gives one turning point whereby a peak is reached which, as we see from the figure,that gives the highest peak and follow the major trend of the data.

Figure 2.3. D-Biopterin Best fit with k=3



The case k = 3 gives one extra turning point, because the extra turning point contributes with the first data point.

Figure 2.4. D-Biopterin Best fit with k=4



The case k = 4 gives one more extra turning point and a peak emerges which, as we see from the Figure 2.4., gives the previous peak and the next largest point that is close to the previous peak.

Chapter III

Estimation of peaks of a NMR spectrum

The section 3.1. we will refer to the database we used to collect the data and give a brief description of that database. The section 3.2. we give a small reference to what data the user enters the program and what output we get. The section 3.3. we describe the spectra obtained by NMR data and explain the peaks generated by y-axis proportions and the position where the peak is found to be more x-axis. The section 3.4. we analyze three NMR data because the objective of NMR spectroscopy is to extract standards so that we can compare quickly and effectively to get useful information for diagnosis, prognosis, and even treat specific diseases.

3.1. Human Metabolome Database

In this chapter are performed an experiment where we used a data set containing NMR spectra through mass spectroscopy which are freely available on the official site - Human Metabolome Database (HMDB). Datasets was downloaded from Hyman Metabolome Database and the results are saved in "NMR-returns.xls on the CD that accompanies the thesis. This database contains detailed information regarding small molecule metabolites found in the human organism and is free available. The database is aimed at bringing together three types of data, chemical data, clinical data and molecular biology / biochemistry data. The HMDB contains mass spectra which can be used by the library to identify metabolites from mixtures by LC-MS / MS spectroscopy or by NMR spectroscopy. It is particularly useful because it can be used in applications related to clinical chemistry, discovery of biological markers and general education. Our goal is to show the depiction of the estimated peaks in a sample NMR spectrum. Mass spectrometry in NMR spectra is quite useful for conducting information without destroying the sample. For this reason it is used in costly biological samples such as DNA nucleic acids, RNA or proteins. In addition, the NMP spectra provide useful information on the three-dimensional structure of proteins (David S. Wishart, 2007). The datasets were downloaded from that site Human Metabolisms Database which have named HMDB0001885, HMDB0000001 and HMDB0012127, HMDB0006050, HMDB0000582, HMDB0001537. The data we can get from the database is a description of the protein found by the spectrum. Similar proteins as well as a chemical formula. In addition, we obtain information on the average molecular weight and the monoisotopic molecular weight. Finally we can get the spectrum and the peak table. The spectrum is found in the MassBank of North America database (MoNa) which contains metadata-centric for design, efficient storage and mass spectroscopy. The database includes over 200,000 mass

spectra from experiments, in-silico libraries and user contributions. This is useful because we can determine the function of our method and show that it is efficient in estimating the peaks and efficiency in computer time. We are going to show that the piecewise Monotonic approximation method also applies efficiently to NMR data**.**

3.2. Applying the piecewise monotonic approximation method to a NMR spectrum

The file we use contains two columns of data. The first column refers to the mass-to-load ratio (m / z) and is represented by the set $\{x_i$: i = 1, 2, . . , n$\}$ and the second column refers to the intensity (DI) where it is represented by the set $\{\varphi_i$: i = 1, 2, . . . ,n$\}$. Spectra gives us useful information from the peaks because the creation of a peak is the molecule that breaks down when it passes in a spectrometer detector. Definitely a spectrum presents some peculiarities such as the presence of noise, the creation of random peaks from chemical-electric noise and the appearance of the baseline by the ionization of the molecules. The aim is to normalize the behaviors and to achieve the goal we use the method we analyzed in Chapter II. (Demetriou I. C. 2015).

The file contains 2289 pairs of data with | U | = 573 and | L | = 574. The second dataset contains 2065 pairs of data with | U | = 584 and | L | = 585 and thirty dataset have 1803 data with | U | = 506 | L | = 507 can visualize our data in figures 3.3., 3.6, 3.9. The forth dataset contains 2044 pairs of data with |U|= 593 and |L|= 594 We will learn about its structure according to increases or decreases in intensity, the creation of discrete or less distinct peaks, and the deviations between the peaks. The fifth dataset contains 3588 pairs of data with | U | = 891 and | L | = 892 and sixth dataset contains 2289 pairs of data with | U | = 658 and | L | = 659. For the data for the first file to which $|L| \approx \frac{n}{2}$ we do not have to calculate a (s, t) and b (s, t) and keep it stored because it is uneconomical for some computer installations. The L2WPMA algorithm is repeated k times and the algorithm is called for $|U|$*-1 and* $|L|$ times as shown in our data. When the minimum squares are presented in the algorithm, the complexity is reduced to $(n|U| + k|U|^2)$ (Demetriou I. C, 1995)

3.3. Structure of protein

A diagram with NMR data of an organic compound shows totally different things from a graph with NMR data of an element. With a compound every change point shows us a different fragment that is produced when a molecular ion is cleaved while with one element we get a different isotope of the element. The pair x and y are the representation of the ions generated by a chemical compound and show the relative percentage of each ion (y- axis) in relation to the mass ratio to

the m / z (x- axis). Disrupted molecules are usually converted to positive ions. The breakdown of the molecule is accomplished by electromagnetic field. According to the paper Κουφόπουλος Ν., (2015), we can distinguish the structure and identify from primary and secondary metabolites databases, oligonucleotides, peptdes, hormones, pharmaceuticals, pesticides, insecticides, food products.

- Mass peaks: are the peaks that make up the mass spectrum of the organic compound. Their position shows the ratio of the mass to the load while the height shows the intensity, whereas the relative percentage of molecular ion in the total ion generated by ionization
- Molecular Ion Peak: is the major ion generated by the loss of an electron of the molecule of a compound and is symbolized by M^+. It is called the mother peak and gives the exact molecular weight of the molecule. It consists of a very simple mass spectrum so all other molecules consist of heavier isotopes and are symbolized as $(M+1)^+$, $(M+2)^+$.
  Most of the time, the isotopes found most often in an association have the smallest mass. A typical example is the water it has $M^+ = 18$, $(M+2)^+ = 20$.

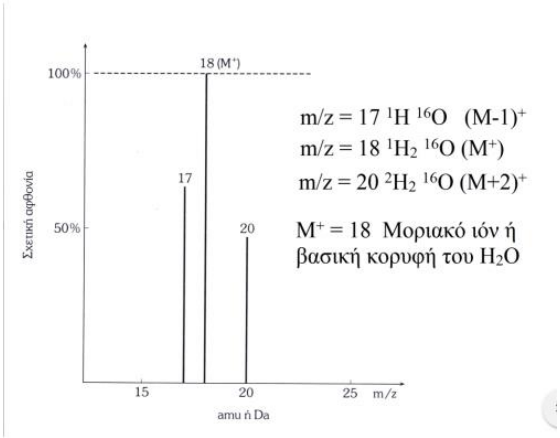


Image 3.2: Mass Spectrum of water (H_2O)

- Base peak: is the largest peak and corresponds to the base ion containing most of the ions produced during ionization. In general, the initial compound is broken down into many fragments and the mother peak appears to barely anywhere. However, peak intensities generated in the spectrum are usually attributed as a percentage of the base peak.
- Fragment peaks: initially a ion is produced which has an excess of energy. This ion breaks down and fragments are created. One of these holds the charge while the others can be either fixed molecules or roots. An important role for the mass spectrum to be produced depends on the strength and chemical nature of the bonds that retain the fragment with the rest of the molecule.

The mass spectrum is used to identify a chemical compound as two different molecules can not be broken up and ionized by an electron beam. While a fragmentation figure derived from NMR data may identify an unknown compound. The procedure followed for emptying the unknown chemical compound is given below.

Initially we try to locate the molecular formula of the compound from the mass spectrum. This is accomplished by identifying the mother peak which can easily detect the molecular weight of the molecule. In addition, it can be detected by the interpretation of the spectrum of fragments. Homogeneous rows can give a break diagram because they have very few bonds between the atoms. Such series are saturated aliphatic hydrocarbons, unsaturated aliphatic hydrocarbons, ketones, alcohols, esters, aromatic carbohydrates, etc. Another way to determine the molecular weight of the compound by means of an isotopic cluster where we can divide the natural isotopes of a compound. In addition, molecules having heavier isotopes have peaks (M + 1, M + 2) relative to the natural isotope. Finally, we can find out which compound is, the ring rule or the nitrogen rule that give us information about the nature of the union. We can learn about the number of unsaturated sites or the number of nitrogen atoms that each molecule has in the compound. Definitely, it is more difficult to identify compounds having high molecular weight and intense ionization and therefore a large number of fragments.

Absolutely, there are modern instruments with standard spectral libraries that can compare the unknown spectra with those of libraries.

Table 3.1: Atoms found in the molecules

Isotope	Atomic Weight	Physical distribution
H^{1}	1.007825	9998.5%
H^{2}	2.014102	0.015%
C^{12}	12.000000	98.9%
C^{13}	13.003354	1.1%
N^{14}	14.003074	99.64%
N^{15}	15.000108	0.36%
O^{16}	15.994915	99.8%
O^{17}	16.999133	0.04%
O^{18}	17.777160	0.2%
Cl^{35}	34.968855	75.8%
Cl^{37}	36.965896	24.2%

Our data obtained from the HMDB database using the Piecewise Monotonic Approximation Method and the L2WPMA algorithm resulted in Figure 3.3 for k = 24. The corresponding diagram also exists in the HMDB database which certifies the operation and the efficiency of the method. The protein produced by the chart is 3-chlorotyrosine.

3-Chlorotyrosine, a specific marker of myeloperoxidase-catalyzed oxidation, is noticeably raised in low - density lipoprotein isolated from human atherosclerotic intima. Specifically, myeloperoxidase halogenates tyrosine residues in plasma proteins and and generates 3-chlorotyrosine (CY). The detection of free chlorotyrosine in blood or urine arises from the disintegration of these chlorinated proteins. CY concentrations can play an important role in activating neutrophils in asthmatic patients. The protein belongs to the classes of organic tyrosine compounds and derivatives. This is the result of its reaction between tyrosine molecules or derivatives together with the amino group or the carboxy group. Definitely, it can also result from the replacement of any glycine hydrogen by a heteroatom. The protein has the chemical formula *C9H10ClNO3* and consists of the molecules *C6H4NO, C6H2CIO, C6H6CIO, C7HCIN, C8H4Cl, C7H2CIO2, C7H6CIO2, C7H5CINO2* (Bouchilloux, Simone, 1955).

In addition, we will represent the data in the CD and lead to figure 3.9 for k = 14. The corresponding diagram is also present in the HMDB database which certifies the operation and efficiency of the method. The protein produced by the diagram is 1-methylhistidine. The protein results from the cysteine reaction with the amino group or with the carboxyl group or from the replacement of the glycine hydrogen by another molecule. One-methylhistidine (1-MHis) is usually found in poultry. The enzyme of carnosinase is divided into b-alanine and 1-Mhis. In people with multiple sclerosis and in Parkinson's, high levels of 1-Mhis are found when eating meat, then the carnosinase enzyme is inhibited and anserine is produced. 1-Mhis is a biomarker for eating meat, especially red meat. The protein has the chemical formula $C_7H_{11}N_3O_2$ (Jain, Rahul; Cohen, Louis A., 1996).

The third dataset refers to the 2-Benzylsuccinate protein belonging to the organic compounds known as phenylpropanoic acid. These compounds contain a benzene ring together with propanoic acid. The corresponding spectrum is also available on the Human Metabolisms Database website that certifies the operation and effectiveness of the method. The third dataset refers to the 2-Benzylsuccinate protein belonging to the organic compounds known as phenylpropanoic acid. These compounds contain a benzene ring together with propanoic acid. The corresponding range is also available on the Human Metabolisms Database website that certifies the operation and effectiveness of the method. 2-Benzyl succinate is an aromatic compound found in the environment and is usually degraded by anaerobic microorganisms. Degradation leads to only a few molecule breaks until the ring breaks. Benzoyl-CoA from the intermediate molecules is the most important since it occurs in a large number of organic compounds such as choro-, nitro-, and amino benzoates, aromatic hydrocarbons, and phenolic. The protein has the chemical formula $C_{11}H_{12}O_4$ (Alfa Johnson Matthey Company Johnson Mattheay Catalog Company, Inc., 2009).

The forth dataset refers to the o- Tyrosine. O-Tyrosine is a human metabolite where your presence occurs due to the hydroxylation of 1-phenylalanine with the hydroxyl radical. O-Tyrosine is proposed as an hydroxy radical biomarker of oxidative damage to proteins. The metabolite appears in patients like Kwashiorkor where there is a lack of pro-energy. Many publications suggest that the effects of o-tyrosine are not inconclusive. (Orhan H, 2004). O-Tyrosine belongs to the phenylalanine group because it contains phenylalanine in the compound. The formation of phenylalanine in the molecule results from the reaction of phenylalanine with the amide or with the carboxyl group or from the replacement of a hydrogen glycine with another atom. The metabolite has the chemical formula $C_9H_{11}NO_3$ (Kokotos, 1986)

Also, we will represent the data in the CD and lead the Tables and Figures that give us the structure and molecules that make up the protein. The protein produced by the Figure 3.16 is 3,5 diiiodothyrosine. 3,5-Diiodothyronine was considered an inactive thyroid hormone metabolite due to its lack of correlation with thyroid receptor thrombin (THR) receptors. Certain recent research has shown that 3,5-Diiodothyronine stimulates cell / mitochondrial respiration. In addition, the molecule affects the transcription of certain genes (Goglia F., 2004). 3,5-Diiodothyronine belongs to the organic compounds known as phenylalanine and its derivatives. As mentioned in the fourth dataset, phenylalanine results from the reaction of phenylalanine in the amide or in the carboxyl group or from the replacement of any hydrogen by the glycine molecule with a heteroatom. The protein has the chemical formula $C_{15}H_{13}I_2NO_4$ (Valashek, I. E., 1995).

Finally, the sixty dataset refers 6-Hydroxydopamine belongs to the class of organic compounds known as catecholamines and derivatives. The corresponding spectrum is also available on the Human Metabolisms Database website that certifies the operation and effectiveness of the method. Catecholamines and derivatives are compounds containing 4- (2- Aminoethyl) pyrocatechol [4- (2- aminoethy) benzene- 1,2-diol (Napolitano, 1995). 6-hydroxydopamine after in vitro and in vivo investigations concluded that it causes the death of dopaminergic neurons. After Elizabeth J., (2002) investigations, we have shown that the endoplasmic stress of the nervous system contributes to neuronal death. These findings, in combination with Parkinson's disease, increased the chances of extensive obstruction of endoplasmic reticulopathy and the non-unfolded proactive response to the pathophysiology of the disease. The protein has the chemical formula $C_8H_{11}NO_3$(Napolitano, 1995).

3.4. Applications of the piecewise motonic method to a NMR data

3.4.1. First dataset representation

In this section we are looking for turning points to get useful information depending on monotonic trends. We will refer to different k such as k = 4, 10, 20, 24 where we see that gives the optimal solution and we will present the corresponding figures. In particular, the turning points are shown in Tables 3.2, 3.3 and 3.4, 3.5. In these tables, we distinguish the dynamic change of the signal according to the number of single parts (k) we give it. The data was entered into the method and the best fit was found. The method then identified turning points for dataset HMDB0001885. The results are saved in HMDB0001885 file.xls on the CD that accompanies the thesis (Bouchilloux, Simone, 1955).

Table 3.2. 3-Chlorotyrosine Turning points for k=4

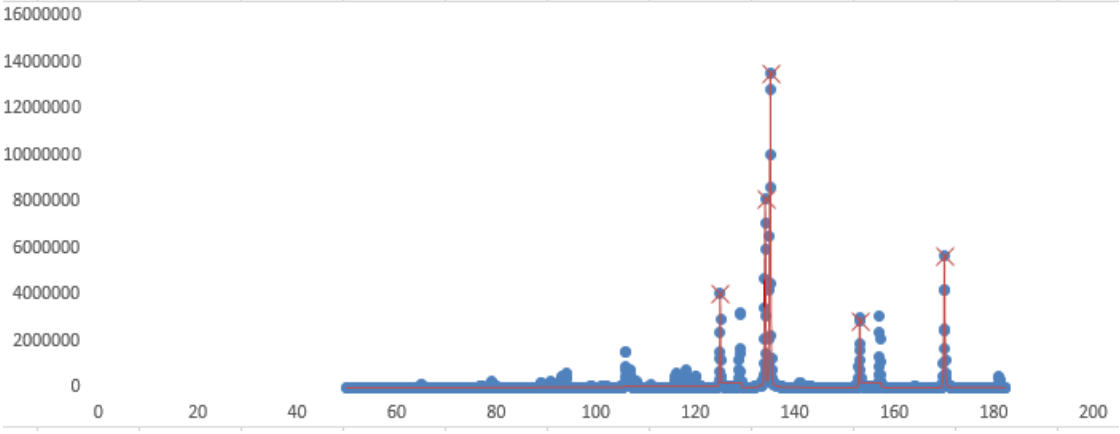
No	tj	ϕ(tj)	y(tj)
0	49.98	165	165
1	134.10	8163367	8163367
2	135	255417	255417
3	135	13550039	13550039
4	250	182	182

In addition to k we also have n-1 peaks. For instance, for k = 4 the turning points are in the positions 1, 1318, 1325, 1334, 2289. We use the algorithm with k = 10. We see that the fit data is smoother display than the raw data and the sum of the squares of the residuals is equal to 58719110960000. The following figure shows the smooth display of the data.

Table 3.3. 3-Chlorotyrosine Turning points for k=10

No	tj	φ(tj)	y(tj)
0	50	165	165
1	125	4087469	212307
2	132	4533	212307
3	134	8163367	8163367
4	135	255417	255417
5	135	13550039	13550039
6	150	80	105123
7	153	3009954	105123
8	165	372	105123
9	170	5640891	105123
10	250	182	182

Figure 3.1. 3-Chlorotyrosine Best fit with k=10

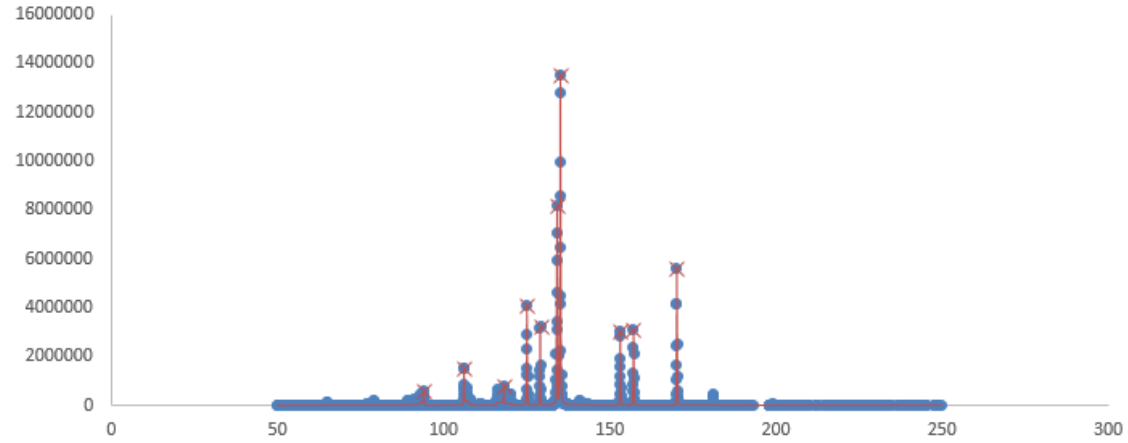


So for k = 10 we have 5 peaks. In a case where we are not satisfied with the depiction of the first attempt to evaluate the correct rotation of these data points and I do not have enough information, we can get more k to see more points of change. In this way we will concentrate on small monotonic tendencies that are not visible with less k. These peaks are created when we change the monotony. The following figure shows the smooth display of data for k = 20

Table 3.4. 3-Chlorotyrosine Turning points for k=20

No	tj	φ(tj)	y(tj)
0	50	165	165
1	94	616141	28113
2	98	183	28113
3	106	1554800	68236
4	113	330	68236
5	118	781810	68742
6	122	1249	68742
7	125	4087469	212307
8	128	4956	212307
9	129	3210380	212307
10	132	4562	212307
11	134	8163367	8163367
12	135	255417	255417
13	170	5640891	105123
14	150	80	105123
15	153	3009954	105123
16	156	1819	105123
17	157	3076792	105123
18	167	430	105123
19	170	5640891	105123
20	250	182	182

Figure 3.2. 3-Chlorotyrosine Best fit with k=20

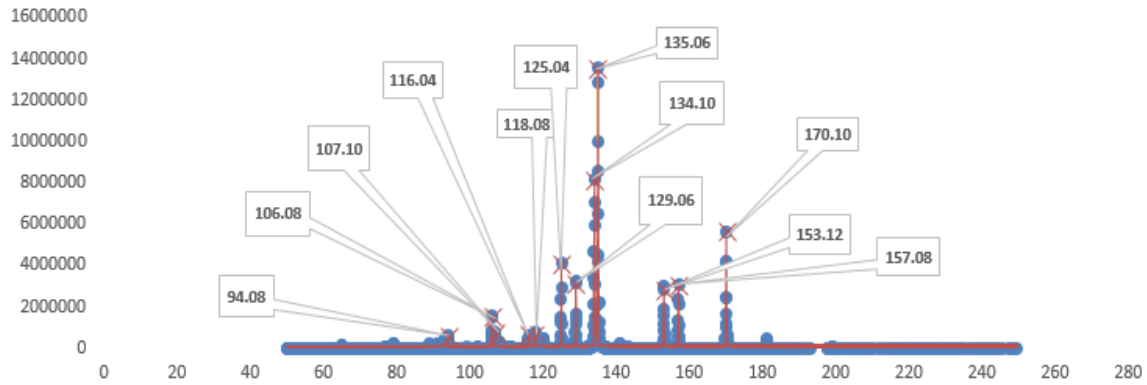


The increase in monotonic trends maintained the turning points presented for k = 10 located at positions 1167, 1276, 1318, 1325, 1334, 1558, 1610, 1807, 1882 while the additional points for k = 20 are in positions 653, 712, 851, 974, 1051, 1110, 1209, 1234, 1656, 1677. In addition, we have a small shift of the relocation from data 1276, 1558, 1807 at data 1281, 1566, 1837. Each new peak gives information about some cleavage that can be done in the molecule that will lead us to the structure of the protein. If we get k = 20 instead of k = 24 we had previously found 19 hat we now have 23 turning points with a sum of squares equal to 3536243130000.

Table 3.5. 3-Chlorotyrosine Turning poins for k=24

No	tj	φ(tj)	y(tj)
0	50	165	165
1	94	616141	28113
2	98	183	28113
3	106	1554800	68236
4	106.50	19702.38	68235.91
5	107.10	749701.56	68235.91
6	113.58	244.57	68235.91
7	116.10	637792.25	68741.86
8	116.46	9958.29	68741.86
9	118	781810	68742
10	122	1249	68742
11	125	4087469	212307
12	128	4956	212307
13	129	3210380	212307
14	132	4562	212307
15	134	8163367	8163367
16	135	255417	255417
17	170	5640891	105123
18	150	80	105123
19	153	3009954	105123
20	156	1819	105123
21	157	3076792	105123
22	167	430	105123
23	170	5640891	105123
24	250	182	182

Figure 3.3. 3-Chlorotyrosine Best fit with k=24



In Figure 3.3 we present 4 extra peaks for k = 24 with 23 turning points with a sum of squares equal to 2292162494000 which are in positions 858, 868, 1018, 1024. In addition, we have a slight shift of the turn point from position 974 to position 976. It is remarkable when the number of peaks increases, so the left side of the chart is enhanced. We notice that we now have the most important turning points while inserting the data in these points. With the increase of k, through the piecewise monotonic approximation, what's why we were able to make the sum of squares of the residuals smaller by holding the most important turning points.

Table 3.6. 3-Chlorotyrosine Turning points positions for several of k are indicated by the spectrum

No	tj	k=	4	10	12	14	16	18	20	24	26
0	1		x	x	x	x	x	x	x	x	x
1	653								x	x	x
2	712								x	x	x
3	851						x	x	x	x	x
4	858									x	x
5	868									x	x
6	971						x	x			
6	974								x		
6	976									x	x
7	1018									x	x
8	1024									x	x
9	1051							x	x	x	x
10	1110							x	x	x	x
11	1167			x	x	x	x	x	x	x	x
12	1209				x	x	x	x	x	x	x
13	1234				x		x	x	x	x	x
14	1276			x							x
14	1281				x	x	x	x	x	x	x
15	1318		x	x	x	x	x	x	x	x	x
16	1325		x	x	x	x	x	x	x	x	x

17	1334		x	x	x	x	x	x	x	x	x
18	1558			x	x	x	x				
18	1566							x	x	x	x
19	1610			x	x	x	x	x	x	x	x
20	1564					x					
20	1656						x	x	x	x	x
21	1677					x	x	x	x	x	x
22	1807			x							
22	1816				x						
22	1837					x	x	x	x	x	x
23	1882			x	x	x	x	x	x	x	x
24	2024										x
25	2045										x
26	2289		x	x	x	x	x	x	x	x	x

The Table 3.6 shows the positions of the bending points according to the different values of k for the 3-Chlorotyrosine molecule. In the table we see in the left part the positions of the turning points in the column marked "tj" and on the left we see the different values of k where k Є {2, 4, 8, 10, 12, 14, 16, 20, 24, 26}. Each time we choose to use k + 2 monotonic changes because we notice that the turning points for some have remained while some others have shifted a bit.It is remarkable that as the k increased, there are additional bend points adjacent to the bending points of the previous k. This shows that the best choice chosen can be a starting point for calculating the best fit for sequential data. We first see that the points found are at points 1318, 1325, 1334 that remain for each k and give us the highest peaks. Finally for k = 24 we have the optimal solution for k = 26, we find that all previous bending points are maintained and another peak is created which may be noise because it does not give any information about the structure of the protein.

The method we used as shown for the different k we used gave the necessary peaks that are useful to distinguish the structure of the molecule. It has been proven that any algorithm with local improvements finds a single local minimum that we do not know if it is the total minimum.We assume, of course, that the method is valid for k+2 while for successive k in the same time it needs special attention and we can lead to the creation of new more efficient algorithms where it may reduce the time needed for a suitable k.

3.4.2. Second dataset representantion

We will present the spectral behavior of another dataset to assert that piecewise Monotonic approximation method is efficacy in estimating the peaks and efficiency in computing time. Our record contains 2065 data with | U | = 584 and | L | = 585. The spectrum consists of vertices representing compounds of the compound. The deciphering of the molecule from the spectrum is based on the location of the peaks and the distance between them according to the intensity we

give to the molecule. The results are saved in HMDB0000001 file.xls on the CD that accompanies the thesis (Jain, Rahul; Cohen, Louis A., 1996).We will refer to various k as k = 4,8,10,14 to see more gives the optimal solution and we will present the corresponding figure. In particular, the turning points are shown in tables 3.7, 3.8, 3.9. In these tables we see the change in the spectrum according to the number of single parts (k). For k = 4 we have 2 peaks. Monotonic changes are very few that it is almost impossible to understand for which protein.

Table 3.7. 1-Methylhistidine Turning points for k=4

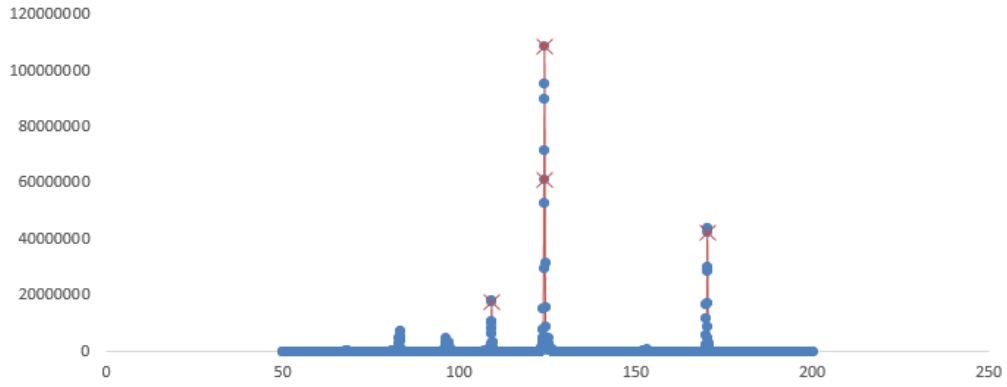
No	tj	φ(tj)	y(tj)
0	50	298	298
1	124	108788680	108788680
2	141	1249	1249
3	170	44215432	44215432
4	200	1114	1114

For k = 8 we have 4 peaks, 7 monotone stresses. The points we have monotonic changes are in positions 1, 515, 616, 949, 1090, 1197, 1462, 1741, 2065. We grow k until we do not get useful information or discern noise in data.

Table 3.8. 1-Methylhistidine Turning points for k=8

No	tj	φ(tj)	y(tj)
0	50	298	298
1	83	7272554	7272554
2	89	298	298
3	109	18176692	18176692
4	118	298	298
5	124	108788680	108788680
6	145	951	1437
7	170	44215432	44215432
8	200	1114	1114

Figure 3.4. 1-Methylhistidine Best fit with k=8

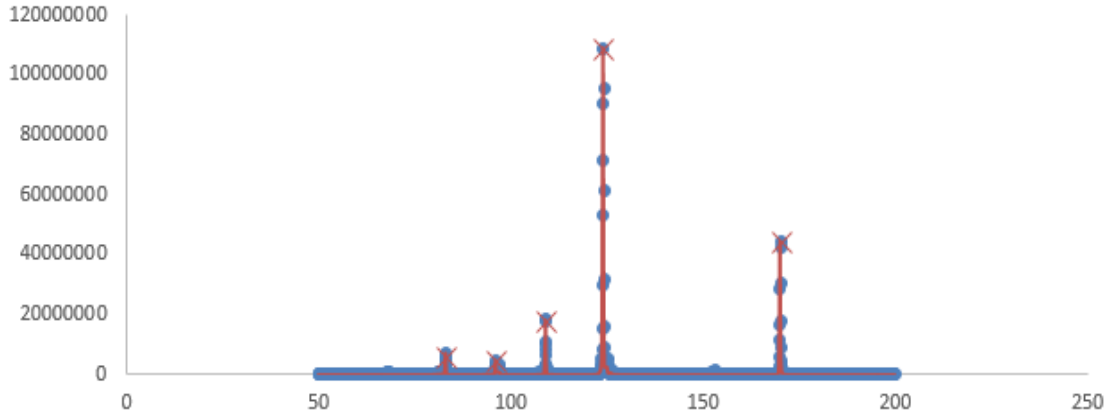


We grow k until we do not get useful information or discern the data. For k = 10 the new positions created are 732, 849. As we have a small shift of position 1462 at position 1502.

Table 3.9. 1-Methylhistidine Turning points for k=10

No	tj	φ(tj)	y(tj)
0	50	298	298
1	83	7272554	7272554
2	89	298	298
3	96	4766357	4766357
4	103	1359	1359
5	109	18176692	18176692
6	118	298	298
7	124	108788680	108788680
8	150	314	314
9	170	44215432	44215432
10	200	1114	1114

Figure 3.5. 1-Methylhistidine Best fit with k=10



If we increase k = 14 the extra positions we get are 738, 749, 1208, 1214. For k = 14 we can get as many peaks as we need to understand the structure of the protein being created. In addition for k = 16 a peak is created which does not give us information about a molecule.

Table 3.10. 1-Methylhistidine Turning points for k=14

No	tj	φ(tj)	y(tj)
0	50	298	298
1	83	7272554	7272554
2	89	298	298
3	96	4766357	4766357
4	96	88746	902662
5	97	3287751	902662
6	103	1359	1359
7	109	18176692	18176692
8	118	298	298
9	124	108788680	108788680
10	125	1190583	2631991
11	125	4966357	2631991

12	150	314	314
13	170	44215432	44215432
14	200	1114	1114

Figure 3.6. 1-Methylhistidine Best fit with k=14

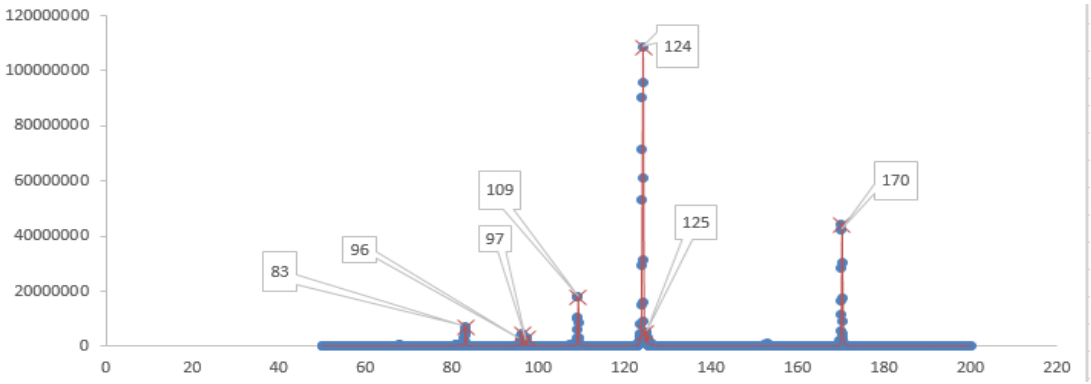


Table 3.11. 1-Methylhistidine Turning points positions for several of k are indicated by the spectrum

No	tj	k=	4	6	8	10	12	14	16
0	1		x	x	x	x	x	x	x
1	515				x	x	x	x	x
2	616				x	x	x	x	x
3	732					x	x	x	x
4	738							x	x
5	749							x	x
6	849					x	x	x	x
7	949			x	x	x	x	x	x
8	1090			x	x	x	x	x	x
9	1197		x	x	x	x	x	x	x
10	1208						x	x	x
11	1214						x	x	x
12	1431		x						
12	1441			x					
12	1462				x				
12	1502					x	x	x	x
13	1542								x
14	1625								x
15	1741		x	x	x	x	x	x	x

As in the previous example, we presented in table 3.10 the turning points for k + 2 monotonic changes for k Є {2, 4, 8, 10, 12, 14, 16}. And in this example we find that usually every time we increase k, the positions of the previous k are retained, while sometimes we have a slight shift. This shift, of course, results in a better adaptation of the method. We also notice that as the k grows so much more turning point we have. As the k grows up, the turning points created are

located near a turning point that was simply created in the previous k. For this reason, the turning point of the previous k can be considered a starting point for calculating the optimal fit. We first see that the points found are at 1197, 1741, 1431. Positions 1197, 1741 are maintained as k grows and these points give us the highest peaks. Finally for k = 14 we have the optimal solution for k = 16, we find that all previous bending points are maintained and two other turning points are created which may be noisy because with the new peak we do not get any information about the structure of the protein. The Piecewise monotonic approximation method we used for different k gave the required peaks that are useful to distinguish the structure of the molecule and can lead us to the treatment of diseases. The protein produced by the diagram is 1-methylhistidine.

3.4.3. Thirth dataset representantion

We will report another dataset proving the piecewise monotonic approximation method is efficient for the peak evaluation and effective to calculate time for NMR data. The dataset that we used contains 1803 data with | U | = 506 | L | = 507. The results are saved in HMDB0012127 file.xls on the CD that accompanies the thesis. Our goal is to look for turning points in order to obtain information on the recognition and the structure of the protein. Depending from the monotones trades which will be entered by the user will be presented in the spectrum of $\frac{k}{2}$ distractions $\frac{k}{2}$ splits the protein. Certainly should be found appropriate the appropriate number k because if less will have incomplete information and whether is greater may be a noise that will lead us to false conclusions. For example we will refer to different k, as k = 4, 8, 12. 16.

Where for k = 16 will show that gives the optimal solution. We will show the corresponding spectra Figure 3.7, 3.8, 3.9 and turning points in the respective tables Table 3.11, 3.12., 3.13 and 3.14. The dataset was downloaded from that site Human Metabolisms Database which has named HMDB0012127 (Alfa Johnson Matthey Company Johnson Mattheay Catalog Company, Inc. 2009).

Initially we start for k = 4 monotonic changes. Turning points are located at positions 1, 838, 1178, 1246, 1803 and two peaks result.

Table 3.12. 2-Benzylsuccinate Turning points for k=4

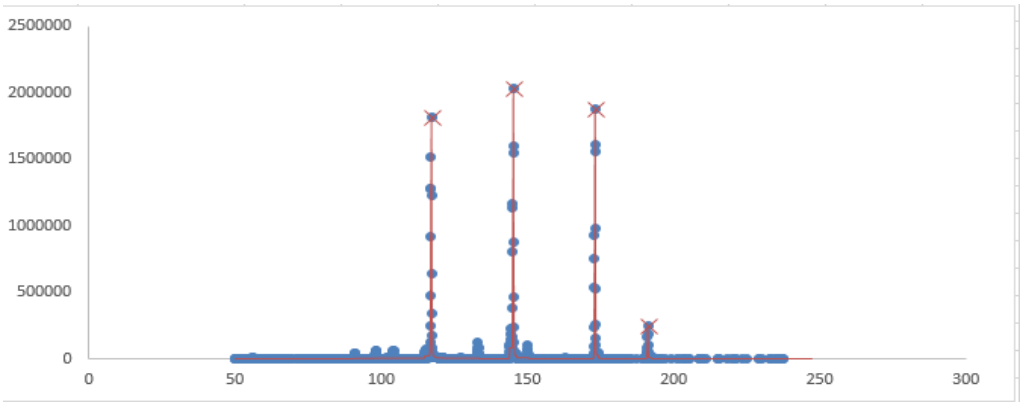
No	tj	φ(tj)	y(tj)
0	50	100	66
1	117	1812022	1812022
2	141	49	49
3	145	2028111	2028111
4	247	99	68

Then we use the algorithm for k = 8 monotonic changes. We note that the sum of the squares of the residuals is equal to 11593519730000 at 164119658500. The new positions of the turning points that are created are the previous and additional positions 1431, 1533, 1681 and 1702. The four peaks do not have all information in the identity of protein.

Table 3.13. 2-Benzylsuccinate Turning points for k=8

No	tj	φ(tj)	y(tj)
0	50	100	66
1	117	1812022	1812022
2	141	49	49
3	145	2028111	2028111
4	161	49	37099
5	173	1875121	37099
6	185	49	7043
7	191	246725	7043
8	247	99	68

Figure 3.7. 2-Benzylsuccinate Best fit with k=8



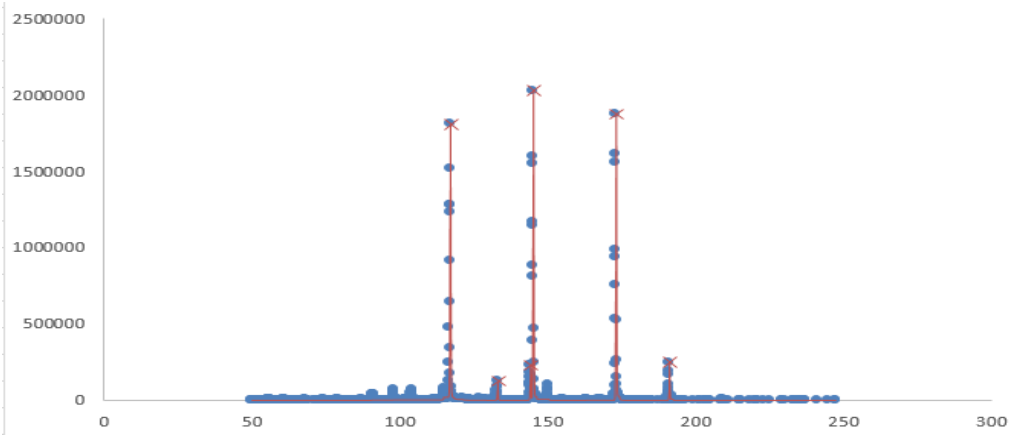
We use the algorithm for k = 12 monotone changes. We note that again the sum of the squares of the residuals is equal to 164119658500 at 68831264770. The new positions created are 1035, 1077, 1230 and 1236. In addition we have a small shift from position 1681 to position 1685. Definitely, we take more monotonic changes to make the right conclusions.

Table 3.14. 2-Benzylsuccinate Turning points for k=12

No	tj	φ(tj)	y(tj)
0	50	100	66
1	117	1812022	1812022
2	130	49	4179
3	133	120688	4179
4	141	49	49
5	144	229816	83150
6	145	16091	83150
7	145	2028111	2028111

8	161	49	37099
9	173	1875121	37099
10	185	49	7043
11	191	246725	7043
12	247	99	68

Figure 3.8. 2-Benzylsuccinate Best fit with k=12



Finally we use the algorithm for k = 16 monotonic changes. We note that again the sum of the squares of the residuals is equal is reduced from 68831264770 to 32429858820. The new turn points we get are 523, 678, 1315 and 1329. We have a small shift from position 1178 to position 1183.

Table 3.15. 2-Benzylsuccinate Turning points for k=16

j	tj	φ(tj)	y(tj)
0	50	100	66
1	98	65442	65442
2	108	49	49
3	117	1812022	1812022
4	130	49	4179
5	133	120688	4179
6	141	49	49
7	144	229816	83150
8	145	16091	83150
9	145	2028111	2028111
10	149	380	380
11	150	102943	102943
12	161	49	37099
13	173	1875121	37099
14	185	49	7043
15	191	246725	7043
16	247	99	68

Figure 3.9. 2-Benzylsuccinate Best fit with k=16

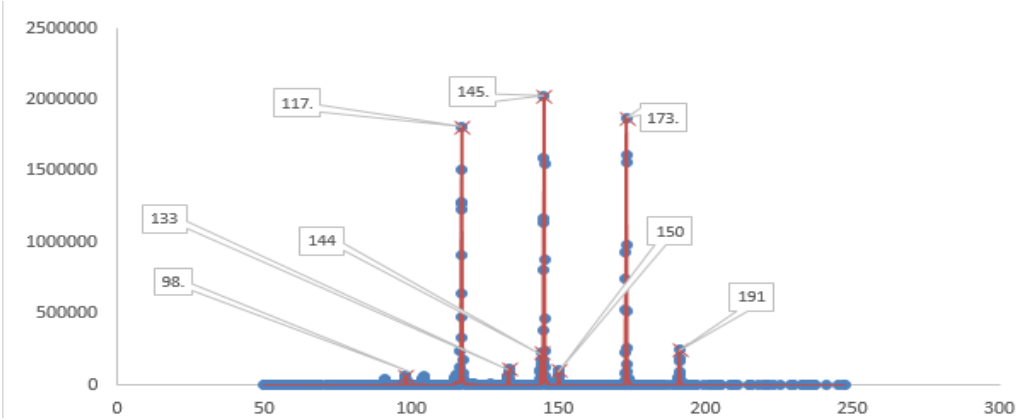


Table 3.16. 2-Benzylsuccinate The turning points positions for several of k are indicated by the spectrum

No	tj	k=	4	6	8	10	12	14	16	18
0	1		X	x	x	x	x	x	x	x
1	523								x	x
2	550									x
3	622									x
4	678								x	x
5	838		X	x	x	x	x	x	x	x
6	1035						x	x	x	x
7	1077						x	x	x	x
8	1178		X	x	x	x	x	x		
8	1183								x	x
9	1230					x	x	x	x	x
10	1236					x	x	x	x	x
11	1246		x	x	x	x	x	x	x	x
12	1315							x	x	x
13	1329							x	x	x
14	1431			x	x	x	x	x	x	x
15	1533			x	x	x	x	x	x	x
16	1681				x	x				
16	1685						x	x	x	x
17	1702		x		x	x	x	x	x	x
18	1803		x	x	x	x	x	x	x	x

As with the previous datasets, the bending points for all k we used until we reach the appropriate result are shown in Table 3.15. We chose to use k + 2 mono changes for k Є {2, 4, 6, 8, 10, 12, 14, 16, 18}. We notice that every time k grows, the positions of the previous k are usually preserved, and in some places we have a small shift that definitely leads to the optimal solution. In addition to increasing k, we get more turning points, so we get more peaks. It is remarkable to note that the new bend points created by the k increase are located near the bending points of the k that were invoked. For this, k can be considered as the starting point for calculating the optimal

fit. Finally for k = 16 we have the optimal solution. For k = 18 we find that it retains all the previous bending points and creates two other bend points which may be noise since it does not give any more information about the identification and structure of the prototype. The Piecewise monotonic approximation method for the various k gave the required peaks that are useful for distinguishing the structure of the molecule and can lead us to the treatment of diseases.

3.4.4. Fourth dataset representantion

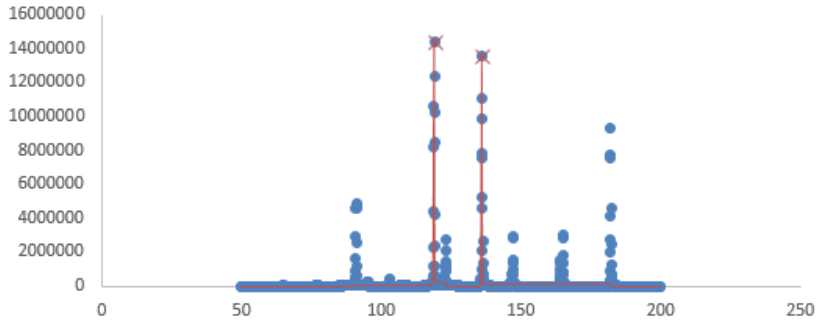
In this dataset we will show that piecewise monotonic approximation method is efficient and valid for another NMR dataset. The dataset that we used contains 2044 data with | U | = 593 | L | = 594. The results are saved in HMDB0006050 file.xls on the CD that accompanies the thesis. We will deal with k + 2 monotone changes in this example. We needed 18 attempts to find more molecular result from this dataset. We will present the Table 3.17, 3.18, 3.19, 3.20 which show us the points of change depending on the k we use. We will recreate the Figures 3.10, 3.11, 3.12, 3.13 to see the positions of the peaks and the distance between them. Finally, we will show the table with all the k and we have changed in Table 3.21.

Initially, we used k = 4 and produced 2 peaks. The turning points are located at positions 974, 1172 and 1256. As shown in Figure 3.10, data at positions 974 and 1256 give the highest peaks.

Table 3.17. O-Tyrosine Turning points for k=4

No	Tj	ϕ(tj)	y(tj)
0	50	196	196
1	119	14352097	14352097
2	131	226	226
3	136	13557751	13557751
4	200	224	212

Figure 3.10. O-Tyrosine Best fit with k=4

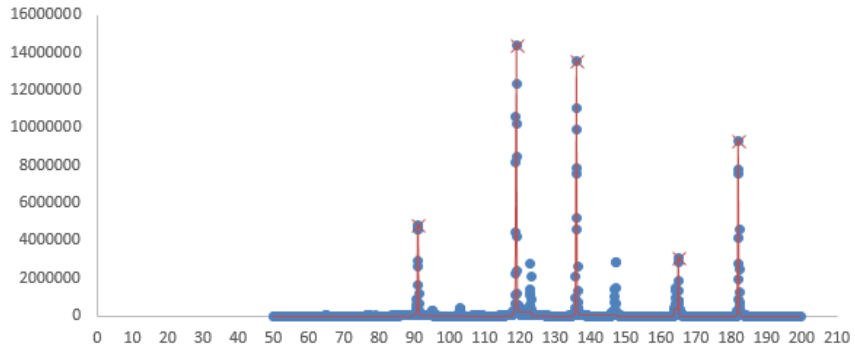


Then we will look for k = 10 monotone changes that give 9 turning points at positions 506, 820, 974, 1183, 1256, 1539, 1654, 1720, 1842. The method maintains points 974, 1256 that the peaks were presented for k = 4. Definitely, we have a shift of point 1172 to position 1183.

Table 3.18. O-Tyrosine Turning points for k=10

No	Tj	ϕ(tj)	y(tj)
0	50	196	196
1	91	4819052	4819052
2	110	208	208
3	119	14352097	14352097
4	132	223	223
5	136	13557751	13557751
6	153	337	337
7	165	3041575	3041575
8	169	171	171
9	182	9304014	9304014
10	200	224	212

Figure 3.11. O-Tyrosine Best fit with k=10



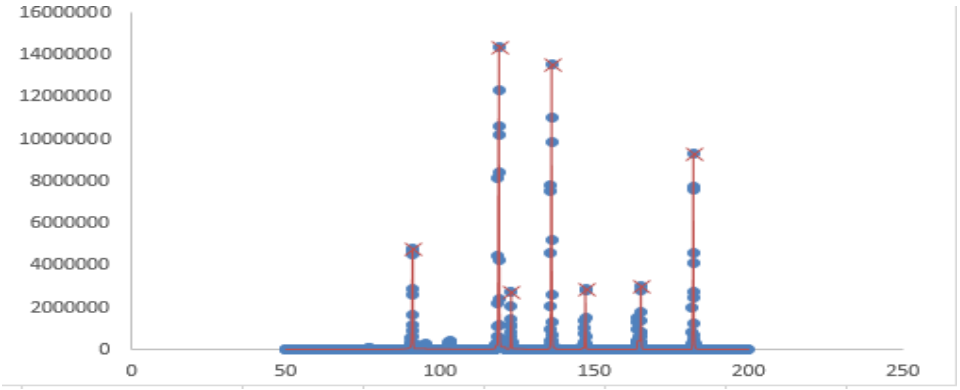
For k = 14, the 13 turning points retain the positions for k = 10 with some displacements in some positions, as well as 4 additional turning points. It holds positions 506, 820, 974, 1183, 1256, 1654, 1842. We have two small shifts from position 1539 to position 1605 and from position 1720 to position 1769. We have new positions at positions 1021, 1040, 1395, 1437 which are in adjacent positions between them.

Table 3.19. O-Tyrosine Turning points for k= 14

No	tj	ϕ(tj)	y(tj)
0	50	196	196
1	91	4819052	4819052
2	110	208	208
3	119	14352097	14352097
4	122	16435	16435
5	123	2763142	2763142
6	132	223	223
7	136	13557751	13557751
8	144	301	301
9	147	2879912	2879912
10	161	191	191
11	165	3041575	3041575
12	175	210	210

13	182	9304014	9304014
14	200	224	212

Figure 3.12. O-Tyrosine Best fit with k=14



The optimal solution was found for k=16. The remaining positions are in the points 560, 820,974, 1021, 1040, 1183, 1256, 1395, 1437, 1605, 1654, 1842. There is a small shift from position 1654 to position 1637. Additionally, two additional points were found at positions 1637 and 1645 adjacent to each other. The 8 peaks that are created are enough to find the structure of protein that is created in the spectrum. For k = 18 we do not get any additional information about its structure and there can be noisy that leads us to the wrong conclusions.

Table 3.20. O-Tyrosine Turning points for k= 16

No	tj	ϕ(tj)	y(tj)
0	50	196	196
1	91	4819052	4819052
2	110	208	208
3	119	14352097	14352097
4	122	16435	16435
5	123	2763142	2763142
6	132	223	223
7	136	13557751	13557751
8	144	301	301
9	147	2879912	2879912
10	161	191	191
11	164	1516598	1516598
12	165	22626	22626
13	165	3041575	3041575
14	176	198	198
15	182	9304014	9304014
16	200	224	212

Figure 3.13. O-Tyrosine Best fit with k=16

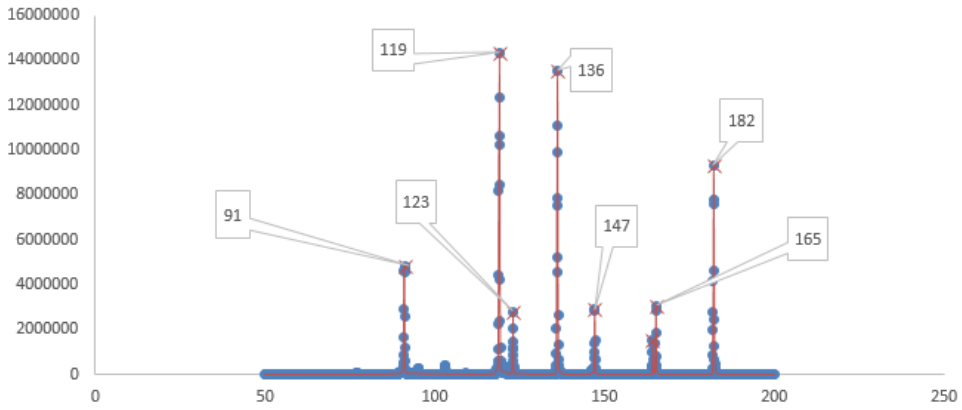


Table 3.21. O-Tyrosine The turning points positions for several of k are indicated by the spectrum

No	tj	k=	4	6	8	10	12	14	16	18
0	1		x	x	x	x	x	x	x	x
1	506				x	x	x	x	x	x
2	621									x
3	708									x
4	820				x	x	x	x	x	x
5	974		x	x	x	x	x	x	x	x
6	1021							x	x	x
7	1040							x	x	x
8	1172		x	x						
8	1183				x	x	x	x	x	x
9	1256		x	x	x	x	x	x	x	x
10	1393						x			
10	1395							x	x	x
11	1437						x	x	x	x
12	1539			x	x	x				
12	1605						x	x	x	x
13	1637								x	x
14	1645								x	x
15	1654					x	x	x	x	x
16	1720					x				
16	1760						x			
16	1769							x		
16	1778								x	
16	1796									x
17	1842			x	x	x	x	x	x	x
18	2044		x	x	x	x	x	x	x	x

As with the previous datasets, the turning points for all k we used until we reach the appropriate result are shown in Table 3.21. We chose to use k + 2 monotonic changes for k Є {2, 4, 6, 8, 10,

12, 14, 16, 18}. We notice that every time k grows, the positions of the previous k are usually preserved, and in some places we have a small shift that definitely leads to the optimal solution. In addition to increasing k, we get more turning points, so we get more peaks. It is remarkable to note that the new bend points created by the k increase are located near the bending points of the k that were invoked. For this, k can be considered as the starting point for calculating the optimal fit. Finally for k = 16 we have the optimal solution. The Piecewise monotonic approximation method for the various k gave the required peaks that are useful for distinguishing the structure of the molecule and can lead us to the treatment of diseases.

3.4.5. Fifth dataset representantion

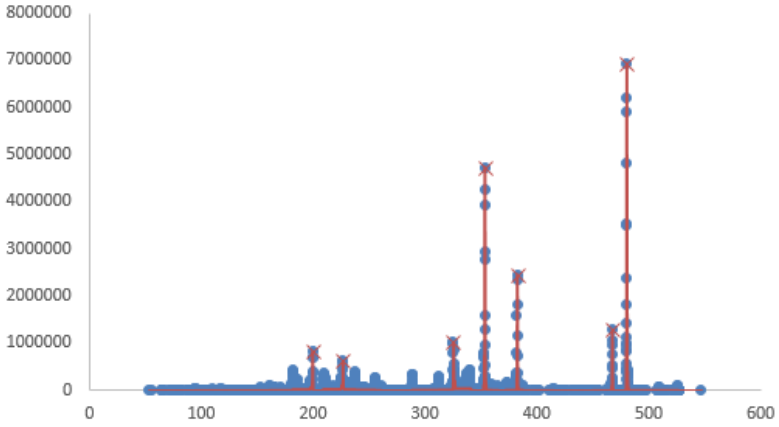
In the fifth dataset, we used to present the spectral behavior of a protein. We used the Piecewise monotonic approximation method and showed that it is also effective in a larger number of monotonous changes. The file contains 3588 data with | U | = 891 and | L | =892 . The peaks created in the spectra show the molecular breakdowns of the molecules constituting the protein. The results are saved in HMDB0000582 file.xls on the CD accompanying the master thesis (Valashek, I. E. ,1995). We have chosen a greater number of monotonic changes for k Є {4, 10, 14, 20, 30, 40, 50} because with a smaller number of k we could not reach the structure of protein.

Initially we start for k = 14 monotonic changes. Turning points are located at positions 1, 1028, 1122, 1398, 18932357, 2649, 2687, 2909, 3033, 3216, 3345, 3387, 3412, 3588 and seven peaks result.

Table 3.22. 3,5-Diiodothyronine Turning points for k= 14

No	tj	ϕ(tj)	y(tj)
0	53	507	493
1	199	828886	828886
2	206	367	367
3	226	617550	617550
4	267	286	286
5	325	1025606	1025606
6	348	478	478
7	353	4720430	4720430
8	370	302	302
9	382	2444820	2444820
10	437	444	444
11	467	1288391	1288391
12	470	369	369
13	480	6920267	6920267
14	547	367	367

Figure 3.14. 3,5-Diiodothyronine Best fit with k=14

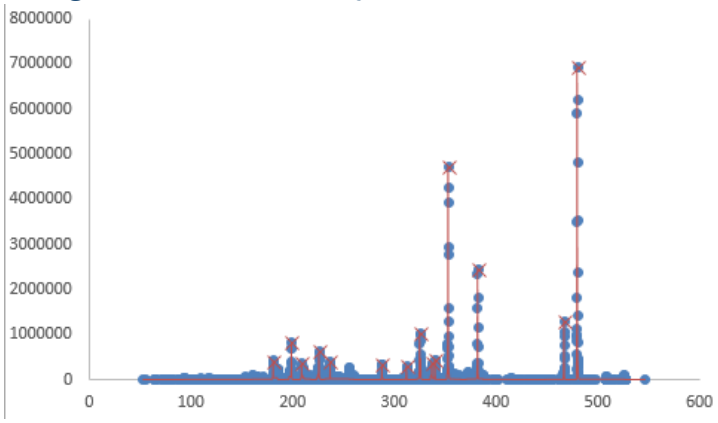


Then we will look for k = 30 monotonic changes that give 29 turning points at positions 1, 764, 889, 1028, 1122, 1168, 1351, 1398, 1456, 1564, 1893, 2005, 2068, 2200, 2246, 2357, 2363, 2374, 2454, 2499, 2528, 2548, 2658, 2687, 2909, 3033, 3216, 3345, 3387, 3412, 3588. The method maintains points 1, 1028, 1122, 1893, 2357, 2687, 2909, 3033, 3216, 3345, 3387, 3412, 3588 that the peaks were presented for k = 14. Definitely, we have a shift of point 2649 to position 2658.

Table 3.23. 3,5-Diiodothyronine Turning points for k= 30

No	tj	ϕ(tj)	y(tj)	j	tj	ϕ(tj)	y(tj)
0	50	200	200	16	326	87957	87957
1	181	420926	420926	17	326	463129	463129
2	189	302	302	18	334	302	302
3	199	828886	828886	19	337	368584	368584
4	206	367	367	20	339	7273	7273
5	209	356698	356698	21	340	447805	447805
6	223	286	286	22	351	369	369
7	226	617550	617550	23	353	4720430	4720430
8	230	302	302	24	370	302	302
9	237	390591	390591	25	382	2444820	2444820
10	267	286	286	26	437	444	444
11	288	338980	338980	27	467	1288391	1288391
12	294	286	286	28	470	369	369
13	312	283946	283946	29	480	6920267	6920267
14	315	302	302	30	547	367	367
15	325	1025606	1025606				

Figure 3.15. 3,5-Diiodothyronine Best fit with k=30



In this case, the method for k = 50 maintains the most points presented for k = 30 in the positions 1, 764, 889, 1028, 1122, 1168, 1351, 1398, 1456, 1564, 2005, 2068, 2200, 2246, 2357, 2363, 2374, 2454, 2499, 2528, 2548, 2658, 2687, 2909, 3033, 3216, 3345, 3412, 3588. While the additional points are located at positions 785, 830, 994, 1004, 1177, 1185, 1381, 1388, 1502, 1541, 1631, 1775, 2482, 2488, 2671, 2676, 2753, 2773, 2954, 2996. Additionally we have a slight shift of data from positions 1893, 3387 at positions 1903, 3389. The peaks created for k = 50 are 25 to 49 turning points while for k = 30 we had 15 peaks with 29 turning points.

Table 3.24. 3,5-Diiodothyronine Turning points for k= 50

j	tj	ϕ(tj)	y(tj)	j	tj	ϕ(tj)	y(tj)
0	50	200	200	26	315	302	302
1	181	420926	420926	27	325	1025606	1025606
2	181	420926	420926	28	326	87957	87957
3	185	240932	240932	29	326	463129	463129
4	189	302	302	30	334	302	302
5	197	216586	216586	31	336	220867	220867
6	198	9013	9013	32	337	20878	20878
7	199	828886	828886	33	337	368584	368584
8	206	367	367	34	339	7273	7273
9	209	356698	356698	35	340	447805	447805
10	210	15008	15008	36	351	369	369
11	210	271315	271315	37	352	234285	234285
12	223	286	286	38	353	35231	35231
13	225	277611	277611	39	353	4720430	4720430
14	226	19081	19081	40	357	367	367
15	226	617550	617550	41	359	123460	123460
16	230	302	302	42	370	302	302
17	233	182523	182523	43	373	165659	165659
18	236	286	286	44	380	302	302
19	237	390591	390591	45	382	2444820	2444820

20	242	286	286	46	437	444	444
21	255	270776	270776	47	467	1288391	1288391
22	271	302	302	48	470	286	286
23	288	338980	338980	49	480	6920267	6920267
24	294	286	286	50	547	367	367

Figure 3.16. 3,5-Diiodothyronine Best fit with k=50

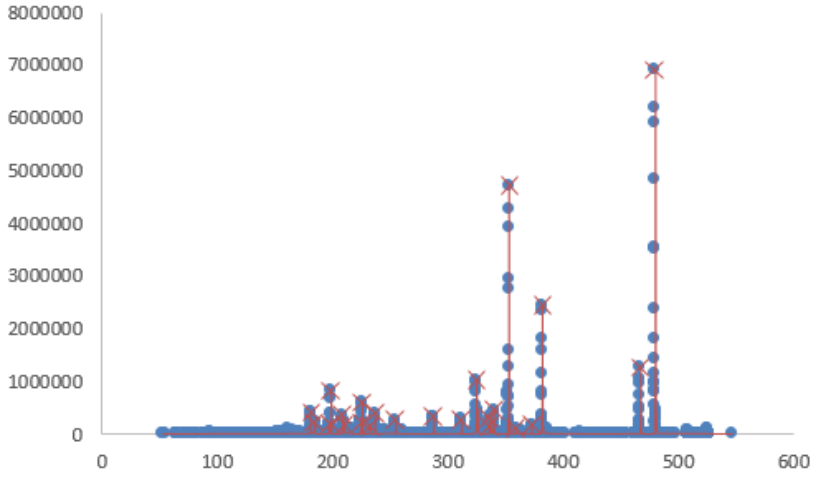


Table 3.25. 3,5-Diiodothyronine The turning points positions for several of k are indicated by the spectrum

No	tj	k=	14	20	30	40	50	No	tj	k=	14	20	30	40	50
0	1		x	x	x	x	x	27	2357		x	x	x	x	x
1	764			x	x	x	x	28	2363				x	x	x
2	785						x	29	2374				x	x	x
3	830						x	30	2454			x	x	x	x
4	889			x	x	x	x	31	2482						x
5	994						x	32	2488						x
6	1004						x	33	2499				x	x	x
7	1028		x	x	x	x	x	34	2528				x	x	x
8	1122		x	x	x	x	x	35	2548			x	x	x	x
9	1168				x	x	x	36	2649		x				
10	1177						x	36	2652			x			
11	1185						x	36	2658				x	x	x
12	1351				x	x	x	37	2671						x
13	1381						x	38	2676						x
14	1388						x	39	2687		x	x	x	x	x
15	1398		x	x	x	x	x	40	2753						x
16	1456			x	x	x	x	41	2773						x
17	1502						x	42	2909		x	x	x	x	x
18	1541						x	43	2954						x
19	1564			x	x	x	x	44	2996						x
20	1631						x	45	3033		x	x	x	x	x
21	1775						x	46	3216		x	x	x	x	x

22	1893		x	x	x	x		47	3345		x	x	x	x	x
22	1903						x	48	3387		x	x	x	x	
23	2005				x	x	x	48	3389						x
24	2068				x	x	x	49	3412		x	x	x	x	x
25	2200				x	x	x	50	3588		x	x	x	x	x
26	2246				x	x	x								

As in the previous examples, we presented in Table 3.25 the turning points for monotonic changes k + 2 for k Є {14,20,30,40,50}. And in this example we find that usually every time we increase k, the positions of the previous k are preserved, and sometimes we have a small shift. This shift, of course, leads to a better adaptation of the method. We also notice that as the k grows up so many more turning points we have. As the k grows, the turn points created are near a turning point created just in the previous k. For this reason, the turning point of the previous k can be considered as a starting point for calculating the optimal fit. Also all the new points created for k = 50 certainly have a new point near them. In this example, while we have the required points for the protein structure, we would have liked the piecewise monotonic approximation method to identify more peaks to provide more information about the structure of the protein called 3,5-Diiodothyronine. In conclusion, the method for one more time demonstrates its effectiveness in terms of time and results.

3.4.6. Sixth dataset representantion

We will study a recent dataset that has yielded usable results for NMR data. We used the piecewise monotonic approximation method and the dataset contains 2289 with | U | = 658 | L | = 659. The peaks created in the spectra show the molecular breakdowns of the molecules constituting the protein. The results are saved in HMDB0001537 file.xls on the CD accompanying the master thesis (Napolitano, 1995). We will refer to various k as k = 4,8,12,16 to see more gives the optimal solution and we will present the corresponding figure. In particular, the turning points are shown in Tables 3.26, 3.27, 3.28.

Initially we start for k = 4 monotonic changes. Turning points are located at positions 1, 1377, 1563, 1675, 2289 and produced 2 peaks result.

Table 3.26. 6-hydroxydopamine Turning points for k= 4

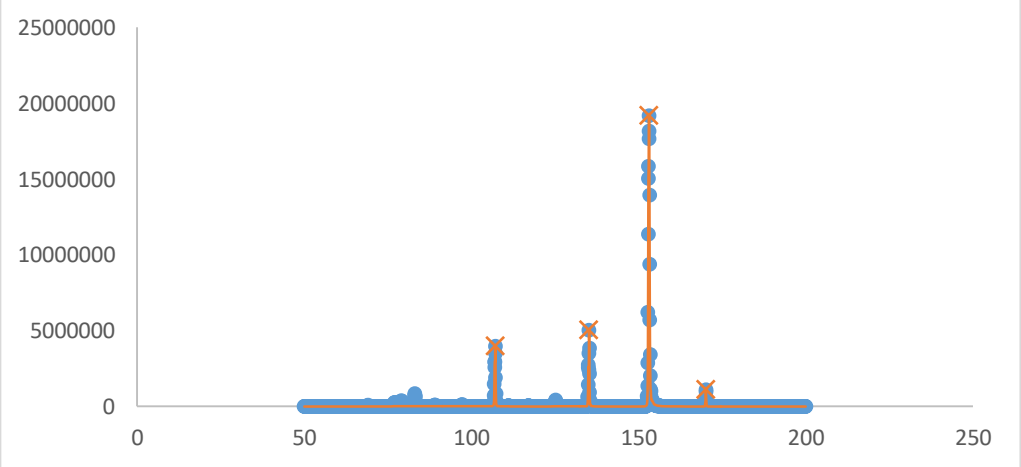
No	tj	ϕ(tj)	y(tj)
0	50	200	200
1	135	5024834	5024834
2	146	194	194
3	153	19204226	19204226
4	200	208	184

Then we will look for k = 8 monotonic changes that give 7 turning points at positions 1, 927, 1111, 1377, 1566, 1675, 1890, 1934, 2289. We have a small shift of the relocation from data 1566 at data 1563. Monotonic changes are very few that it is almost impossible to understand for which protein.

Table 3.27. 6-hydroxydopamine Turning points for k= 8

No	tj	ϕ(tj)	y(tj)
0	50	200	200
1	107	3992676	3992676
2	118	196	196
3	135	5024834	5024834
4	147	391	391
5	153	19204226	19204226
6	167	182	182
7	170	1117308	1117308
8	200	208	184

Figure 3.17. 6-hydroxydopamine Best fit with k=8



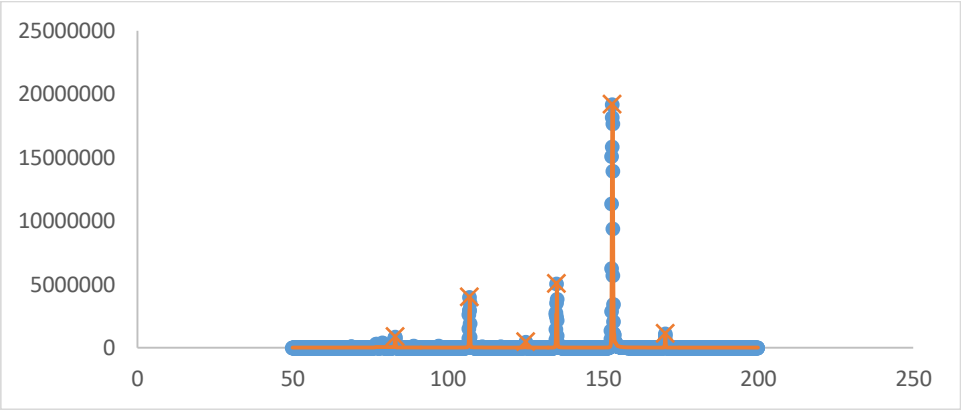
For k = 12, the 11 turning points retain the positions for k = 8 with some displacements in some positions, as well as 4 additional turning points. It holds positions 1, 927, 1377, 1675, 1934, 2289. We have two small shifts from position 1111 to position 1566, from position 1890 to position 1897 and from position 1568 to position 1769. We have new positions at positions 528, 782, 1213, 1309 which are in adjacent positions between them.

Table 3.28. 6-hydroxydopamine Turning points for k= 12

No	tj	ϕ(tj)	y(tj)
0	50	200	200
1	83	841457	841457
2	98	167	167
3	107	3992676	3992676
4	119	196	196
5	125	439149	439149

6	131	178	178
7	135	5024834	5024834
8	147	204	204
9	153	19204226	19204226
10	167	160	160
11	170	1117308	1117308
12	200	208	184

Figure 3.18. 6-hydroxydopamine Best fit with k=12



The optimal solution was found for k=16. The remaining positions are in the points 1, 528, 782, 927, 1213, 1309, 1377, 1568, 1675, 1897, 1934, 2289. There is a small shift from position 1115 to position 1145. Additionally, four additional points were found at positions 427, 438, 460 and 488 adjacent to each other. The 8 peaks that are created are enough to find the structure of protein that is created in the spectrum. For k = 18 we do not get any additional information about its structure and there can be noisy that leads us to the wrong conclusions.

Table 3.29. 6-hydroxydopamine Turning points for k= 16

No	tj	ϕ(tj)	y(tj)
0	50	200	200
1	77	281269	43006
2	78	5259	43006
3	79	395566	43006
4	81	2634	43006
5	83	841457	841457
6	98	167	167
7	107	3992676	3992676
8	121	164	390
9	125	439149	439149
10	131	178	178
11	135	5024834	5024834
12	147	204	204
13	153	19204226	19204226
14	167	160	160
15	170	1117308	1117308
16	200	208	184

Figure 3.19. 6-hydroxydopamine Best fit with k=16

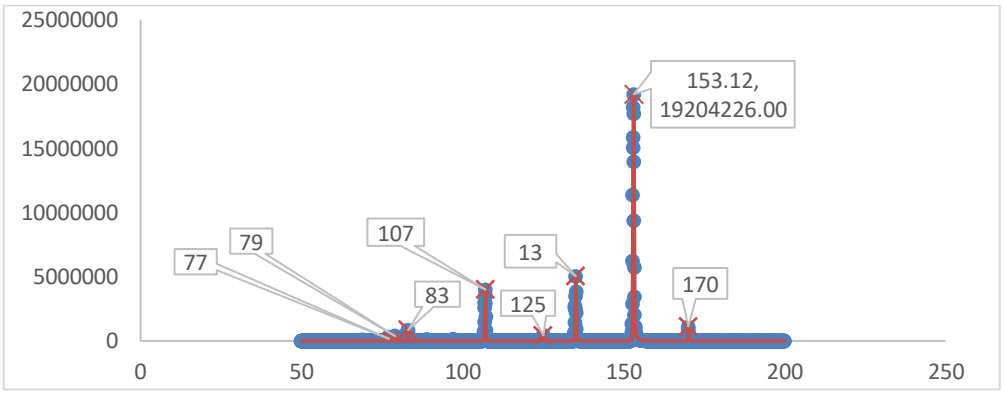


Table 3.30. 6-hydroxydopamine The turning points positions for several of k are indicated by the spectrum

No	tj	k=	4	6	8	10	12	14	16	18
0	1		x	x	x	x	x	x	x	x
1	427								x	x
2	438								x	x
3	460							x	x	x
4	488							x	x	x
5	528					x	x	x	x	x
6	782					x	x	x	x	x
7	927			x	x	x	x	x	x	x
8	1111			x	x					
8	1115					x	x			
8	1145							x	x	x
8	1213						x	x	x	x
9	1309						x	x	x	x
10	1374									x
10	1375									x
11	1377		x	x	x	x	x	x	x	x
12	1563		x							
12	1566			x	x					
12	1568					x	x	x	x	x
12	1675		x	x	x	x	x	x	x	x
13	1890				x					
13	1892					x				
13	1897						x	x	x	x
14	1934				x	x	x	x	x	x
15	2289		x	x	x	x	x	x	x	x

In the last dataset, we used 9 k Є {2, 4, 6, 8, 10, 12, 14, 16, 18}. We choose k+2 monotonic change and the results representation in the Table 3.29. We used 9 replicates to have suitable results for protein identification and structure. It is noteworthy that whenever k is grown, the positions of previous k remained. Definitely, we have a small shift in a few positions that leads to the optimal solution. Finally, for k = 16 we have the optimal solution. For k = 18 we find that it retains all the previous bending points and creates two other bend points which may be noise since it does not give any more information about the identification and structure of the prototype. The Piecewise monotonic approximation method for the various k gave the required peaks that are useful for distinguishing the structure of the molecule and can lead us to the treatment of diseases.

Chapte IV

Discussion and Conclusions

In this chapter we will present conclusions that may be useful for some future research according to our findings.

In the first chapter we discussed introductory terms and information according to the utility of spectroscopy and we analyzed various types of spectroscopy such as IR spectra, NMR spectra, Raman spectrum and Magnetic resonance spectrometers. We have also dealt with the behavior of proteins according to the spectroscopy we use.

In the second chapter we started the piecewise monotonic method and it is properties that make it valuable to peak finding in spectroscopy. We clarified the use of the method on a dataset that is easy to follow. As we provided an example to explain the creation of peaks.

In the third chapter we applied the piecewise monotonic method to three datasets as follows. We obtained from the HMDB database by LS MS / MS spectroscopy and referred to the 3-chlorotyrosine protein, one-methylhistidine protein and 2-Benzylsuccinate protein. We placed the data from the L2WPMA software package as we mentioned its operation in the previous chapter, and we ran it for 24 single modules. We performed the piecewise monotonic method in (Powell, M.J.D. ,1970) data which starting with an estimate of $k = 4$ to $k = 24$ with step 2 on each increase for 3-chlorotyrosine protein. In addition we used for the second dataset we started with an estimate of $k = 4$ to $k = 14$ with step 2 at each increase for one-methylhistidine protein. Finally we used the third dataset that we started with an estimate of $k = 4$ to $k = 16$ with step 2 in each increase for 2-Benzyl protein. We have noticed how the spectrum changes depending on the value of k and which protein is produced by the spectrum. We have explained how from a spectrum we can understand which protein will be produced.

Following the above analysis, we have reached the following conclusions. We first found that when the number of individual segments increases, the average square error of the data decreases. Payed special attention to using successive k values for the same data. If we knew from the beginning the required number of peaks, we could quickly and efficiently have spectrum imaging through the method. Observing Figures 3.1., 3.2., 3.3, 3.4., 3.5., 3.6., 3.7, 3.8, 3.9 we lee that subtle trends in data are not detected for the smaller k values, because their values are rather conservative. The piecewise monotonic approximation method can smooth the data as little as possible to avoid spectrum alteration. In addition, the display of a peak that does not give us any information lies beyond useful tops.

Also this peak does not transform the spectrum and does not alter the results. Finally, detecting the proton structure with NMR data.The goal of NMR spectroscopy analysis is to extract prominent profiles so that we have prototype standards to be able to compare them with other data as well as to find information that will provide us with important information on the diagnosis, prognosis and treatment of specific diseases.

The work of this thesis started on 14-10-2018 and finished on 14-03-2019. The time was very limited for the whole project, because we had to collect data, explain the rationale of the spectra according to the spectroscopy we choose and to attribute the behavior of the protein molecules to the spectra. Also, we have performed the experiment on the data sets and analyze the results. Besides, we have been working on a completely new subject and I needed to develop skills both in the analysis of spectroscopy and the application of the piecewise monotonic method. Finally, I hope that my research will help to find effective ways to identify and structure the proteins. It also helped to find new methods that will improve the algorithm L2WPMA. For example of making the algorithm more efficient is that it does not take it from the user but find the algorithm on its own.

Bibliography

1. Alfa Johnson Matthey Company Johnson Mattheay Catalog Company, Inc. (2009) Alpha-Benzulsuccinate acid (HMDB0012127). Retrieved Apr. 06, 2009, from htpp//www.hmdb.ca.
2. Andrews, Kenneth J. M.; Barber, W. E.; Tong, B. P., (1969).Synthesis of biopterin. Journal of the Chemical Society, (6), 928-30. Retrieved Nov 16, 2005, from http//www..hmdb.ca.
3. Ashenhurst J., (2016). Infrared Spectroscopy: A Quick Primer On Intrpreting Spectra, Master Organic Chemistry.
4. Ball D. W., (1962). The Basics of Spectroscopy, University of Central Florida, SPIE Publications
5. Bouchilloux, Simone, (1955) Chlorination of L-tyrosine.Bulletin de la Societe de Chimie Biologique (HMDB001885).Retrieved Febr. 28, 2006, from htpp//www.hmdb.ca.
6. David S. Wishart, (2007). HMDB: the Human Metabolome Database (PMID:18953024). Retrieved Oct 17, 2006, from http//www.ncdi.nlm.nih.gov.
7. Demetriou I. C. (2015). An Example of Peak Finding in Univariate Data by Least Squares Approximation and Restrictions on the Signs of the First Differences. In Proceedings of the World Congress on Engineering, vol 2, London, UK.
8. Demetriou, I. & Koutoulidis, V. (2013). On Signal Restoration by Piecewise Monotonic Data Approximation, Lecture Notes in Engineering and Computer Science.1. 268-273.
9. Demetriou I. C.(2007), "Algorithm 863: L2WPMA, a Fortran 77 package for weighted least-squares piecewise monotonic data approximation", ACM Trans. Math. Softw., vol 33, 1, pp. 1-19.
10. Demetriou I. C., and M. J. D. Powell(1991), "Least squares smoothing of univariate data to achieve piecewise monotonicity", IMA J. of Numerical Analysis, vol 11, pp. 411-432.
11. Demetriou I. C., (1995). Discrete piecewise monotonic approximation by a strictly convex distance function. Mathematics of Computation, Vol 64, pp 157-180.
12. Elizabeth J., (2002). Endoplasmic Reticulum Stress and the Unfolded Protein Response in Cellular Models of Parkinson's Disease. Journal of Neuroscience, 22 (24) 10690-10698; DOI: https://doi.org/10.1523/JNEUROSCI.22-24-10690.2002. Retrieved Dec 15, 2002, from https://www.ncbi.nlm.nih.gov.
13. Goglia F., (2004). Biological effects of 3,5-diiodothyronine (T(2)), Dipartimento di Scienze Biologiche ed Ambientali, Universita degli Studi del Sannio, Benevento 82100, Italy. Retrieved Feb, 2005, from https://www.ncbi.nlm.nih.gov.
14. Harris RK, (1983). Nuclear Magnetic Resonance Spectroscopy, Longman Scientific Technical, NY.
15. Hill HC, London AG (1975). Interpretation of IR Spectra A Programmer and Introduction.
16. Jain, Rahul; Cohen, Louis A., (1996). Regiospecific alkylation of histidine and histamine at N-1 (t).Tetrahedron (HMDB0000001). Retrieved Nov. 16, 2005, from http//www.hmdb.ca

17. John Wiley & Sons, Chicheste (2004). Stuart B. Infrared Spectroscopy: Fundamentals and Applications.
18. Kokotos (1986). Synthesis and study of substituted coumarins. A facile preparation of D,L-o-tyrosine. Journal of Heterocyclic Chemistry, 23(1), 87-92 (HMDB0006050). Retrieved Apr. 12, 2007, from http//www.hmdb.ca.
19. Napolitano, (1995).Generation of the Neurotoxin 6-Hydroxydopamine by Peroxidase/H2O2 Oxidation of Dopamine. Journal of Medicinal Chemistry, 38(6), 917-22 (HMDB0001537). Retrieved Nov 16, 2005 from from http//www.hmdb.ca.
20. Nicholas C. Thomas, (1991). The Early History of Spectroscopy. Ayburn University at Montgomery, Montomery.
21. Richard O.C. Norman, (2016). Chemical compound. Retrieved Aug. 27, 2008, from htpps//www.britannica.com.
22. Powell, M.J.D., (1970). Curve fitting by splines in one variable. In: Numerical Approximation to Functions and Data (J.G. Hayes, ed.). The Athlon Press, London, pp. 65-83.
23. Orhan H, (2004). Simultaneous determination of tyrosine, phenylalanine and deoxyguanosine oxidation products by liquid chromatography-tandem mass spectrometry as non-invasive biomarkers for oxidative damage. Department of Pharmacochemistry, Division of Molecular Toxicology, Leiden/Amsterdam Center for Drug Research (LACDR), Vrije Universiteit, De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands. Retrieved Jan. 25, 2004, from http//www.ncbi.nlm.nih.gov.
24. Valashek, I. E. (1995). Synthesis of 3,5-diiodo-DL-thyronine. Khimiko-Farmatsevticheskii Zhurnal (HMDB0000582), 29(6), 44. Retrieved Nov 16, 2005, from http//www.hmdb.ca
25. William Reusch, (2013). Spectroscopy. Michigan state university, from http//www.chemistry.msu.edu.
26. Αγγελική Καλπαξή (2012). Μελέτη ομοιομάτων προσομοίωσης χρόνων μαγνητικής αποκατάστασης T_1, T_2 στην απεικόνιση μαγνητικού συντονισμού, Πανεπιστήμιο Πατρών.
27. Δημητρίου Κ. Ιωάννης, (1997). Εμπειρία με τις κατα τμήματα μονότονες προσεγγίσεις διακριτών δεδομένων,Πανεπιστήμιο Αθηνών. Πρακτικά 10ου Πανελλήνιου Συνέδριου Στατιστικής Πειραιάς, ΕΣΙ, 109-123
28. Ζαφειρόπουλος & Βασίλειος (2015). Μέθοδοι Τρισδιάστατης Απεικόνισης: MRI & CT, ΤΕΙ Κρήτης, Εκδόσεις Κάλλιπος.
29. Δρίτσα Αθηνά (2002). Φαματοσκοπία πυρηνικού μανγητικου συντονισμού-NMR, ΤΕΙ Καβάλας.
30. Κουφόπουλος Ν., (2015). Ανακάλυψη βιοδεικτών με τη χρήση NMR και MS μεταβολομικής. Εφαρμογή στο διαβήτη. Εθνικό Μετσόβιο Πολυτεχνείο. Άρτεμις +
31. Λιαροκάπης Ε. & Τσιακλάγκανου Π. (2012). Φασματοσκοπική μελέτη Raman νημάτων πολυεθυλενίου, Nulon και νανοσωληνών άνθρακα, Αθήνα.
32. Μιχάλης Φαρδής (2017) . Εισαγωγή στον παλμικό πυρηνικό μαγνητικό συντονισμό (NMR), Ινστιτούτο Νανοεπιστήμης & Νανοτεχνολογίας ΕΚΕΦΕ Δημόκριτος.
33. Ράπτης Ι. (2016). Φαματοσκοπία Raman, Εθνικό Μετσόβιο Πολυτεχνείο.