4-2016

Compressive Sensing Framework for Mass Spectrometry Data Analysis

Khalfalla Ahmad Kh. Awedat
Western Michigan University, khalfec2oo1@gmail.com

COMPRESSIVE SENSING FRAMEWORK FOR MASS SPECTROMETRY DATA ANALYSIS

by

Khalfalla Ahmad Kh. Awedat

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Electrical and Computer Engineering
Western Michigan University
April 2016

Doctoral Committee:

Ikhlas Abdel-Qader, Ph.D., Chair
Johnson Asumadu, Ph.D.
James Springstead, Ph.D.

COMPRESSIVE SENSING FRAMEWORK FOR MASS SPECTROMETRY DATA ANALYSIS

Khalfalla Ahmad Kh. Awedat, Ph.D.

Western Michigan University, 2016

Mass Spectrometry (MS) data is ideal for identifying unique bio-signatures of diseases. However, the high dimensionality of MS data hinders any promising MS-based proteomics development. The goal of this dissertation is to develop an accurate classification tool by employing compressive sensing (CS). Not only can CS significantly reduce MS data dimensionality, but it also will allow for full reconstruction of original data. The framework developed in this work is based on using L2 and a mixed L2-L1 norms, allowing an overdetermined system to be resolved. The results show that the L2-based algorithm with regularization terms has a better performance than that of L1 and Q5 algorithms under all applicable assumptions. Performance was measured using overall success rate, sensitivity, positive predictive value and specificity. The regularization parameters and sensing matrix were optimized to achieve a robust classification method. Additionally, the Block Sparse Bayesian Learning (BSBL) algorithm was used to reconstruct MS data using a fingerprint-based technique. The simulation results validate the proposed framework and indicate the potential for a

successful prostate cancer detection technique using MS data. The proposed framework can be a useful tool for assessing patient risk of disease and will aid in paving the way for personalized medicine.

ACKNOWLEDGEMENTS

I would like to being by thanking God for providing me with the strength and patience to work towards fulfilling my goals now, and throughout my life. It is due to His blessings that I am here today.

An immense thank you and deep heart-felt appreciation goes to my supervisor, Dr. Ikhlas Abdel-Qader , who has constantly guided me during all phases of my research. Her continuous support, encouragement and astute recommendations have all played an invaluable role in my research.I would like also to thank my committee members: Dr. Ikhlas Abdel-Qader, Dr. Johnson Asumadu and Dr. James Springstead for their support and supervision during this dissertation process.

Additionally, I must thank my mother who has tirelessly supported me during my studies throughout my years. Thank you, Mom, for your continuous prayers and words of comfort, both of which have given me the courage to follow my dreams and always reach for success.

Finally, I would like to thank my beloved wife, Shirihan, who has always been there for me, both in good times and difficult times. I would like to thank her for her unwavering patience, support and encouragement. She has always believed in me and has sacrificed a lot to allow my academic dreams to come true. Thus, I am dedicating this dissertation to her and our lovely children: Abdlukdoos, Renad, Robeen and Taher.

Khalfalla Awedat

TABLE OF CONTENTS

Table of Contents- Continued

CHAPTER

Table of Contents- Continued

LIST OF TABLES

LIST OF FIGURES

List of Figures- Continued

List of Figures- Continued

CHAPTER I

INTRODUCTION

High-throughput proteomics techniques, such as mass spectrometry (MS) based approaches, are increasingly gaining interest in many applications, including bio-marking discovery and drug devolvement. They provide a powerful technique for analyzing protein mixtures [1] [2]. Protein identification is an essential step in the proteomics field. The identification pattern of protein can lead to important discoveries such as the classification of sample bases of a particular pattern. The general procedure of MS-based approaches to protein identification involves digesting protein into small pieces (peptides) to simplify the protein chemistry. The resulting peptide mixture is separated in time according to the peptide's chemical and physical properties. Next, the MS analysis is performed on the individual peptides, and a sub-sequence (spectrum) that represents a large portion of full data is selected. The data analysis task consists of interpreting the sequences, by comparing the spectra to candidate peptides [3].

The mass spectrometry (MS) produces very high-dimensional data-sets. Therefore, an essential feature of a robust interpretation algorithm is the accurate evaluation of the quality of the matches between spectra and peptides; that is, an estimation of the likelihood of correctness. The reliable identification of proteins from mixtures using mass spectrometry would provide an important tool in both biomedical research and clinical practice.

- Given a complex protein mixture, this method identifies the proteins using mass spectrometry techniques by matching the spectra to the theoretical spectra which are produced from candidate peptides.
- In a clinical setting one is often interested in how MS spectra differ between patients of different classes, for such as spectra from healthy patients vs. spectra from patients having a particular disease. Many disease relevant mechanisms are controlled by proteins (e.g. hormones) which can be detected in biological samples (blood, urine, etc.) using mass spectrometry (MS). Mass spectrometry allows (potentially) for monitoring the entire set of proteins in a given sample.

  To identify unknown proteins and locate multiple proteins present in a complex mixture, or to find significant differences in the data between samples from healthy and diseased individuals, the following steps have to be achieved:

  1. Given an organ or non-organ sample; find the spectrometry representation which depends on the sample.
  2. Classify unknown spectra using a database, which identifies according to a prior knowledge of the sample.

The high dimensionality of MS data and noisy spectra produced from MS are two challenges one has in utilizing this procedure. Since the acquired data is usually noisy,

the algorithms should be robust to noise, and the identified feature set should be as small as possible [4].

1.1. Significance of the Research

Since mass spectrometry data is biologically valuable for its ability to define protein peptides and clinically valuable for its ability to differentiate between healthy and diseased samples, MS data acquisition and analysis has gained significant importance over the past years. Of paramount significance and challenge is that MS data comes with high dimensionality: it consists of tens of thousands of m/z ratios and an intensity level for each m/z ratio. For example, currently, a low resolution SELDI-TOF MS (Surface Enhanced Laser Desorption/Ionization Time of Flight Mass Spectrometry) can measure up to 15500 data points that record data between 500 and 20000 m/z ratios. With a high resolution MS, the data points could be 400000 [5]. Being of such high dimensionality, MS data classification is computationally complex. Efforts are focused on improving classification while reducing computations. In addition, MS compressive sampling can be a major change in the field, and thus researchers are targeting techniques for MS reconstruction using fewer data [4] [6] [7].

1.2. Target Goals

MS sensing data analysis using compressive sensing (CS) techniques is the focus of this dissertation. Compressed data will have a very low dimensionality when compared

with the original MS data. The ability to produce an accurate prediction of class content in a sample of compressed MS data is one of the main goals of this work. Moreover, since the original has been projected onto a different low-dimension space, the second goal is to investigate the recovery of an estimate of the high dimensionality data based on the prior knowledge of MS sensing data. Since MS data is not sparse and the CS is guaranteed to recover the sparse solutions only under certain constraints, the sparse difference (SD) method will be applied to get a sparse representation of MS data.

1.3. Research Questions

In order to achieve these goals, this dissertation addresses the following questions:

- Can the classification process of MS data be achieved under CS framework, and how much accuracy can be achieved?
- Can we find a good method to reconstruct high quality CS data from sparse MS data?
- What are the advantages of a proposed algorithm over the present techniques?
- What are effective parameters in proposed algorithms that can play a role in drawing a robust classification and recovery? Can these parameters be optimized?

1.4. Organization

Chapter 1 introduces the objectives, significance of research, target goals and research questions of this project. Introduction to Mass spectrometry, Mass Spectrometry for Proteomics and Compressive Sensing are presented in Chapter 2. Chapter 3 explores the pertinent literature of general methods that have been used for MS classification and some classification methods based on a CS framework. Chapter 4 introduces the proposed classification and recovery algorithms while Chapter 5 discusses the dissertation's results followed by a conclusion and future work in Chapter 6.

CHAPTER II

INTRODUCTION AND BACKGROUND

2.1. Mass Spectrometry (MS)

2.1.1 Mass Spectrometry Overview

Mass spectrometry (MS) is an analytical tool used to measure or determine the elemental composition of biological or chemical compound in order to measure the molecular mass of a sample. This task is achieved by generating ions related to their mass-to-charge ratio and detecting them qualitatively (with high resolution) and quantitatively by their respective mass (m) -to-charge (z) ratios (m/z)* and abundance [8]. All mass spectrometers consist of three essential parts: ionization, separation and detection of ions in their gas phase. Since the analyte in gas form is not always available, a wide variety of mass spectrometers have been created to deal with different needs and applications. For example, the MS coupled to liquid chromatograph (LC) has enough separating power to be highly efficient and to identify the mass characteristic of a complex mixture. The ionization source in most cases is the interface between LC and MS. Figure (1) shows the schematic diagram of the LC-MS instrument [9].

* Usually use Dalton (Da) or Thomson (Th) as the atomic mass unit. 1 Th=$1Th = 1\frac{u}{e} = 1\frac{Da}{e} = 1.036426x10^{-8}kgC^{-1}$.

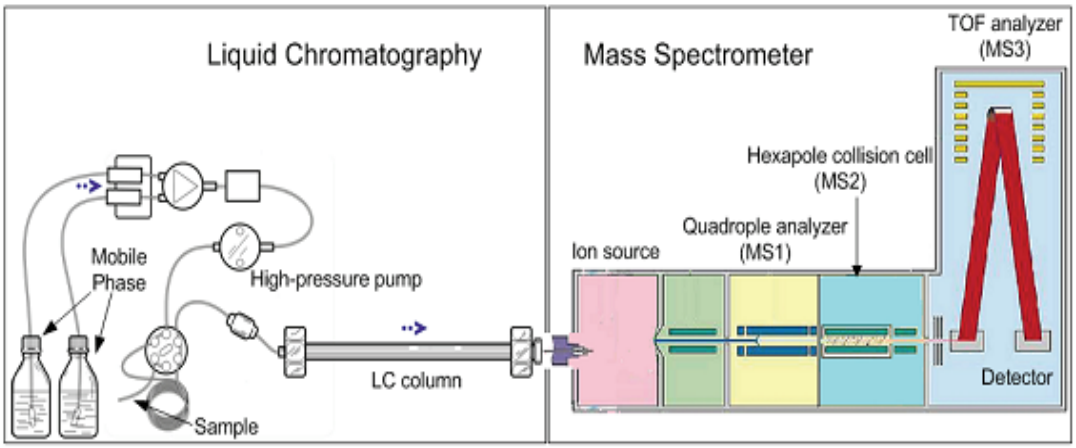


Figure 1. A Schematic diagram of the LC-MS instrument. Source [6]

Depending on classes compounds and type of information needed, several ion sources and mass analyzers have been established and applied to LC/MS [9]. The separation sample components are generated using the LC-column, which has certain hydrophobic properties, so the peptides of samples are kept according to their physiochemical properties. Depending on the chromatography mode, the peptides elute from the column at different times and separate from the sample. This is called retention time (RT). Next, the separated sample species entered MS, and the following processes are achieved for each eluted peptide:

1. The peptides are converted to ions in the gas form by spraying them into an atmospheric pressure ionization source (API).

2. The ionized peptides are accelerated so that they all have the same kinetic energy.
3. The peptides are then deflected by a magnetic or electrical field according to their mass-to-charge ratio. The lighter they are the more they are deflected, and a mass analyzer sorts the ions.
4. The detector is used to count the ions which emerge from the mass analyzer.

The mass spectral data provides valuable information about identity, quantity, structure and purity of sample [10].

The measured data can be depicted in a 2D or 3D image where the retention time is the first axis, the m/z value is the second axis and the third axis represents the abundance intensity as shown in . The completed procedure is called an LC/MS run [1] [2].

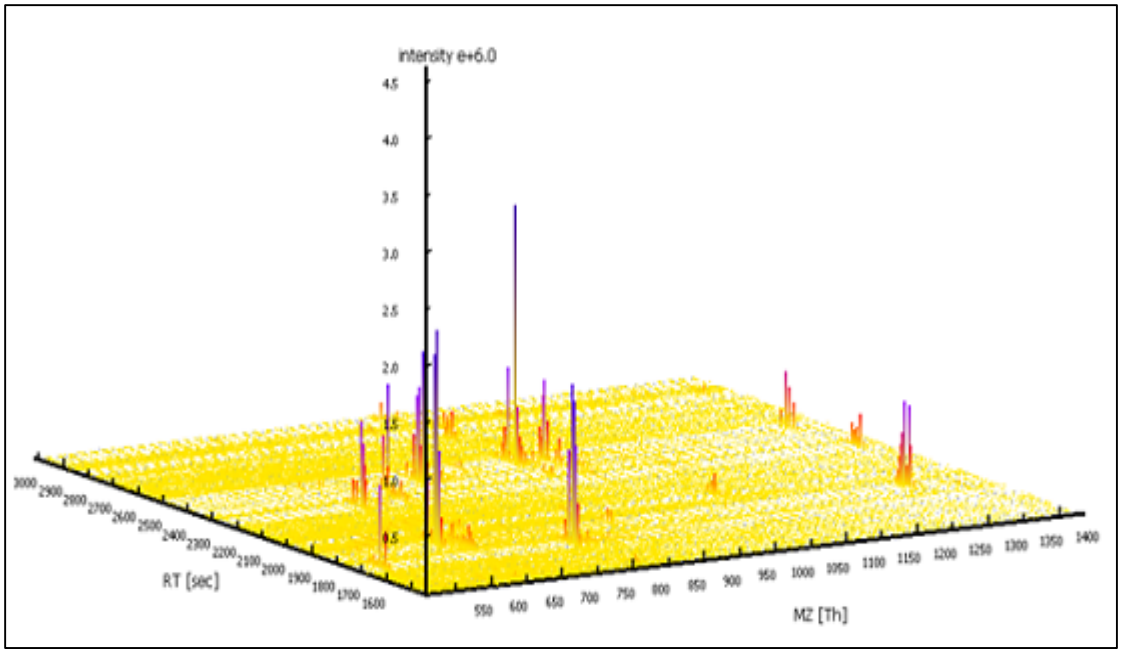


Figure 2. LC/MS dataset showing abundance versus time (sec) and m/z (Th) in a 3D illustration using OpenMS software

2.1.2Mass Spectrometry Applications

The mass spectrometry has been engaged in many applications and fields such as clinical testing, drag discovery, biotechnological development and environmental analysis. Some common applications of MS are listed in Table (1).

Table 2. Common applications of MS source [11]

Field of study	Application
Proteomics	▪ Determine protein structure, function, folding and interactions ▪ Identify a protein from the mass of its peptide fragments ▪ Detect specific post-translational modifications throughout complex biological mixtures ▪ Quantitate (relative or absolute) proteins in a given sample ▪ Monitor enzyme reactions, chemical modifications and protein digestion
Clinical Testing	▪ Perform forensic analysis such as confirmation of drug abuse ▪ Detect disease biomarkers (e.g., newborns screened for metabolic diseases)
Genomics	▪ Sequence oligonucleotides
Environment	▪ Test water quality or food contamination
Geology	▪ Measure petroleum composition ▪ Perform carbon dating

2.1.3. The Tandem Mass Spectrometer (MS/MS)

Tandem mass spectrometry (MS/MS) refers to more than one stage of MS analysis, with molecular fragmentation occurring between stages. The basic form of

MS/MS combines two mass spectrometers as shown in Figure 3. The first system (MS-1) is used to select mass (parents) that are a characteristic of target ions from the stream of ions of a mixture. The parents pass to the region where they are broken down to produce the fragment ions. The parent ions collide with the neutral gas molecules (usually argon or nitrogen) which are filled in the collision cell. This process, called Collision Induced Dissociation (CID), generates a series of product (daughter) ions by breaking certain chemical bonds in the molecular ions. The second MS system (MS-2) is used to separate the fragment ions according to the mass [1]. The resulting MS/MS spectrum provides information about the remaining parent and product ions. The MS/MS spectrum is of particular importance in that it provides a characteristic fingerprint for different compounds even if these compounds have the same elemental formula.

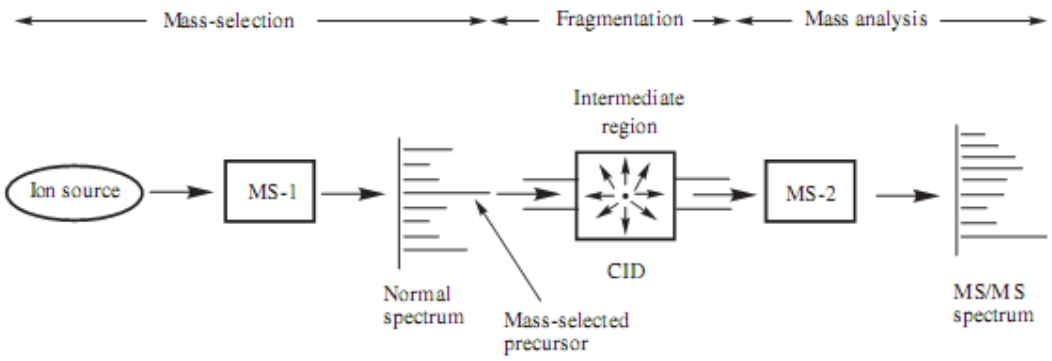


Figure 3. Basic principle of tandem mass spectrometry [1]

Usually, the analyte sample is very complicated, and protein is one such example. In this case the produced spectrum from the second system run consists of masses of a large

number of small fragmented peptides. These peptides are composed of a list of peaks, and each peak is represented by the measured m/z and an intensity value that represent the number of fragments. Moreover, those peaks identify the proteins to which each peptide belongs [11].

2.2 Mass Spectrometry for Proteomics

Proteomics refers to a systematic identification of protein structures and their quantity, activity and molecular interaction. The mass spectrometry is a power tool for protein identification in a biological system. Proteins are composed of peptides, which are chains of Amino acids. There are 20 basic amino acids, uniquely abbreviated with a single letter. Peptides thus can be described as a string of the letters corresponding to the amino acids. Most of the amino acids have distinguishable masses (Table 2), which makes peptide identification by MS/MS spectra possible. Proteins and peptides are defined by their unique sequences of amino acid residues [12].

Table 3. Names and abbreviations of the standard amino acids with their masses

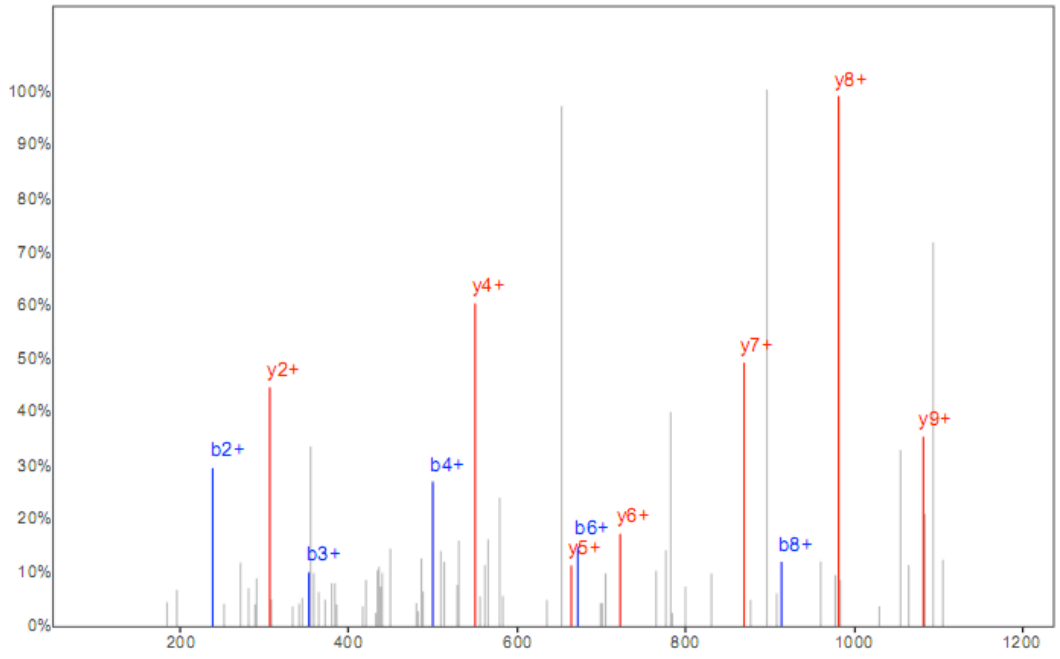
Amino acid	Abbreviation	mass	Amino acid	Abbreviation	mass
Alanine	**A**	**71.04**	**Methionine**	**M**	**131.04**
Cysteine	**C**	**103.01**	**Asparagine**	**N**	**114.04**
Aspartic acid	**D**	**115.03**	**Proline**	**P**	**97.05**
Glutamic acid	**E**	**129.04**	**Glutamine**	**Q**	**128.06**
Phenylalanine	**F**	**147.07**	**Arginine**	**R**	**156.10**
Glycine	**G**	**57.02**	**Serine**	**S**	**87.03**
Histidine	**H**	**137.06**	**Threonine**	**T**	**101.05**
Isoleucine	**I**	**113.08**	**Valine**	**V**	**99.07**
Lysine	**K**	**128.09**	**Tryptophan**	**W**	**186.08**
Leucine	**L**	**113.08**	**Tyrosine**	**Y**	**163.06**

2.2.1 Fragmentation Process

After the spectra have been generated in MS/MS using CID, the peptide in each observed MS/MS spectrum is identified using a computational method. As the fragmentation process will occur at different bond positions, the multiple peptides will be generated. The acquired peptides result in the formation of two fragments. One contains an N-terminus, known as the b-ion, and the other is a C-terminus called a y-ion. Each consecutive peptide fragment is called a ladder; they differ in mass by a single amino acid. Figure 7a-b illustrates the ladder of b- and y-ions, using a peptide of ten amino acids as an example along with the experimental spectrum.

mass	ion	Sequence	Sequence	ion	mass
138.0662	*b1*	**H**	TLFGDELCK	y9	
239.1139	b2	**HT**	LFGDELCK	y8	1082.5187
352.1979	b3	**HTL**	FGDELCK	y7	981.4710
499.2663	b4	**HTLF**	GDELCK	y6	868.3869
556.2878	b5	**HTLFG**	DELCK	y5	721.3185
671.3148	b6	**HTLFGD**	ELCK	y4	664.2971
800.3573	b7	**HTLFGDE**	LCK	y3	549.2701
913.4414	b8	**HTLFGDEL**	CK	y2	420.2275
1073.4721	**b9**	**HTLFGDELC**	**K**	**y1**	**307.1435**

(a) The ladder of b- and y-ions



(b) An observed spectrum

Figure 4. An illustration of fragmentation for peptide HTLFGDELCK from atlas data (a) the ladder of b- and y-ion, (b) the observed spectrum of this peptide

2.2.2 Peptides Identification

The peptides are molecules of a chain of amino acids; they are similar to protein but shorter. The MS/MS is the most common tool to identify the peptides. An unknown peptide is registered in fragment masses as a spectrum. Then the computational methods conclude the peptide sequence from its spectrum [13]. There are two main search methods: De novo and database search methods. De novo sequencing methods utilize the approach to deduce the sequence of a peptide or part of it directly from MS/MS by finding the mass differences between peaks that correspond to the amino acids. There are a few factors that can cause difficulty, but mainly include missing fragment ions and the existence of noise peaks in the spectrum [14] [15]. The database search algorithm is the most common strategy used for peptide identification. For each protein sequence in a database, it is possible to completely collect peptide sequences with their theoretical fragmentation spectra, which happens by comparing each acquired MS/MS spectrum against those obtained from the sequence database and using various scoring schemes to find the best match peptides. The result is a list containing the top scoring solution [13]. Figure 8 illustrates the database procedures.

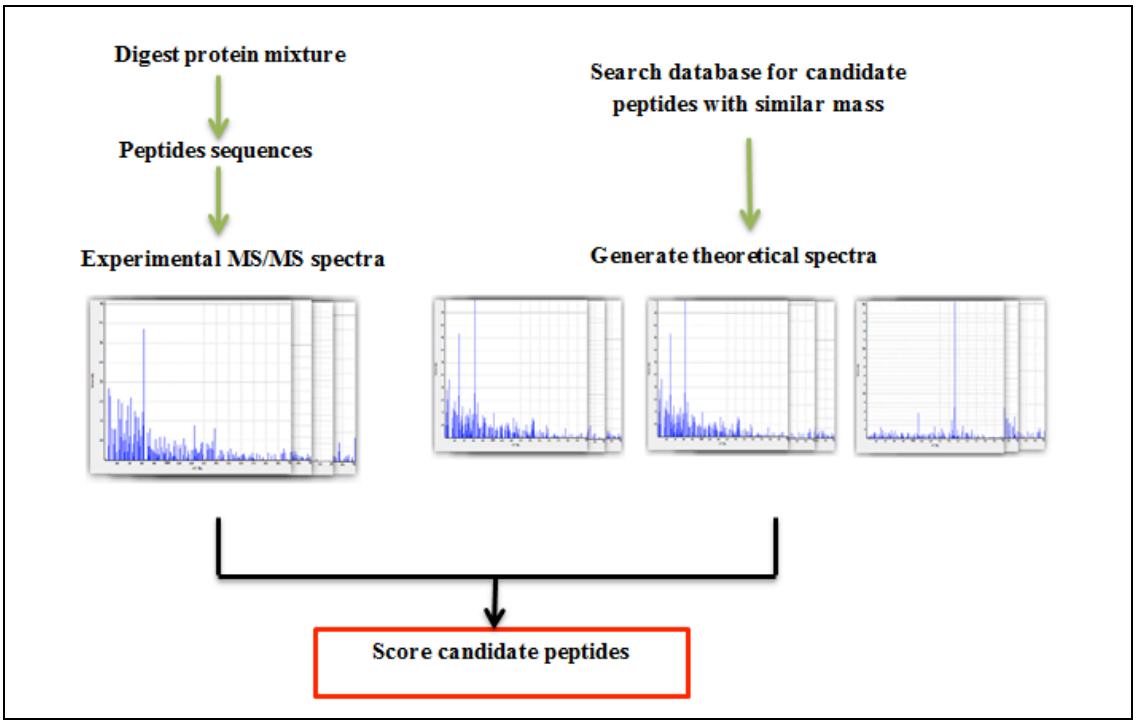


Figure 5. General framework for peptide identification using database search revised from [16]

2.2.3 Protein Identification

The ultimate goal of a proteomic approach is to identify the proteins that are presented in a biological sample. In a database approach, based on peptide sequences which are selected as candidate peptides, the experimental spectra using MS/MS are matched to theoretical spectra. As a result, each observed spectrum is mapped to one or more peptides. These peptide-spectrum matches (PSM) are grouped according to their corresponding protein entries in the proteins database. For each protein, its identified

peptides are used to assess the evidence of its presence in the sample. The proteins can be ranked according to the number of peptide matches [17] [18].

2.3 Compressive Sensing (CS)

2.3.1 Introduction

In general, compressive sensing (CS) has been built upon the fact that signals can be represented by only a few coefficients in a suitable basis. A nonlinear optimization algorithm is applied to recover the signals from very few measurements. The signals which have this representation are called compressible or sparse signals. In order for digital systems to be able to recover or demodulate the transmitted signal, the Nyquist rate should be satisfied (F_s of at least the double of the signal's highest frequency); however, the nonlinear recovery using CS requires the sample frequency to be smaller than F_s because only a few coefficients of the signal will transmit [19].

In high dimensionality data, compression is most commonly approached by finding a new basis set where the signal can have a sparse representation; that is the signal in $\mathbb{R}^N$becomes in $\mathbb{R}^K$ where K<< N are the highest coefficients. The small coefficients (N-K) will be discarded; the number of neglected coefficients depends on the target level of acceptable distortion. For instance, JPEG standard indicates that the small percentage of measurements coefficients are needed to reconstruct the signals such as in the methods used in transform coding.

The CS comes as a smart and inexpensive method that only requires a few measurements from a large amount of data, and the reconstruction process with high quality has been guaranteed. The CS has emerged as a new framework for signal acquisition and sensor design. The CS is able to reduce the measurements and sampling rate. The scenario is to sample the signal in compressing form [20]. Limited bandwidth signals having sparse coefficients can be recovered from a small set of linear, non-adaptive measurements. The challenges in CS are designing these measurements to fit with the practical models and acquisition systems [19] [20].

2.3.2 The Classical Sampling Versus Compressed Sensing

There are three differences between the sampling scheme and CS [21]:

a) Sampling theory focuses on continuous-time signals. In contrast, the CS is a mathematical model focused on measuring finite-dimensional discrete vectors.
b) Rather than sampling a signal at specific points in time, the CS system typically acquires the samples (measurements) using the inner product between the signal and coding function (sensing matrix), and the measurements are selected randomly.
c) The two frameworks are different in the manner in which they deal with signal recovery. The recovery is achieved by a linear process through $sinc$ interpolation that requires a simple interpretation for the sampling scheme while non- linear recovery methods are used in CS.

2.3.3 Signal Representation

The signal $x \in R^N$ can be represented by a complete dictionary $D = \{d_1, d_2, \ldots\ldots\ldots, d_{Md}\}$, as columns with N length, $\|d_i\| = 1\ \forall i$ using a linear expansion of d_i by the following equation

$$x = D\boldsymbol{\alpha} = \sum_{i=1}^{Md} \alpha_i d_i \qquad (1)$$

Where,M_d is the number of dictionary elements, $\alpha's$ are coefficient vectors which are limited by the number of M_d columns or atoms in D.

Note that Md may be larger than N, and the dimensionality of the original signal space [20]. In general, for the decomposition represented in the equation (1), we can identify a signal characteristic from the large number of dictionary atoms. Dictionaries in which the number of structures is higher than the dimension of the signal are called over-complete dictionaries; in this case, the dictionary is redundant because there is more than one way to represent the same signal. If $Md = N$, then we have the complete dictionary, and the coefficient vector represents a solution of decomposition of x by the basis D. The complete dictionaries have been performed in many decomposition algorithms such as Fourier Transform, Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT). However, there is not enough flexibility to guarantee a perfect representation for the signals that have vary wide localization in frequency and time. For

instance, the DTW has a good high frequency representation, but it has a poor representation in low frequency during a high frequency basis function. DCT basis functions have a fixed time frequency resolution. The over-complete dictionary is one tool to deal with this kind of problem. In this way, one can have elements representing several compromises between time and frequency resolutions or signals in other space providing more flexibility [19].

2.3.3.1. Adaptive Representation of Signals

Adaptive representations are referred to when choosing from a large dictionary where atoms are used to represent a signal. The term adaptive is for a choice which is signal dependent. Coding, compression to signal enhancement, de-noising and pattern recognition are some applications where the adaptive representations have been applied [22]. The adaptive representation is achieved by decomposing the signal into a set of waveforms. This set is often referred to as a dictionary; all the waveforms have the same number of elements or atoms. Depending on the atoms being larger than or equal to the dimension of a signal, the dictionaries can be over-completed or completed as dependent on a dictionary type. The dictionaries contain an infinity atom for the continuous signal.

The over-complete dictionary elements could be represented with other elements; therefore, the dictionary's decomposition is non-unique. For that property, the adaptation comes from choosing the representations among many choices. In over-complete dictionaries; the signal can be expressed using a small number of coefficients. However,

choosing these coefficients is difficult and needs a complex algorithm. Finding the best representation of a signal using a redundant dictionary, the dictionary's size must be analyzed. The richer dictionary guarantees a small number of values necessary to represent a signal [23]. Due to redundancy, there are many ways to represent the same signal. To develop most techniques is to find a representation where the small number of coefficients contain the majority of the signal's energy; this is called sparse representation, i.e. representation with a larger number of zero coefficients.

2.3.3.2. Approximation on an Over-Complete Dictionary

The sparse representation using redundant dictionaries can improve compression schemes and noise reduction, but it also improves the resolution of an inverse problem such as source separation and compressive sensing [22]. For an N dimensional signal, $x \; and \; M_d$ is a size of a dictionary$D$. If M is the size of the output signal, where $M < N < M_d$ then finding the representation of the approximating signal takes the form:

$$x_M = \sum_{m=0}^{M-1} \alpha_m d_m \quad that\ minimize\ \|x - x_M\| \tag{2}$$

However, this problem is combinatorial and NP-hard. Thus, several methods have been developed to reduce computational complexity by searching for the efficient but not the

optimal approximation. The following list describes the two most popular algorithms: Basis Pursuit (BP) and Matching Pursuit (MP).

- **Basis Pursuit (BP)**

The principle of the BP technique was proposed by Chen and Donoho [24] for solving the following convex optimization problem for a signal whose coefficients have minimal L_1:

$$\min\|\alpha\|_1 \qquad subject\ to \sum_{m=0}^{M_d-1} \alpha_m d_m = x \tag{3}$$

where α is a vector $\in \mathcal{R}^{M_d}$ that contains the α_{M_d} coefficients. The BP decompositions are much sparser because L_1 is non-differentiable, and the decomposition on an optimal basis and the orthogonal basis are not necessary. Therefore, the idea behind this technique is that L_1-norm enhances sparsity. A good approximation strategy provides the optimal value of α vector, $M_d - size$ for M largest coefficients.

- **Matching Pursuits (MP)**

Mallat and Zhang proposed an MP algorithm [25], which decomposes a signal into a linear expansion of waveforms that are selected from the redundant dictionary. An atom or element selection step and a residual update step are completed in each iteration of the algorithm. The atom selection step finds the element that has the highest

correlation with the current residual error while in the update step; we update the residual error by subtracting the correlated rom it. The algorithm terminates when a halting criterion is satisfied, such as when the norm of the residual falls below a desired approximation error bound [26].

2.3.4. Signal Compression and Sparsity

The uncompressing signals such as audio, video and images require a vast amount of data to be stored and transmitted which are usually limited resources. The task is to find a sophisticated scheme to represent the signal with essential measurements or coefficients which satisfy an acceptable and reliable resolution level [27]. Fortunately, this is possible because the signals in general contain redundant information. The approximation signals are done using techniques, such as a compression scheme, that change the signal representation to minimize the redundancy. The solution is to find a sparse representation where the information is concentrated in few coefficients, and the remaining coefficients are zeros. In general, the signals are not sparse, so one may use the threshold value to discard the coefficients below the value. If that is accomplished, the dimension of stored or transmitted signals will be significantly reduced [23] [28].

Fourier and wavelet bases are the starting point to decompose the signals in a sparse representation while revealing the whole signal properties. Over a discrete signal, Fourier transform is decomposition in a discrete orthogonal basis $\left\{e^{j2\pi kn/N}\right\}\ for\ 0 \leq k < N$, which has properties similar to Fourier transform on function. For a signal x with

dimension N and $\psi \in \mathcal{R}^{NxN}$ with orthonormal basis (i.e DFT, DCT, DWT), decompose s as superposition atoms in a new domain

$$s = \psi x \tag{4}$$

where $s \in \mathcal{R}^N$ is the new base with sparse representation. For example, the 259x194 grayscale is shown in Figure 6a. Figures 6 b, c and d show the compressible signal in different bases: DFT, DCT and DWT. What one can observe from this representation is that the most significant coefficients are concentrated in a few non-zero elements where the most of the remaining elements are very small.

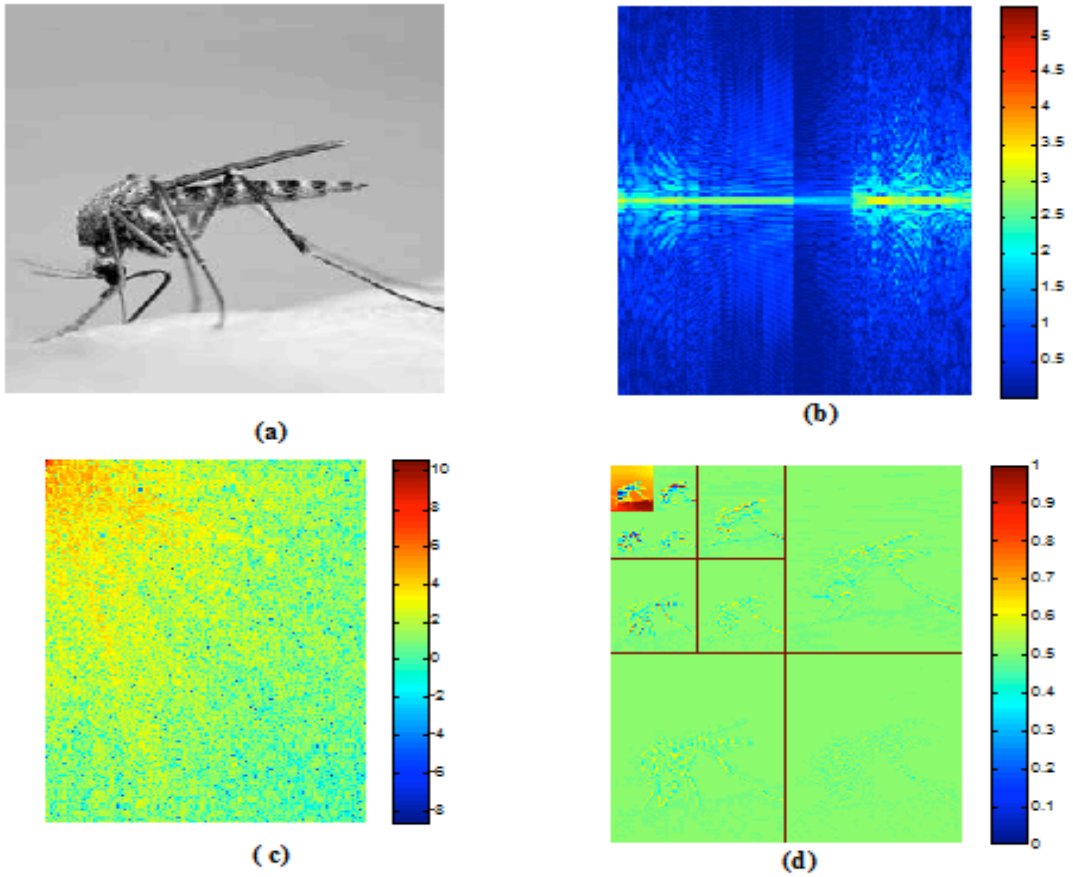


Figure 6. Signal sparse representation (a) the original signal,(b,c,d) DFT, DCT and DWT coefficients for the signal

Figure 7 shows the wavelet coefficients, which are very clear with just a few large elements comparing with all coefficients. When the coefficients values are sorted from the largest to smallest, there is a sharp descent. A power low decay is usually exhibited for compressible signals.

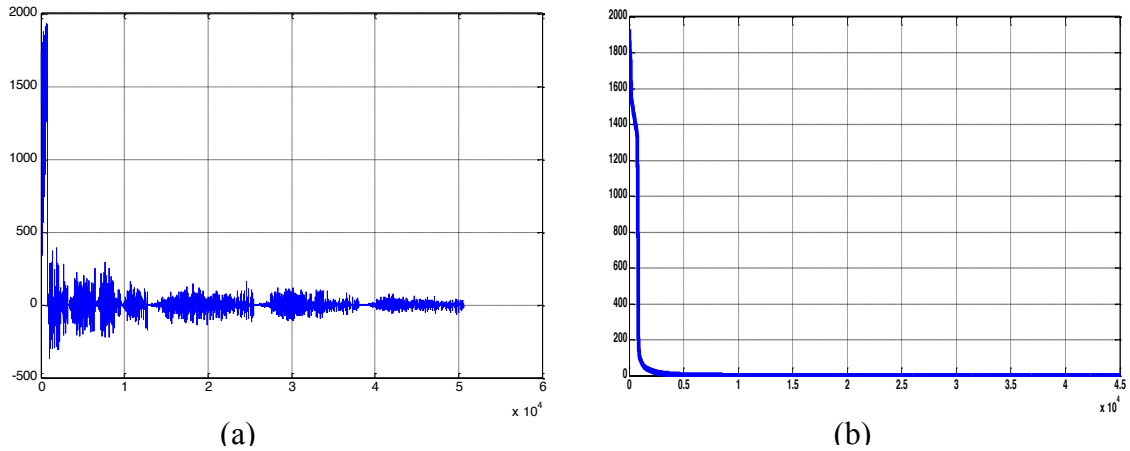


Figure 7. (a) Shows the wavelet coefficients of the image in figure (8) whereas (b) shows the pixel in descent

In the transform-based function, as shown in equation (4), we assume that the representation signal, which usually is a sparse signal, is given and the goal is to find the approximation signal after discarding some coefficients. Therefore, for a given orthonormal basis ψ, we represent $x := \psi^T s$ because by choosing ψ , the decay rate will be affected. More generally, $x(m)$ is selected with m-term as a best approximation vector where $m < N$; then the retained energy ratio can be calculated for m-term entries of a vector x [20].

$$e_m(\psi x) = \frac{\|x(m)\|_2}{\|x\|_2} \tag{5}$$

The error of approximation can be expressed in m-term of $x(m)$ as:

$$\|x(m) - x\|_2 = \|\psi(x(m) - x)\|_2 = \|s(m) - s\|_2 \leq \boldsymbol{\varepsilon}_m(\psi x) \qquad (6)$$

Figure (8) illustrates the compressed image in Figure 6 with different bases ψ. Where the vertical axis is $e_m(\psi x)$ as a function in m, the horizontal axis is the percentage of $\frac{m}{N}$. As one can observe, in DCT and DWT the energy concentrates in just a few large coefficients. With the same number of coefficients, the orthonormal matrix is more accurate than the identity matrix.

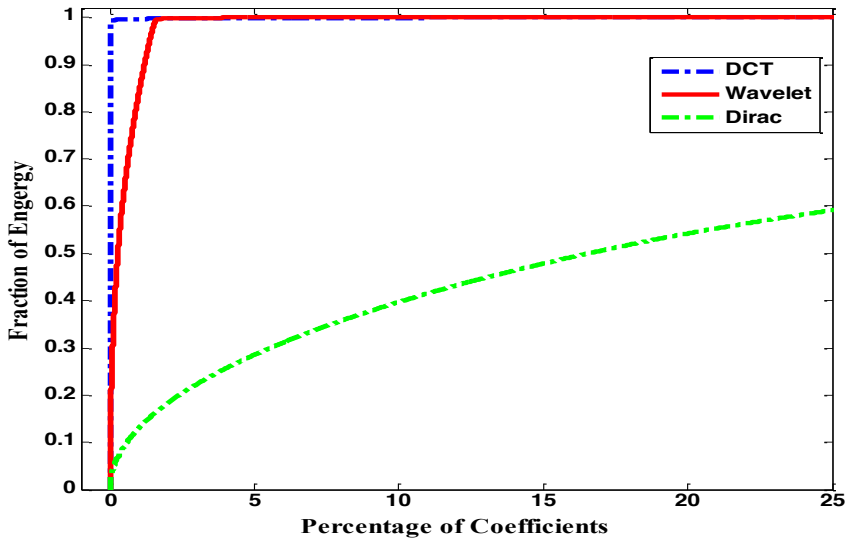


Figure 8. The cumulative energy contributed by sorted coefficients in different bases

2.3.5. Compressive Sensing Overview

In most applications of signal processing, the compression scheme is done by collecting data. The first step is sampling according to the Nyquist criteria and then

projecting the signal into another domain (e.g., Fourier, Wavelet and Cosine transform) where it has simple representation. This simple representation can be obtained because most coefficients of the signal in a new domain are very small compared with large coefficients (Sparse). This is what happens in the most popular compression standard and acquisition instruments. That indicates that only a small percentage of measured coefficients are required to reconstruct a signal and its lossy technique; therefore, the efficiency is lost. In addition, the sampling procedure in many practical analog instruments is too high so that many samples must be taken to get an exact recovery. It may be too costly to build a capable device to acquire the sample at the necessary rate [19] [29].

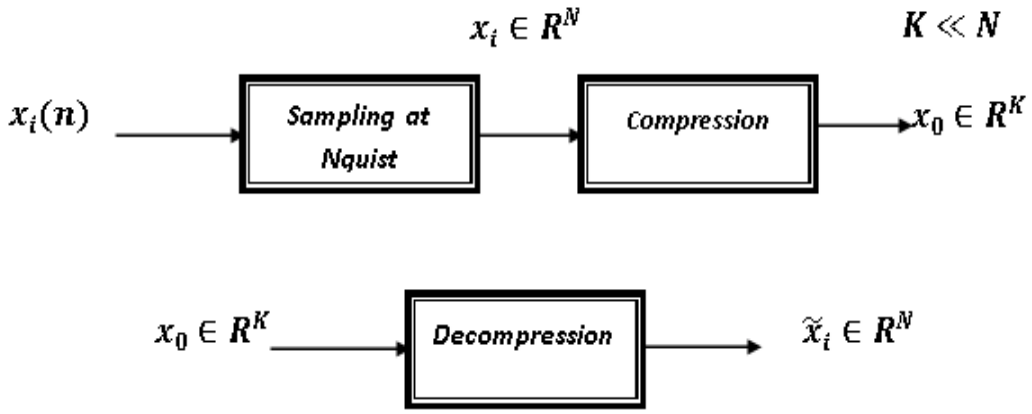


Figure 9. Compression procedures

The sampling procedure would be more effective if it is taken directly from a signal at a sparsely transformed domain. Once that is provided, the performance of the system will improve. However, the transformer domain should be known prior to this

procedure; thus,, a sophisticated and smart algorithm will be required to reconstruct the original signal from a few coefficients with the optimization procedure.

2.3.5.1. Essential Aspects

Building a scheme captures the data already in a compressed form which starts with a general linear measurements of a signal [20]. The inner product is used to acquisition any measurement y_m between the signal $x(n) \in \mathbb{R}^N$ and test function ϕ_m .

$$y_m =< x, \phi_m > \quad where\ m = 1,2,3, \dots \dots N \tag{7}$$

Even though we know that x is sparse in some domains, the number and position of sparse coefficients are not known. If $\phi_m$is a unity matrix with an NxN dimension, then the input and output are the same $y = x$. Moreover, a non-adaptive solution is needed to apply the same mechanism to collect the data from any signal.

2.3.5.2. Algebraic Formulation

Assume that we have $x(n)$as signal and a ψ is the transformation matrix that makes $\boldsymbol{s}$ sparse in another domain

$$y = \phi x \tag{8}$$

$$s = \psi x \tag{9}$$

To take small measurements, let M be the random number of measurements, $M \ll N$ in a test function ϕ_Ω matrix (number of columns are larger than rows). From equations (8) and (9) we can write that:

$$y = \theta_\Omega \boldsymbol{s} \ \boldsymbol{where} \ \theta_\Omega = \phi_\Omega \psi^T \tag{10}$$

where $\phi_\Omega \ and \ \theta_\Omega$ are the measurement signals in matrices formed in raw data and transform domains respectively, and MxN is the size of these matrices where $M \ll N$. Hence, our reconstruction problem can be focused on finding x for a given measurement y. This problem is ill posed (a problem which may have more than one unique solution) because there are an infinite number of possible solutions. All the same, not all solutions satisfy the sparsity property of $\boldsymbol{s}$. Therefore, we need to search among all solutions for the sparsity property [30].

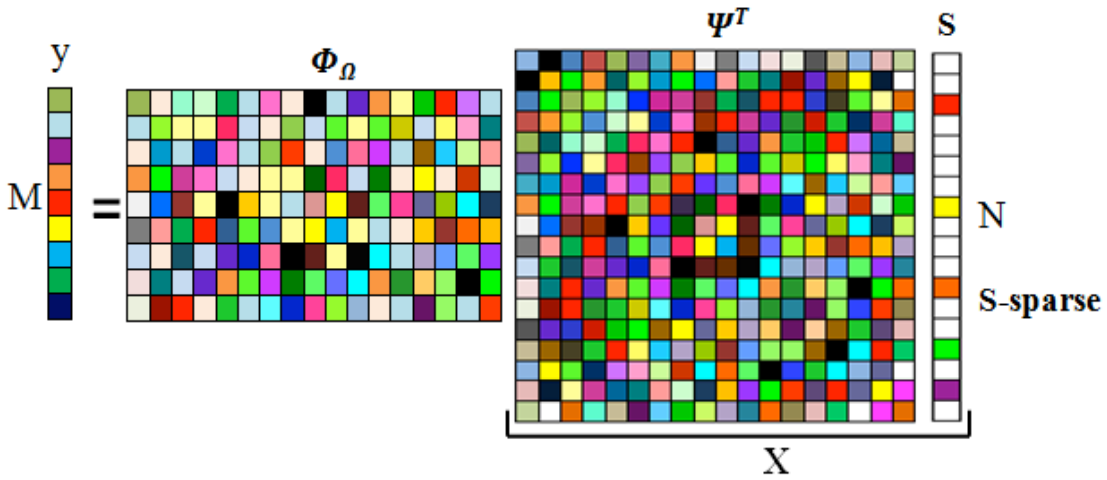
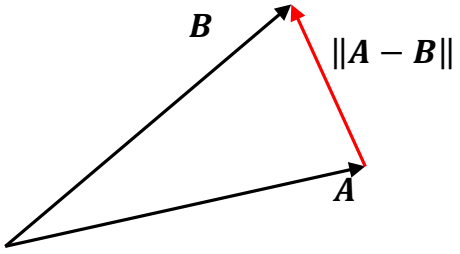


Figure 10. Acquisition matrices (adopted from [31]), add definitions for the matrices

2.3.5.2. Norms

The norms usually are a way to measure the distance between two vectors. Depending on the applications, the norm can be used to measure the size of the error (as the least mean square method).

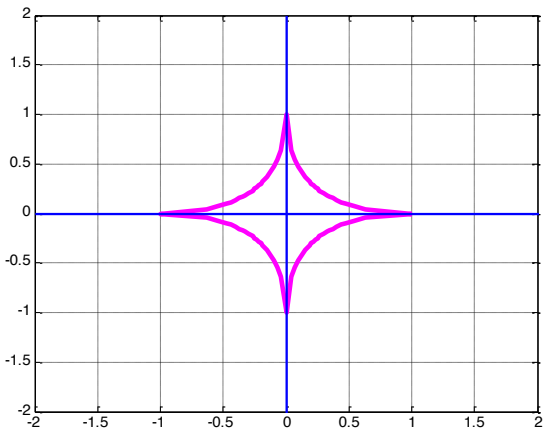


The general formula of the norm is:

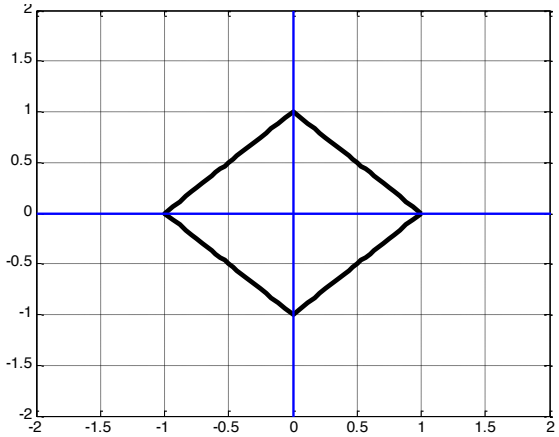
$$\|x\|_p = \left(\sum_{i=1}^{M} (x_i)^p\right)^{1/p} \tag{11}$$

Three conditions should be satisfied for the norm:

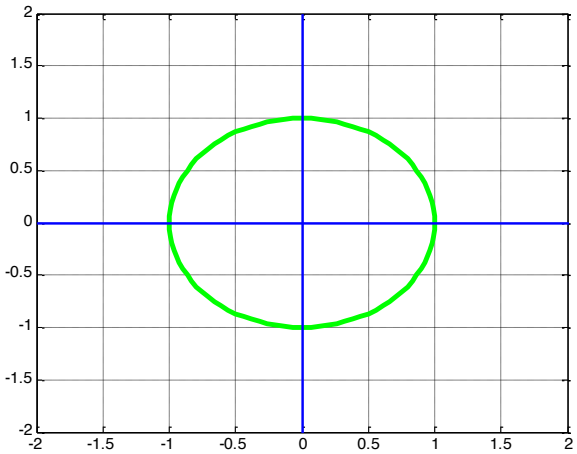
1. $\|x\| \geq 0$
2. $\|x + y\| \leq \|x\| + \|y\|$
3. $\|\alpha x\| = |\alpha|\|x\|$



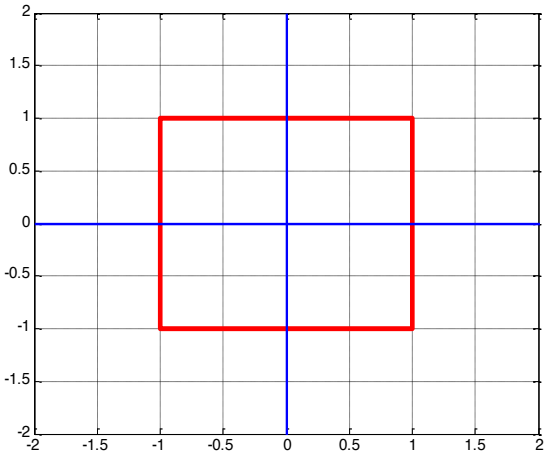
$p = \frac{1}{2}$



$p = 1$



$p = 2$



$p = \infty$

Figure 11. Different norm forms ($p = \frac{1}{2}, 1, 2 \ \ and \ \infty$)

2.3.5.3. Sparsity and the L1 Norm

The sparsity can be obtained by solving the following optimization problem:

$$\min\|s\|_0 \quad subject\ \ to \qquad \phi_\Omega x = y \tag{12}$$

However, this problem is NP-hard. But it has been observed that the sparse signals have a small $\boldsymbol{L1}$ norm relative to their energy.

$$min\|x\|_1 \ \ subject\ to\ \ y = \phi_\Omega x \tag{13}$$

When $L0 = L1$ holds, that is key to CS.

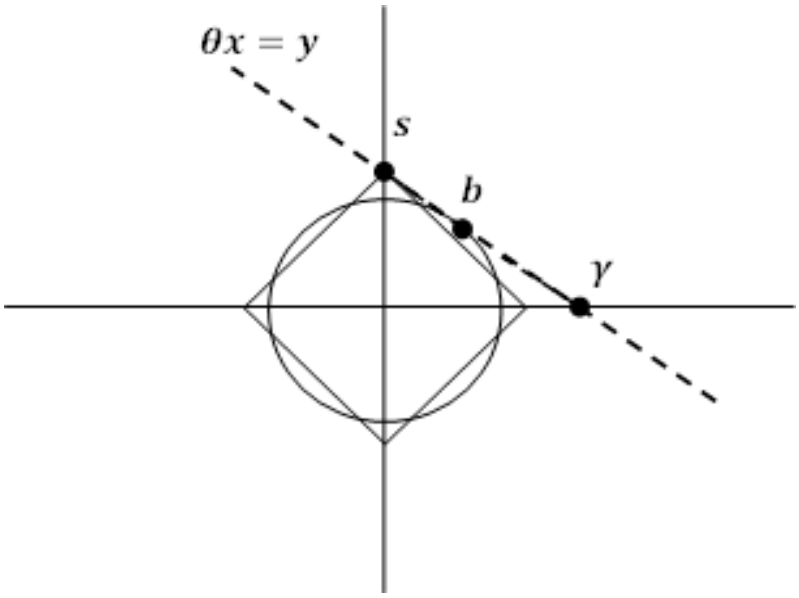


Figure 12. Comparison between L1 and L2 norm for finding the sparsest solution

2.3.6. The Recovery Algorithm

The idea of compressing sensing is in terms of its recovery assumption if we do not follow sampling theorem requirements. Therefore, CS is based on the assumption of a signal undersampling, but reconstruction is secured using methods of convex optimization [32].

$$\min\|s\|_1 \quad subject\ to \qquad \theta_\Omega s = y \tag{14}$$

Now, many conditions should be understood to ensure that the sparsity solution for reconstruction is found. The sensing matrix θ_Ω is the key of the recovery procedure.

2.3.6.1. Incoherence

The coherence is a measurement of the correlation between the measurement matrix ϕ_Ω and the transform matrix ψ where the signal can have a sparse representation

$$\mathbb{\Gamma}(\phi, \psi) = \sqrt{N}\ \max_{i,j} \frac{|< \phi_i, \psi_j >|}{\|\phi_i\|\|, \psi_j\|} \tag{15}$$

In general, $\mathbb{\Gamma}(\phi, \psi)$ measure the minimum angle between ϕ, ψ. The high incoherence means that these two vectors are far apart [33].

$$1 \leq \mathbb{\Gamma}(\phi, \psi) \leq \sqrt{N} \tag{16}$$

If we apply incoherence definitions on Fourier transform (where $\psi(t)_k = \frac{1}{\sqrt{N}} e^{\frac{-j2\pi kt}{N}}$ is composed of the DFT coefficients and $\phi(t)_k = \delta(t-k)$ is the canonical or spike basis) the coherent between ϕ and ψ yield $\mathbb{\Gamma}(\phi, \psi) = 1$. Therefore, we have a maximal incoherent (minimum coherent). The small coherence is required between two domains to get better performance. The advantage of incoherence is that for all random vector's combination measurements, there is a relation between the sparse and measurement vector; each of the measurement vectors (rows of ϕ) must be 'spread out' in the ψ domain. In general, the random matrices are largely incoherent with any fixed basis ψ. Therefore, we select an orthonormal basis ϕ uniformly at random, which can be done by orthonormalizing N vectors sampled independently and uniformly on the unit sphere. By extension, random waveforms ($\phi_k(t)$) with independent identically distributed (i.i.d.) entries, e.g. Gaussian or ±1 binary entries, will also exhibit a very low coherence with any fixed representation ψ [34]. Incoherence is based on θ (mutual coherence)

$$\mathbb{\Gamma}(\theta) = \sqrt{N} \max_{i,j} |\theta_{ij}| \tag{17}$$

where each column is a L2-normalized $\|\theta\|_2 = 1$

$$\theta = \begin{bmatrix} \phi_1^T\psi_1^* \dots\dots\dots\dots\dots\dots\dots \phi_1^T\psi_N^* \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \phi_N^T\psi_1^* \dots\dots\dots\dots\dots\dots \phi_N^T\psi_N^* \end{bmatrix}$$

For a compressive sensing, we only consider matrix $\theta \in \mathbb{R}^{MxN}$ with $M \ll N$. The analysis of recovery algorithms usually involves a quantity that measures the suitability of the measurement or sensing matrix. The coherence is a very simple measure of quality. In general, the smaller the coherence, the better the recovery algorithms perform [35].

2.3.6.2. The Restricted Isometry Property (RIP)

The RIP establishes a condition for θ_Ω matrix to guarantee recovery of a sparse vector. For each integer s=1,2,….N, the s-restricted isometry constant $\delta_s \in (0,1)$ as a smallest number satisfying that

$$(1-\delta_s)\|s\|_2^2 \leq \|\theta_\Omega s\|_2^2 \leq (1+\delta_s)\|s\|_2^2 \quad with \ \|x\|_0 \leq k \tag{18}$$

RIP restricts the energy of the signal to the set of Ω and is proportional to the size of Ω. If θ_Ω satisfies RIP in order of $2k$, then θ_Ω approximately preserves the distance between any pair of k- sparse vectors [36] [37].

Theorem: [23] Let θ be an NxN orthogonal matrix and $\mathbb{\Gamma}(\theta)$ satisfy the last formula in equation (18). Fix a subset $\Bbbk$ of a signal domain and choose a subset Ω from the measurement domain size M. The number of measurement can be written as:

$$M \geq C_0 . |\Bbbk| . \mathbb{\Gamma}(\theta)^2 . \log(N) \tag{19}$$

C_0 is fixed numerical constant. Suppose that there are two s-sparse signals, $s_1 and\ s_2$, such that $\theta_\Omega s_1 = \theta_\Omega s_2 = y$. Then let $r = s_1 - s_2 \in \mathcal{R}^{2s}$ be sparse:

$$\theta_\Omega r = \theta_\Omega (s_1 - s_2) = \theta_\Omega s_1 - \theta_\Omega s_2 = 0 \tag{20}$$

Use the RIP condition to get that:

$$(1 - \delta_{2s})\|r\|^2 \leq \|\theta_{\Omega\Bbbk}\|^2 = 0 \tag{21}$$

Since$(1 - \delta_{2s}) > 0$, therefore $\|r\|^2 = 0$ must be satisfied; that means s_1, s_2 are not distinct, but they are one sparse pair.

As a result, the incoherence and RIP are the most the important conditions for a CS algorithm to guarantee recovery of the original signal represented only by sparse coefficients. The transform matrix columns are linear independent. Due to the matrix being fat, the sparse will add an advantage; the only thing one needs is for the θ_Ω

columns to behave like linear independents. The RIP provides that the sparse linear combination involve no more than the **s** vector. Then every signal ***s*** **is** supported on ***ƙ*** with signs matching z; the recovery from $y = \theta_\Omega s$ occurs by solving

$$\tilde{s} = \min_{s^*} \|s^*\|_1 \quad subject\ to\ \ y = \theta_\Omega s^* \tag{22}$$

2.3.7. Compressive Sensing Procedures

Recording only the necessary data to reconstruct the signal is the main idea behind CS. The CS proceeds by recording $y = \phi s$, where ϕ is the measurement matrix mxn . Choosing $M \ll N$ immediately gives a compressed measurement vector y of length m instead of n. In compressing sensing technique, the most effort is spent to find an optimizing solution for the linear system equation$y = \phi x$. The encoding phase is non-adaptive and does not need analysis in order to find the final encoding [20]. Figure 13 summarizes the main CS procedure steps

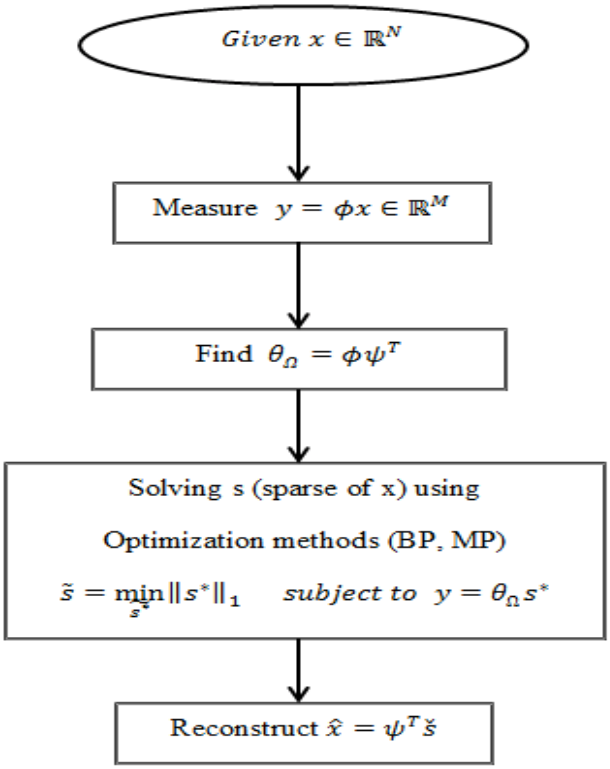
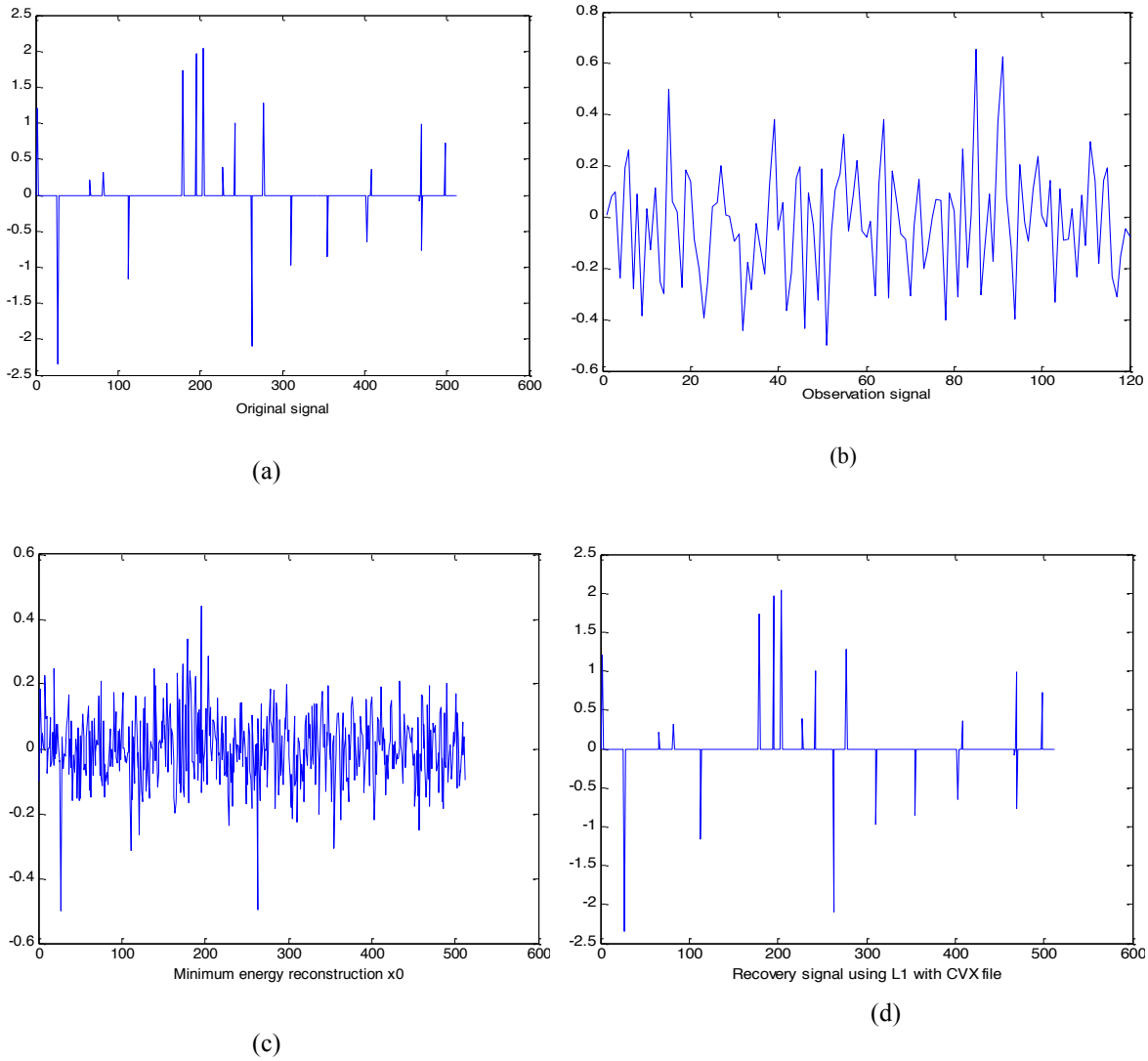


Figure 13. CS procedure. starting with the x as the original signal and finding s which is a sparse representation of x ($\boldsymbol{s = \psi x}$)

Figure 14 shows the sample numerical example of how CS can be applied to reconstruct the sparse of the signal. The input signal is a vector of the 512 signal level; the signal is already sparse, but the number of sparse representations and their positions are randomly selected. In this example, the sensing matrix θ_Ω can be chosen as $\boldsymbol{MxN}$ random distribution matrix (120x512). The BP method has been applied to solve the

convex optimization as a function with L1. The solution of convex formula is not easy to program; therefore, this study used the license as a new MATLAB tool box [38].

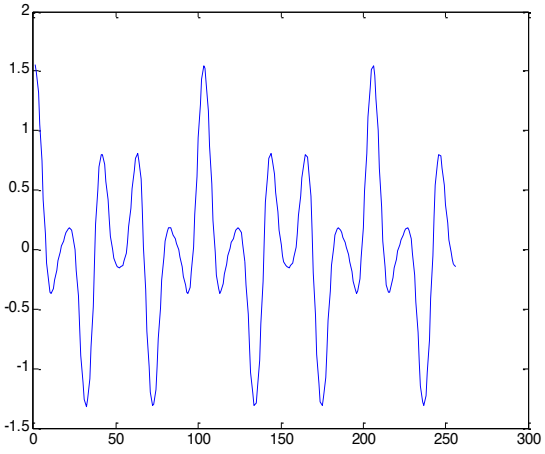


(a) (b)

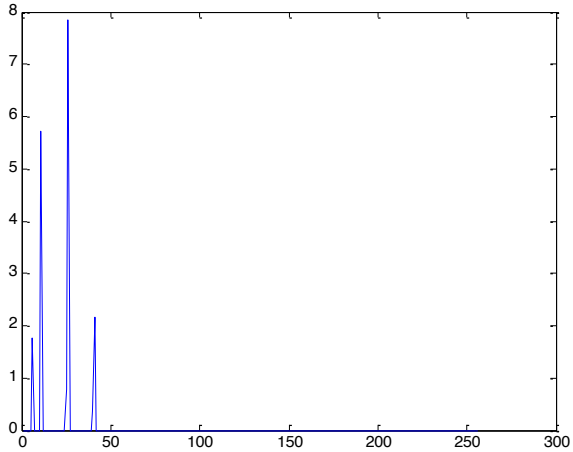
(c) (d)

Figure 14. (a) The input single which is constant only as a sparse coefficient, (b) the input signal after multiplied with the transformation matrix, (c) the minimum energy x0 and (D) the recovery signal

To illustrate the compressed sensing methodology as flow diagram shown before, this study chose a signal x , which has a sparse representation. The sinusoidal signal with length **256** is shown in the figure below, and its sparse representation is in the DCT domain. First, this study creates a restriction or measurement matrix: $\phi_\Omega \, is \;\; MxN$, where $N = 256 \;\; and \; M$ is the number of measurements that should be chosen to satisfy that $M \geq C_0 |\mathrm{k}| \log(N)$, and it will be represented in rows. We can observe $y = \phi_\Omega x$; the measurements' locations are plotted in Figure c . Meanwhile, the $\psi \, (NxN)$ DCT coefficients matrix is generated, then the transform matrix $\theta_\Omega = \phi_\Omega \, \psi$ is calculated. The recovery of the original signal is done using the Basis- Pursuit method, shown in Figure d, which gives an exact reconstruction.



(a)



(b)

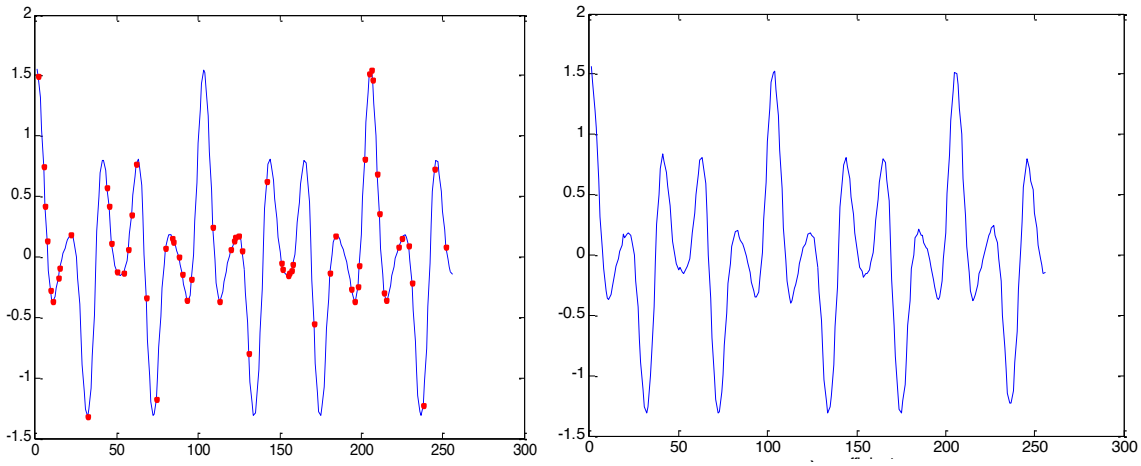


Figure 15. Reconstruction of the signal from a few measurement elements: (a) is the original signal vector with 256 elements, (b) is DCT representation of the original signal. In (c) measurements elements are obtained using the 50x256 matrix (80.468%) missing elements, and (d) displays the BP recovery signal for measurements signals

2.3.7.1. Compressed Sensing for a Non-Exact Sparse Signal

Most signals are not exactly sparse; they can be approximately sparse in a specific domain such as in the power law when the coefficients are rapidly decaying. Now, this study applies the CS technique on those kinds of signals. If the $\boldsymbol{x}$ signal has an approximate sparse representation that is derived by making the coefficients, which are less than a certain level, equal to zeros. If s_b is the best S-sparse approximation, i.e (N-S), the number of coefficients are to be zeros.

Theorem: assume that s is approximately sparse, and s_b can be defined as above [23]. Then if $\delta_{2s} < \sqrt{2} - 1$, the solution $\tilde{s}$ is

$$\tilde{s} = \min_{s^*} \|s_b\|_1 \;\; subject\ to \;\; \theta_\Omega s_b = y \tag{23}$$

Which follows that

$$\|\tilde{s} - s\|_1 \leq C. \|\tilde{s} - s_b\|_1$$

and

$$\|\tilde{s} - s\|_2 \leq C_0 s^{-1/2}. \|\tilde{s} - s_b\|_1$$

According to this theorem, the CS provides the exact recovery scheme even for the signals which are approximately sparse. Furthermore, CS is a universal algorithm that can be used for all sparse or compressible signals without any chances of failure.

2.3.7.2. Compressed Sensing for Corrupted Signals

Another very important and realistic scenario is to consider that the acquired data is contaminated with noise [39] [40]. In this case, the measurements will be affected by the noise. If the noise is n with a power$\|n\|_2 \leq e$, then one can write that:

$$y = \phi x + n \quad with \ \|n\|_2 \leq e \tag{24}$$

If one has a noisy sparse signal s_n with support of r (support of a function is the set of points where the function is not zero value $(s_n) = \|s_n\|_0 = \#\ of\ non\ zero\ elements$) then one uses the least-mean square method to find $\tilde{s}$.

$$\tilde{s} = \begin{cases} (\theta_{\Omega r}^T \theta_{\Omega r})^{-1} \theta_{\Omega r}^T y & on\ r \\ 0 & elsewhere \end{cases} \quad (25)$$

If the signal is approximated well using enough numbers of elements, then

$$\|\tilde{s} - s\|_2 \cong \|\theta_{\Omega r} n\|_2 \cong e \quad (26)$$

The condition to guarantee $\tilde{s}$ is that

$$\|\tilde{s} - s\|_2 \leq C_1 e \quad (27)$$

Theorem [20]: Assume $y = \theta_\Omega s + n, \quad \|n\|_2 \leq e$; then if $\delta_{2s} \leq \sqrt{2} - 1$, the solution $\tilde{s}$ is

$$\tilde{s} = \min_s \|s\|_1 \quad subject\ to\ \|\theta_\Omega s - y\|_2 \leq e \quad (28)$$

follow that

$$\|\tilde{s} - s\|_2 \leq C_0 s^{-1} \|\tilde{s} - s\|_1 + C_1 e \quad (29)$$

The reconstruction error is the superposition of two factors, the additive noise error and the sparsity approximation error.

2.3.8 Multiple Measurement Vector (MMV)

The MMV is generalizing the notation from that the signal be a single vector and sparse to an ensemble of signals to be jointly sparse. Unlike the single vector, in the MMV

model, a given ensemble's measurement vectors $Y \in \mathbb{R}^{MxL}$ and a dictionary $\boldsymbol{\Phi}$ where M is the number of measurements that are taken from each signal vector in X and L is the number of ensemble signals. The recovered signal as joint sparse is the matrix $X \in \mathbb{R}^{NxL}$.

$$\boldsymbol{Y} = \boldsymbol{\Phi X} \tag{30}$$

When $\mathrm{L} = 1$, we have the single vector case. The $\mathbf{X}$, $\mathbf{Y}$ can be represented as $\mathbf{X} = \left[\mathrm{x}^{(1)}, \mathrm{x}^{(2)}, \ldots\ldots, \mathrm{x}^{(\mathrm{L})}\right]$, $\mathrm{Y} = \left[\mathrm{y}^{(1)}, \mathrm{y}^{(2)}, \ldots\ldots, \mathrm{y}^{(\mathrm{L})}\right]$. Where $\mathrm{x}^{(\ell)}, \mathrm{y}^{(\ell)}$ for $1 \leq \ell \leq \mathrm{L}$ are columns vectors and $\Phi \in \mathbb{R}^{MxN}$. Figure (16) shows applying MMV for multiple vectors $x^l, l = 1, 2, \ldots\ldots, L$.

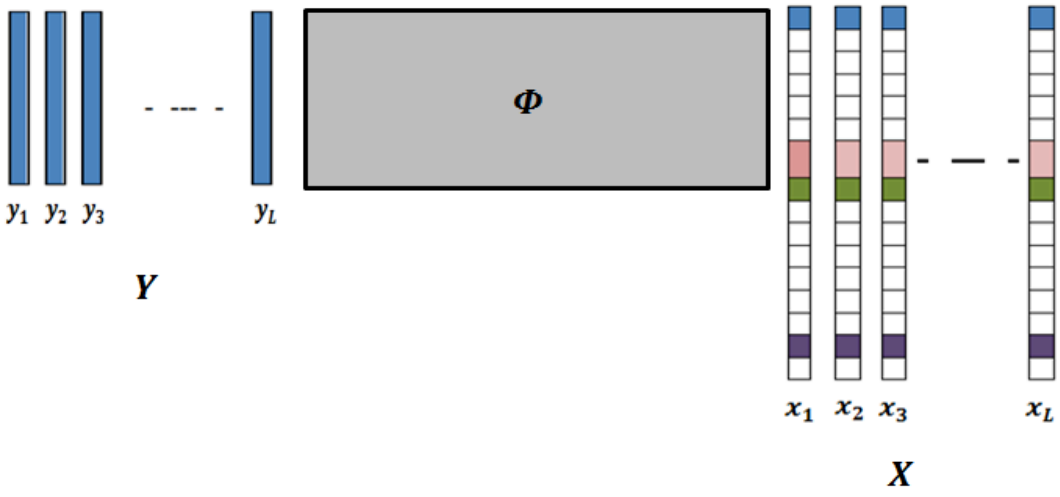


Figure 16. General structure for MMV algorithm for L signals

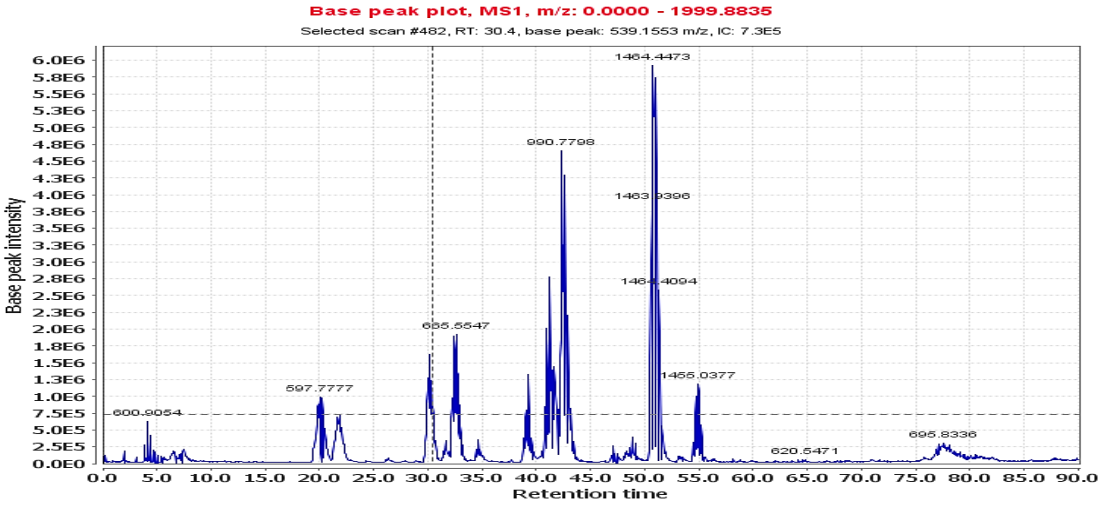
There are two conditions for MMV,

i. Each column of $\mathbf{X}$ which represent a single signal should be sparse. This requirement is the same in the single measurement vector case.
ii. The matrix $\mathbf{X}$ has a common sparse with a few number of rows that contain nonzero entries.

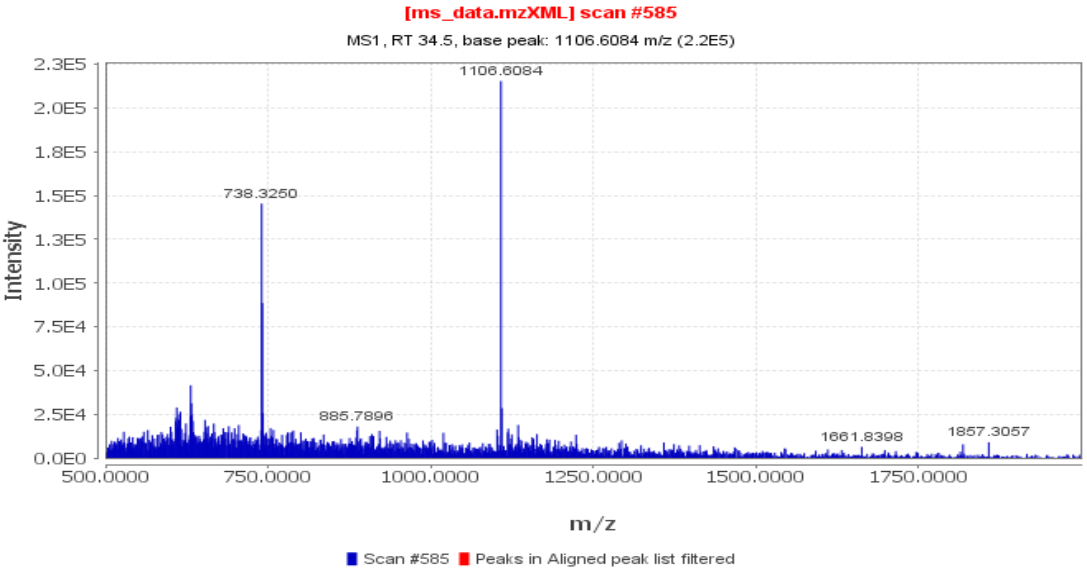
To recover the jointly sparse matrix from the system $\boldsymbol{Y} = \boldsymbol{\Phi}\mathbf{X}$, the problem can be formulated as [41].

$$\widehat{\mathbf{X}} = \min\|\mathbf{X}\|_{1,2} \quad \text{subject to} \quad \mathbf{Y} = \boldsymbol{\Phi}\mathbf{X} \tag{31}$$

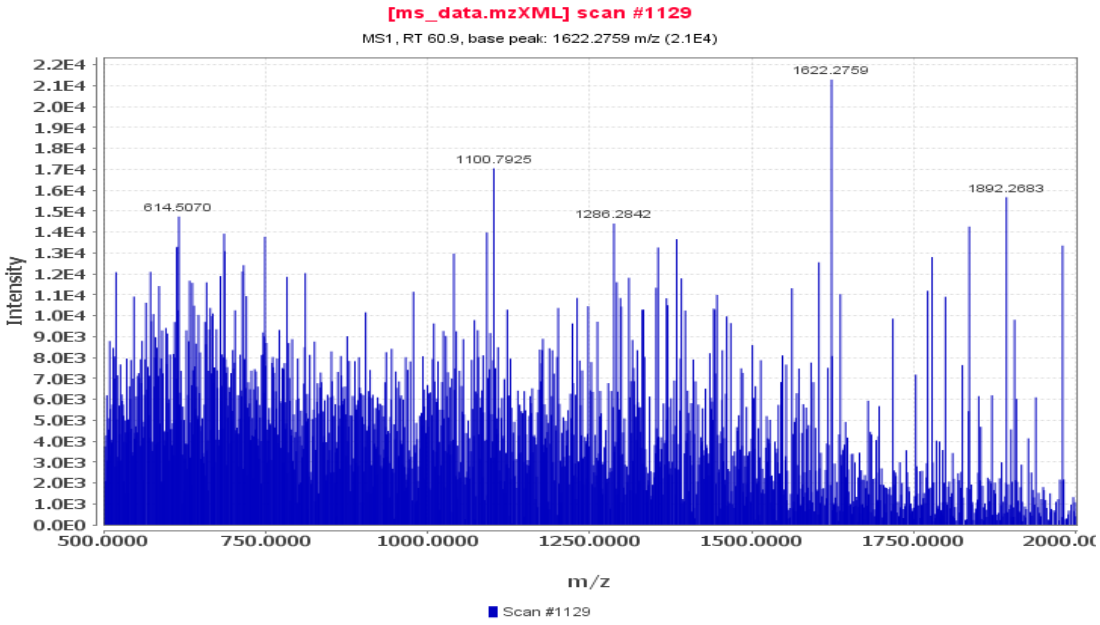
In this section, we presented reconstruction results using MMV for an example LC/MS dataset. The data set was acquired using an LCQ Advantage mass spectrometer for a protein that was treated with and without a lipid that covalently modifies proteins at specific cysteine binding sites. The data consists of 1371 runs (scans) with a total number of mass/charge ratio and intensity pair $1473x10^6$ for the first level. The data range is m/z (500 - 2000). Figure 17) below shows the TIC of the entire row data.


Base peak plot, MS1, m/z: 0.0000 - 1999.8835
Selected scan #482, RT: 30.4, base peak: 539.1553 m/z, IC: 7.3E5
Base peak intensity
Retention time
1464.4473
990.7798
1463.9396
1464.4094
655.5547
1455.0377
597.7777
600.9054
620.5471
695.8336

(a)


[ms_data.mzXML] scan #585
MS1, RT 34.5, base peak: 1106.6084 m/z (2.2E5)
Intensity
m/z
1106.6084
738.3250
885.7896
1661.8398
1857.3057
Scan #585
Peaks in Aligned peak list filtered

(b)



(c)

Figure 17. LC/MS image view, TIC for the entire data has been shown in (a) whereas two different samples at a specific time has been selected in (b), (c) using MZmine 2.11

The MMV sparse recovery problem will be applied [42] on MS data. We need to reform the MS data as a joint sparse matrix; where the retention time represent the ensemble data as columns of a matrix and M/Z will be the symbol for the rows. Each element in the matrix has a value which is represented by intensity as shown in Figure (18). However, the number of M/Z s vary from one sample time to another. In other words, each $x_j, 1 \leq j \leq T$ has its own dimension and T is the total sample scans. Therefore, to reform the joint sparse matrix, we assume that the number of rows is equal

to the maximum scan dimension n_{max} and set the other rows that are less than this number as equal to zeros.

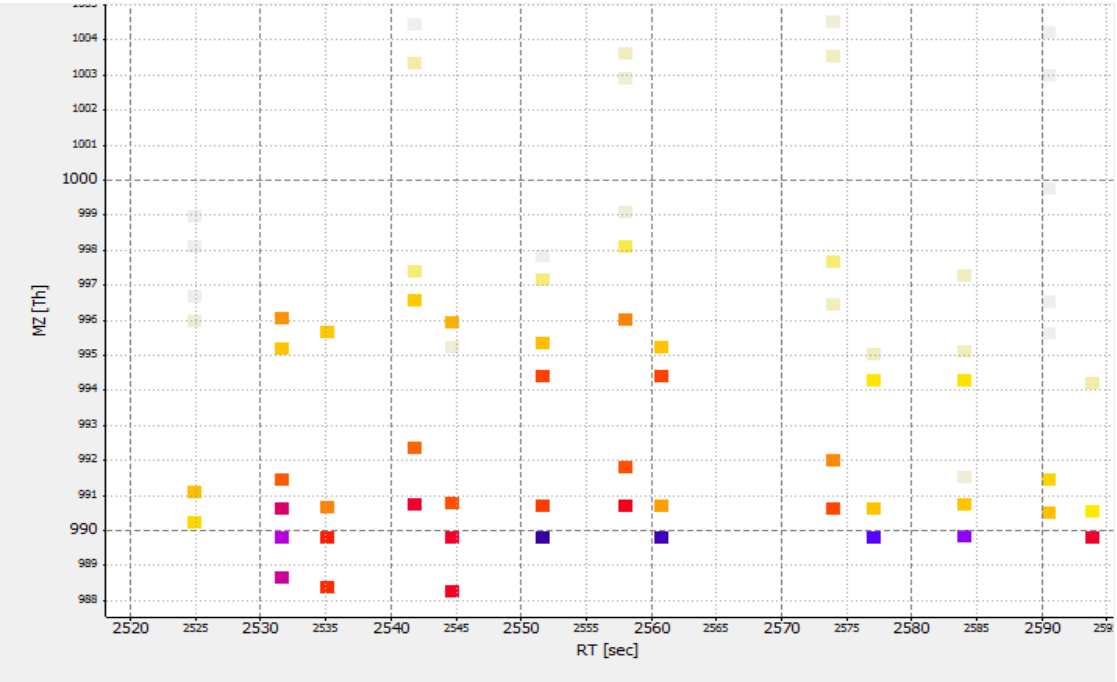


Figure 18. The 2D matrix Positions data points in a small range of m/z and retention time of the whole profile raw data. The colored squares represent the intensity value

To satisfy the sparse condition, we assume that there exists a known sparse basis $\psi \in \mathbb{R}^{n_{max}}$ in which x_j be sparse. Due to varieties of x_j, the sparse signal $s_j = \psi x_j$ having different nonzero K_j. We select the maximum value of nonzero K_{max}. Therefore, the common measurement can be found, thus

$$y_j = \psi \phi_j x_j \quad , \qquad j = 1,2, \ldots\ldots, T \tag{32}$$

Denote $\phi_j \in \mathbb{R}^{m_j x n_{max}}$ is the measurement matrix of the j signal. The model including additive noise can be written as:

$$\boldsymbol{Y} = \boldsymbol{\Phi\Psi X} + \boldsymbol{Z} \tag{33}$$

The dimensionality of $\boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{R}^{\boldsymbol{mxT}}$ and $\mathbf{Z}$ is a noise with $\|Z\|_2^2 \leq \epsilon$.

According to jointly sparse conditions, our formulated matrix has sparse columns and few rows are nonzero, however, the common sparse of all rows is not satisfied. To achieve that, the data matrix divides into segments with each segment is being treated as an MMV model as shown in figure(19). In order to make cluster across a different samples, we need to align the peaks which might be shifted due to instrument.

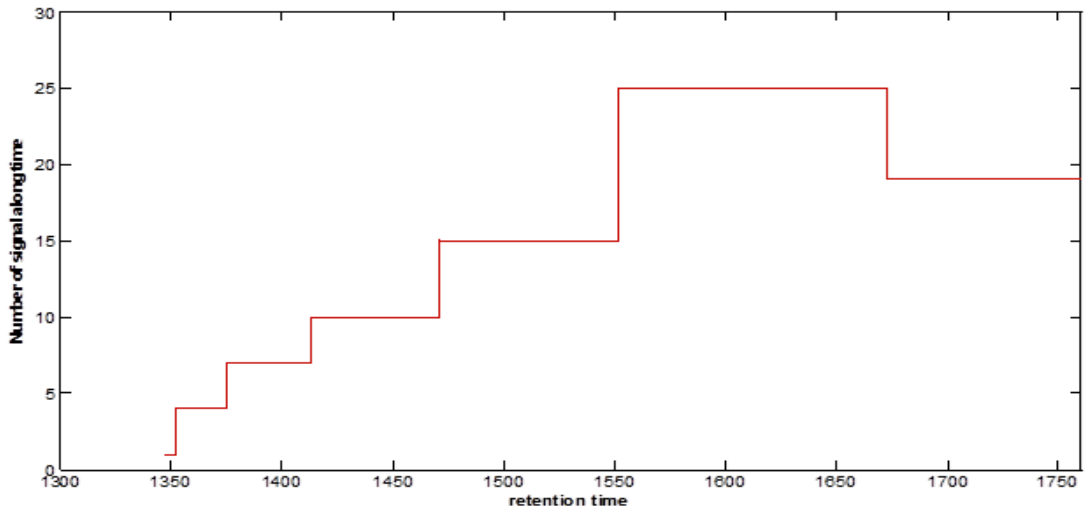


Figure 19. The data segmentation where there are number of peaks at the same m/z of each retention time range

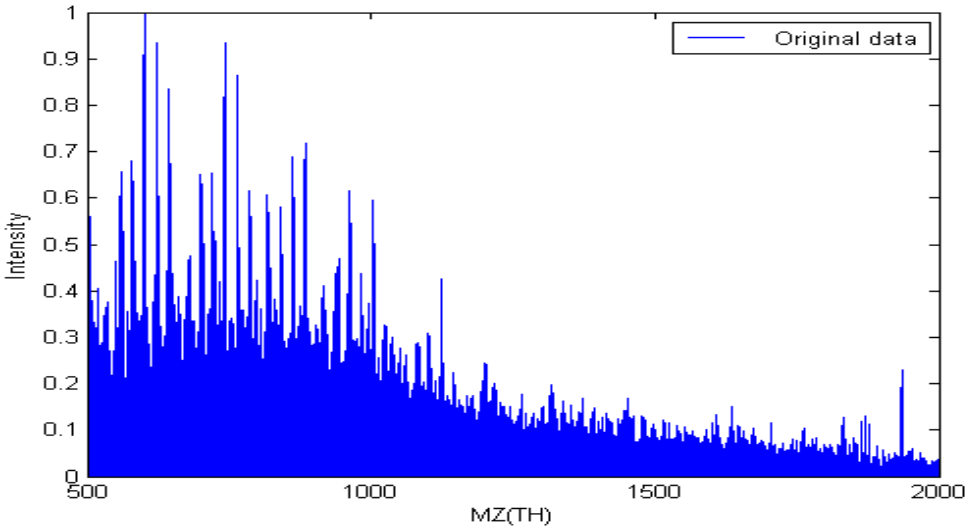
Now, with measurement matrix where the dimension of the original data has been severely reduced; the next procedure is recovery the original data **X** from **Y.** There are several efficient algorithms have been proposed to improve BP for MMV in equation (31) mentioned in [43]. We select the MMV version of BPDN as one methods of a spectral projected gradient (SPGL1) [44].

$$minimize \quad \|\boldsymbol{S}\|_{1,2} \qquad subject\ to \qquad \|Y - \Psi^T \Phi X\|_F \leq \epsilon \tag{34}$$

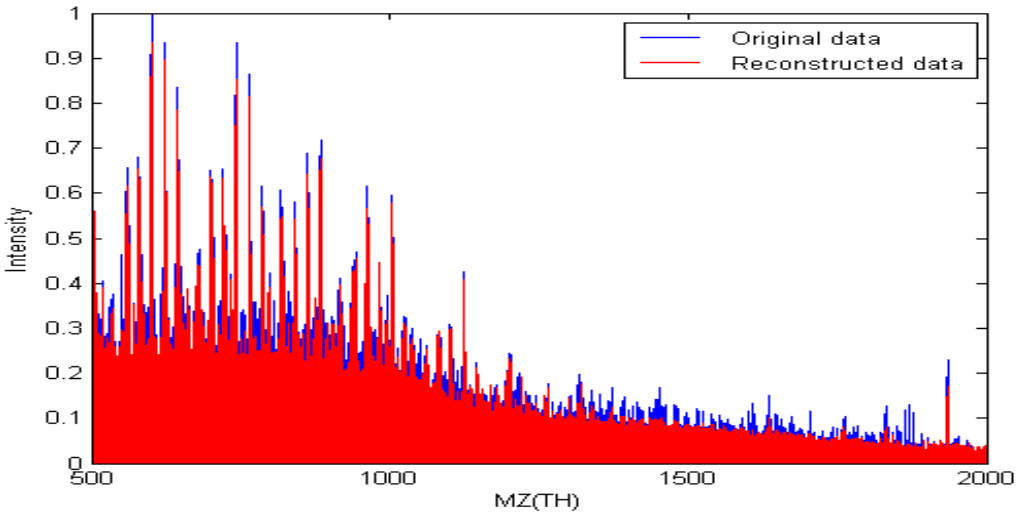
Where the mixed (1,2) norm $\|S\|_{1,2}$ is defined by the sum of two norms of the rows of S and $\|\boldsymbol{S}\|_F$ is The Frobenius norm and **S** is a recovery matrix of sparse signal which has been obtained in a specific domain $\boldsymbol{\Psi}$ such as discrete cosine transform (DCT), wavelet. We select the spares coefficients based on the power ratio of signal. For evaluation purpose, the normalized mean squared error (NMSE) to evaluate the individual columns from the reconstructed matrix

$$\frac{\left\|\widehat{x_j} - x_j\right\|_2^2}{\left\|x_j\right\|_2^2} \tag{35}$$

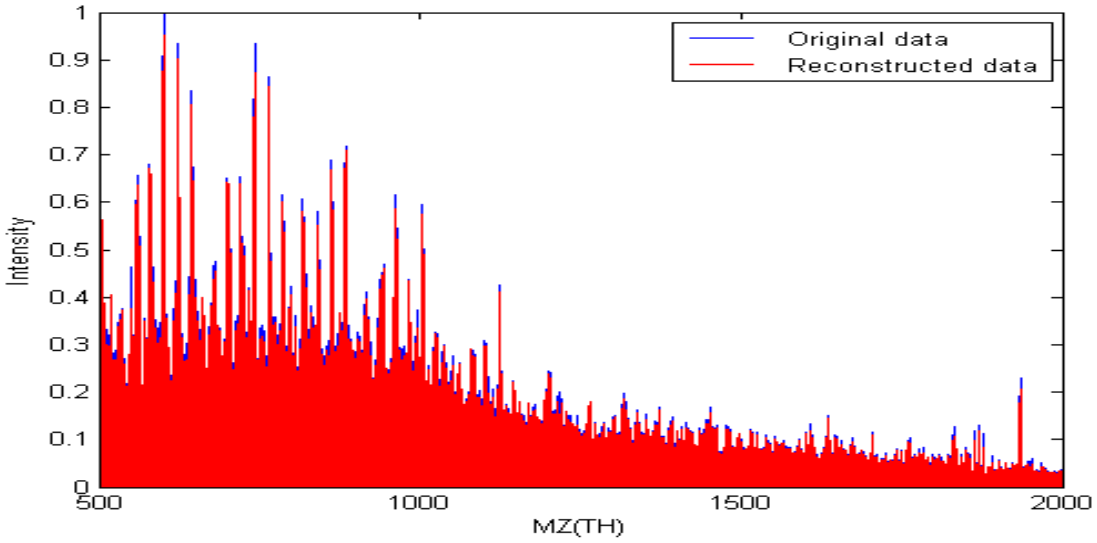
Figure(20) demonstrates the reconstruction data with a different compressive sensing ratio M/N. Figure 20(a) shows the projection of the source data and figure 20(b),(c) and (d) show the reconstruction data with 6%, 12% and 35%. The data recovery is still possible with a very low (NMSE) even with a compression ratio as deep as 6%.



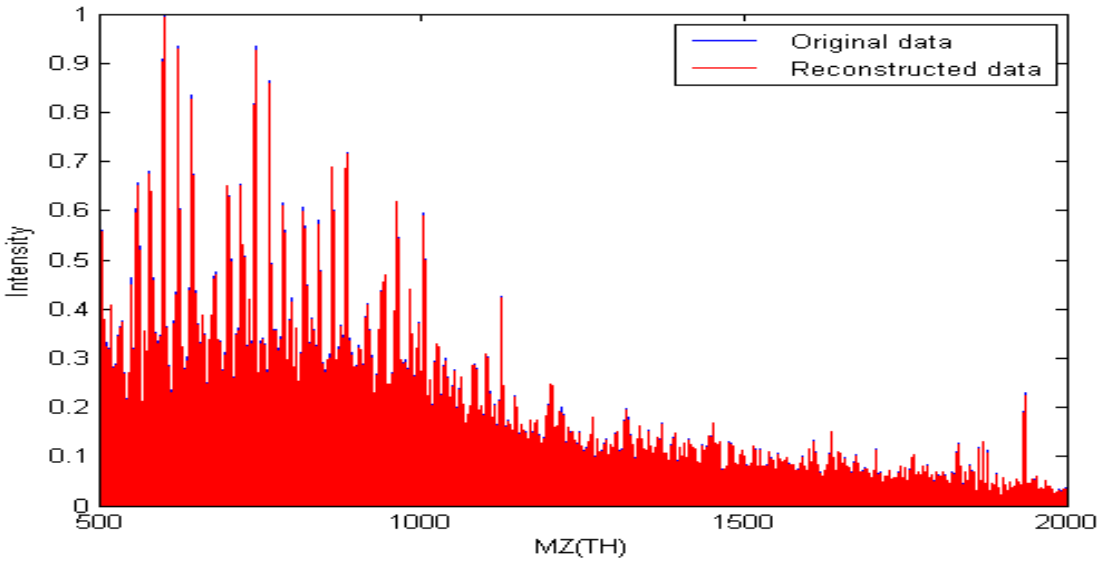
(a) The original sample



(b) $MSE = 7.2297x10^{-5}$



(c) $NMSE = 3.387x10^{-6}$



(d) $NMSE = 6.592x10^{-7}$

Figure 20. Reconstruction data using MMV. Where (a)shows the original data whereas (b),(c) and (d) show the comparison or original and recovery data with a 6%, 12% and 35% portion of data and a very low NMSE ratio

CHAPTER III

PERTINENT LITERATURE

Mass spectrometry data is one of the most valuable types of biological and clinical data. An example of biological data is to define protein peptides, and an example of clinical data is to recognize the health and disease sample. Therefore, the classification of MS data is very important. The challenge in using MS data usually is its high dimensionality with a classification process that is computationally intensive. The dimensionality reduction must be accomplished first, and classification is sought as a second step. In the recent studies, there is not enough attention given to MS based on CS. Therefore, this survey first presents the major classification methods of MS data and some methods based on CS. In the second part, the recovery methods of MS data from sensing data will be considered. Generally, most classification strategies of MS are based on database analysis, and in some cases, the test sample does not exist in the database. Consequently, the recovery of MS data can give an advantage for updating the database; since the dimension of features has been reduced, more samples can be added to the database. In addition, the predictor will materialize faster than before.

3.1 Mass Spectrometry Data and Classification

The MS classification starts by providing a given set of MS training spectra with correct classifications. The output of the classifier (predictor) depends on the classifier's ability to discriminate (classify) a new test sample into one of the classes. Many

algorithms have been proposed to classify MS data. In general, the main aspects for those schemes can be summarized in the following groups:

3.1.1 A Complete MS Data Classification

The classification performs according to the whole MS data where the entire peak intensities are considered. In [45], the linear discriminator analysis (LDA) had been explored for MS classification. The classification decision has been achieved according to the linear combination of the features, by dividing the M/Z range into bins then selecting the midpoint to represent each bin and then using F-statistics to select the important peaks. The selected peaks produce a low dimension set. The LDA predictor is applied to this dataset. The sensitivity (sen.) and specificity (spec.) and positive protect value (PPV) show an excellent performance (>97%) when applied to the SELDI spectra of ovarian cancer.

In reference to [46], the Principle Component Analysis (PCA) has been used to classify the gas chromatograph MS (GC-MS) of unleaded gasoline samples with an octane rating of 89. Using a piecewise retention time alignment algorithm connected with an analysis of variance (ANOVA) feature selection as prior preprocessing steps before the classification, the PCA performance has been improved compared to applying PCA on GC-MS data directly. The proposed algorithm can be summarized as follows.

- .Each individual ion of the (m/z) chromatogram in the GC-MS feature is aligned with all other chromatograms of that m/z.

- The feature selection ANOVA algorithm is applied to each individual m/z to find the important features based on the Fisher ratio threshold.
- Those selected features are submitted to PCA for classification.

In reference to [47], the discrimination decision is based on a tree-classification. Serum samples were taken from malignant prostate cancer cells, benign prostate hyperplasia cells and healthy (Control) samples. The peaks selection, peaks alignment and feature selection algorithm (AUC)[‡] were applied before the classification process. Once the preprocessing procedures were completed, the tree-classification took place. This process is applied to distinguish cancer from non-cancer samples as well as identifying all genetic subtypes of the cancer and their biological activity. The tree algorithm starts with splitting the training set into two pins according to the presence or absence of intensity levels from one peak. The splitting process continues as long as terminal nodes have no gain. The classification decision is based on the majority of samples in the nodes. The performance has a high throughput proteomic classification sensitivity of 83%, a specificity of 97%, and a positive predictive value of 96%.

Reference [48] considers the peak detection as one of the important preprocessing techniques in proteomic data analysis. In contrast to many algorithms, the shape of the peak can provide additional information. The noise in MS data can generate high spikes which will lead to peak miss-selection. The authors propose to transform the

[‡] Area Under Curve.

MS spectrum into continuous wavelet (CWT) space. This transformation provides an advantage for recognizing MS peaks from noisy peaks. In addition, using this projection, the baseline removal step does not need to be applied. The summary of steps in this algorithm are as follows:

- Perform the CWT over the entire the MS data spectrum.
- According to 2D produced CWT coefficients, identify the ridge lines based on the local maxima point from the CWT coefficients.
- Based on the ridge lines, the peaks can be selected corresponding to not only the amplitude, but also the amplitude with width peaks as well.
- Since there some peaks with low amplitude and width, the adapting threshold values are applied.

In a further study [49], the Q5 algorithm was proposed for the probabilistic classification of a serum sample using mass spectrometry. The Q5 algorithm, built upon a dimension, was reduced using PCA by projecting the spectra-space into a lower dimension, where the cross class variance is maximized. Then, LDA is applied to classify the projecting data. This algorithm has been applied to ovarian and prostate cancer data sets. The method achieved a good performance for sensitivity, specificity, and positive predictive values of $\geq 97\%$.

The study in [5] used the three stages proposed for prostate cancer classification from the MS data. The filters using t-testing, wavelet analysis and statistical moment are

used as a dimensionality reduction for the MS spectrum. For the classification purpose, the kernel partial least square algorithm was implemented. As a result of this comparison, the wavelet analysis with kernel partial least square gives a better performance for classifying prostate cancer. The maximum values of accuracy are (95.8%), specificity (96.8%) and sensitivity (96.8%).

3.1.2. Manually Preprocessing MS Classification

Partial features are candidates for classification where some peaks or ranges of spectra, such as alignment or filter, are excluded during the preprocessing procedures. In addition, some peaks can be eliminated regarding the amplitude of intensity based on prior knowledge.

The studies cited in [50] [51] consider support vector machines (more details in [52]) as an effective method for classification with low dimension data. The study cited in [51] proposes that the Recursive Feature Elimination (SVM-RFE) algorithm should select important genes/biomarkers for the classification of noisy data. The selecting features were done according to the feature's weight as ranking criterion in the SVM classifier. Whereas, the authors in [50] propose the recursive support vector machine (R-SVM) algorithm as an improvement to (SVM-RFE) by changing the features selection process. This process is done recursively so that the privileged aspects for the classification model can be obtained in a recursive manner, at different levels of gene selection.

In [53], two types of filters were used for feature selection gain ratio (GFS) and correlation filter feature selection (CFS). The filter approach starts with selecting features once these features can be applied to the classifier. In GFS, the information gain was used to select attributes having a large number of values. However, CFS determines the subset attributes by considering the individual productivity ability of each feature along with the degree of redundancy between them. Correlation coefficients were used to estimate the correlation between the subset attributes and class and inter-correlation between features. The C4.5 [54] [55] has been used as a predictor decision for the gain filter method whereas the Genetic algorithm is used as a way of research for a correlation filter. The experiment results show that the classification accuracy with a CFS filter is better than applying a GFS filter.

Sparse proteomics analysis (SPA) is another way to complete feature selection based on the compressive sensing idea [4]. The sparse features are a small number of features that can be used to accurately predict unknown proteomic data. The algorithm used in this study is able to define significant features' positions. These features are the maximum of some peaks selected to the close peaks. The dimensionality reduction is done by projecting the data to select the features' positions.

3.1.3 Sparse Representation for MS Classification

In recent years, Wright J. &Yang A. [56] proposed a new theory that classifying among multiple linear regressions can be achieved from sparse signal representation.

The test sample can be represented from linear representation of all training samples as a vector. The coefficients vector entries are zeros except those associated with a particular class or category. For a training set with n samples labeled into k distinct class, the arrangement of ni training sample of i^{th}class for $[i = 1,2, \dots \dots . k]$ as columns of matrix is $A_i = [v_{i.1}, v_{i,2}, \dots \dots v_{i,ni}]$ where $v \in \mathbb{R}^m, A_i \in \mathbb{R}^{mxni}$. For the entire training samples with k classes, the matrix is $A = [A_1, A_2, \dots \dots . . A_k]$ where$n = nixk$. Given a test sample $\in \mathbb{R}^m$can be represented by a linear representation of all training samples

$$s = Ax \ \in \mathbb{R}^m \tag{36}$$

where $x = [0, 0, 0, 0, x_{i,1}, x_{i,2}, \dots \dots \dots . x_{i,ni}, 0, 0, 0, \dots .]^T$ is the coefficient vector with entries of zeros except those associated with the i^{th} class. The following statements take place:

- If $m > n$, the system equation $\boldsymbol{s = Ax}$ will be overdetermined, and $\boldsymbol{x}$ can be determined correctly as a unique solution.
- In an under-determined case, the solution is not unique, and a nonlinear method must be used to find the nearest solution.
- When x is sparse enough, the solution can be found through an ℓ_1 minimization problem as in equation 31.

$$\hat{x} = min\|x\|_1 \quad subject\ to \quad Ax = s \tag{37}$$

Due to noise, some non zero entries are associated with the multiple object class; the test sample can be approximated as $\hat{s} = A\hat{x}$, so the sample is categorized more accurately based on how well the coefficients from each category are assigned to the object with minimum residual:

$$\min_i r_i(s) = \|y - A\hat{x}_i\|_2 \qquad i = 1,2,\ldots..k \tag{38}$$

The study in reference [6] applies this theory on MS data. The training set has been arranged in a matrix whose columns represent the intensity spectra for all training set, and its rows represent the features. Since the number of features are much larger than the training sample, the authors propose, instead, to deal with high dimension features by projecting these features to low dimension space using a sensing matrix. Using this projection, the information is preserved as an advantage from the CS paradigm. Thus equation (29) will be as follows:

$$\phi s = \phi A x$$

$$y = \theta x \square (39) \square\square \text{where} \tag{39}$$

where $y = \phi s$ and $\theta = \phi A$. The authors claim that even with this low dimension representation, the classification using a sparse representation still fits. The results show

that, by using this predictor, the recognition rate will be better than PCA and wavelet algorithms.

The authors in [57] proposed sparse representation classification (SRC) for feature selection on mass spectrometry proteomics data. The proposed algorithm starts with the k-cross validation, preprocessing steps for entire dataset. Then, the authors split the dataset to training and test sets. Each subset is split again to d sets. Finally, they randomly selected the features from the training set and divided into the set containing h features. A decision tree, which can provide the relative importance of features in a particular subset, is based on the score of features.

In [7] the authors propose using the fractal and entropy space as a recognition scheme, and then they applied NN classifier to classify the test signal and find a train set sample close to it. The fractal and entropy feature factor for a signal $x = [x_1, x_2, \dots \dots x_N]^T$

$$Fractal = Fx = 2 - \frac{log(\sum_{i=2}^{N} |x_i - x_{i-1}|)}{-log(M/N)} \tag{40}$$

$$Entropy = Ex = -\sum_{i=1}^{N} |x_i| log(|x_i|) \tag{41}$$

where N is the dimension of x; and M is the number of measurements which have been taken. Then, the fractal entropy feature vector has been formed as

$$FEx = \begin{bmatrix} F_x \\ E_x \end{bmatrix} \quad (42)$$

Now, the proposed algorithm can be stated as the following:

- Assume the input MS data is x; the sensing data y will be calculated using the sensing matrix ϕ.

$$y = \phi x \quad (43)$$

- $F_y \; and \; E_y$ are to be computed for input sensing data and all prior classificatory training sets.
- The Euclidean distance y and all train sets will be calculated as:

$$\left\| FE_y - FE_{yi} \right\|_2, i = 1,2, \dots \dots . D \quad (44)$$

where D is the number of train set.

- Finally, use the NN classifier to find the minimum distance to the train set y_s.

3.2. Recovery Data

The second issue is recovering the MS data from sensing data. Since the MS is high dimension data, we consider the dimensionality reduction of this data under the condition that the original data has been preserved to recover with the original dimension.

For instance, clinically, if the sample recognized is a disease sample, it is necessary to recover data and get more details such as stage of disease. In addition, projecting the database in low dimension is good for updating the database and adding more samples.

There are many applications where the CS has been proposed for biometrics. The authors in [58] applied CS as a technique for reducing the huge amount of data. This is accomplished by peak picking in spectra using the L-norm of the Image Mass Spectrometer (IMS), while considering the TV norm as the deciding algorithm. In [59] many recovery formulas were created for IMS data. A Kronecker Compressive Sensing (KCS) for multidimensional signal was invented as an example for IMS [60] [61]. However, in the recent studies, there is not enough attention given to MS based on CS.

Since the MS data is not sparse, the sparse difference (SD) [6] [7] will be applied. The SD, or disease fingerprint (data fingerprint), can be established by finding the difference between disease and health features. The difference between two MS samples from the same category can be considered as sparse.

If $x \in \mathbb{R}^n$ represents the original MS test sample, $x_s \in \mathbb{R}^n$ is the closest sample to x from the training set in the same category. By projecting the MS in low dimension using sensing matrix $\phi \in \mathbb{R}^{mxn}$ where $m \ll n$, the new projected data will be as follows:

$$y = \phi x \tag{45}$$

$$y_s = \phi x_s \tag{46}$$

Subtract equations (45), (46) and we will get

$$y - y_s = \phi(x - x_s) \tag{47}$$

According to the authors, y_s is derived from the classification process, and y is derived from equation (45); then SD is applied:

$$\Delta y = y - y_s \ \ and \ \ \Delta x = x - x_s \tag{48}$$

Then the sparse difference compressive sensing (SDCS) recovery takes place:

$$min\|\Delta x\|_1 \quad s.t \quad \phi\Delta x = \Delta y \tag{49}$$

Once Δx has been recovered, the MS can be found as follows:

$$x = x_s + \Delta x \tag{50}$$

CHAPTER IV

PROPOSED FRAMEWORKS- CS TOOLS FOR MS CLASSIFICATION AND RECOVERY ALGORITHM

4.1 MS Classification Based on CS Framework

The MS data is usually a very high dimension, and the classification process is computational. The main objective is to propose a classifier that will be accurate and built according to the low dimension for MS data classification. In this algorithm, the entire MS spectra is involved in the classification decision. This involves either selecting some features from the MS or considering that all MS features in the high dimensional spectra have disadvantages. By acquiring the MS data through CS sampling, the sensing data has not only a lower dimension than the original data, but also the whole set of information is preserved. The classification will be made according to the low dimension data leading to faster processes without losing accuracy.

For the accuracy of classification prospective, this dissertation's proposed algorithm is expected to achieve higher performance when compared with other common classification techniques such as PCA, PCA/LDA and feature selection classifiers. One reason for this is that only the high dimension and a few selection feature issues are considered in the proposed algorithm. Furthermore, the reconstruction of original MS from low dimension data using a CS technique is the second purpose that the predictor has to achieve. Since the MS data is not sparse, BSBL will be used for the recovery prospective.

Applying the compressive sensing algorithm is the essential step to produce a low dimension presentation of MS data. This dissertation is particularly interested in a method that produces optimal and robust solutions in the MS data case where the following assumptions are considered:

1. The data is noisy.
2. The collected data (MS sample) is of high dimension [typically 10^5 to 10^8].
3. The number of samples in the database is relatively small [typically 10^2to10^4].
4. Selecting a number of measurements from the data in low dimension, this method will produce a robust recognition and high quality recovery data.

The orthonormal $L2$ norm method takes advantage of the first assumption where the entire MS spectrum has been considered. In addition, the low dimension takes advantage of the second assumption that has been addressed by projecting the MS data in low dimension. More details about the proposed method are described in the next section.

4.1.1 The Orthonormal $L2$ Norm Method

The first step is rearranging the training set of MS data in a proper way. Each sample is represented by a vector pair $\{{}^{m}/_{Z}, I\} \in \mathbb{R}^N$ where m/z is the mass to charge ratio and I is the spectral intensity. Then we stack ni columns of i^{th} class as $x_i = \{I_{i,1},\ I_{i,2}, \dots \dots I_{i,ni}\} \in \mathbb{R}^{Nxni}$. Then the training set containing the n samples belonging to K classes can be represented as $X = [x_1, x_2, \dots \dots . x_k] \in \mathbb{R}^{Nxn}$, thus $n = \sum_{i=1}^{K} n_i$. Any

test sample, $x \in \mathbb{R}^N$, can be a subset of the training set of an unknown class. In [62], the test sample can be represented as a linear combination of the training set.

$$x = Xr\,, \qquad x \in \mathbb{R}^N \tag{51}$$

where $r = [0,0,0,\ldots\ldots, r_{i,1}, r_{i,2}, \ldots. r_{i,ni}, 0,0,0]^T \in \mathbb{R}^n$ represents the coefficients vector (all zeros except those associated with i^{th}class) that needs to be estimated. When $N < n$, then there are fewer constraints than unknowns and the system equation $x = Xr$ is underdetermined, with an infinite number of solutions; r can be found with a non-unique solution. That means that many choices of r lead to the same x [63]. While the smallest (sparsest) solution can be found using L_1 norm, others chose to use nonlinear methods to find the nearest solution such as convex optimization [20] and Newton methods [64]. It is proposed to project the original high dimensionality data to a much lower one using a sensing matrix and taking advantage of CS framework both in this work as well as in [62]. However the MS data is an overdetermined system. Therefore, by taking advantage of a CS framework, the sensing data can be acquired through CS. Instead of dealing with the X matrix, our MS data set, a new sensing data that it will be generated as:

$$y = \phi x = \phi Xr = Yr \tag{52}$$

where $Y = [y_1, y_2, \ldots\ldots y_k] \in R^{Mxn}$ where $\phi \in \mathbb{R}^{MxN}$ is the transformation matrix($\mathbb{R}^N \rightarrow \mathbb{R}^M$). In general, $\boldsymbol{M}$ has to be much smaller than $\boldsymbol{N}$, to satisfy the underdetermined condition. From that projection, r can be found by applied an $L1$ minimization problem:

$$\hat{r} = \min\|r\|_1 \qquad subject\ to\ \ Yr = y \tag{53}$$

Due to high dimensionality of MS features comparing number database samples, we still have an overdetermined system. In contrast to the L1 case in [62], it is possible to estimate *r* using L_2 norm by solving:

$$min\|y - Yr\|_2^2 \tag{54}$$

The $L2$ advantage over $L1$ is that it can be applied in an overdetermined case. The $L2 - norm$ is not robust for sparse detection due to overfitting. To overcome the limitation of L1 and L2 overfitting, the regularized regression method that linearly combines the L1 and L2 penalties has been applied [65]. Therefore, the equation(48) can be written as:

$$min\|y - Yr\|_2^2 \ + \lambda_1\|r\|_1 + \lambda_2\|r\|_2^2 \tag{55}$$

where the term $\lambda_1\|\mathrm{r}\|_1 + \lambda_2\|\mathrm{r}\|_2^2$ is known the Elastic net penalty and both the trade-off parameters λ_1 and $\lambda_2 \geq 0$ compromise between model complexity and results accuracy. The equation (49) is equivalent to the optimization problem:

$$\underset{r}{\mathrm{argmin}}\ \|y - Yr\|_2^2 \qquad s.t \qquad \lambda_1\|r\|_1 + \lambda_2\|r\|_2^2 \tag{56}$$

Due to noise, some non-zero entries are associated with the multiple object class; therefore, we can approximate the test sample $\hat{\mathrm{y}}_\mathrm{i} = \mathrm{Y}\hat{\mathrm{r}}_\mathrm{i}$ to create a better classifier based

on how well the coefficients from each category are assigned to the object that minimizes the residual.

$$\min_i r_i(y) = \|y - Y\delta r_i\|_2 \qquad i = 1,2,\ldots..K \tag{57}$$

where $\delta r_i \ for \ (\forall \ i)$ are the sparse coefficients of class i. This is derived by inputting all elements of $\hat{r}$ zeros except the coefficients of class i. The flowchart below shows the main steps of the proposal L2 algorithm.

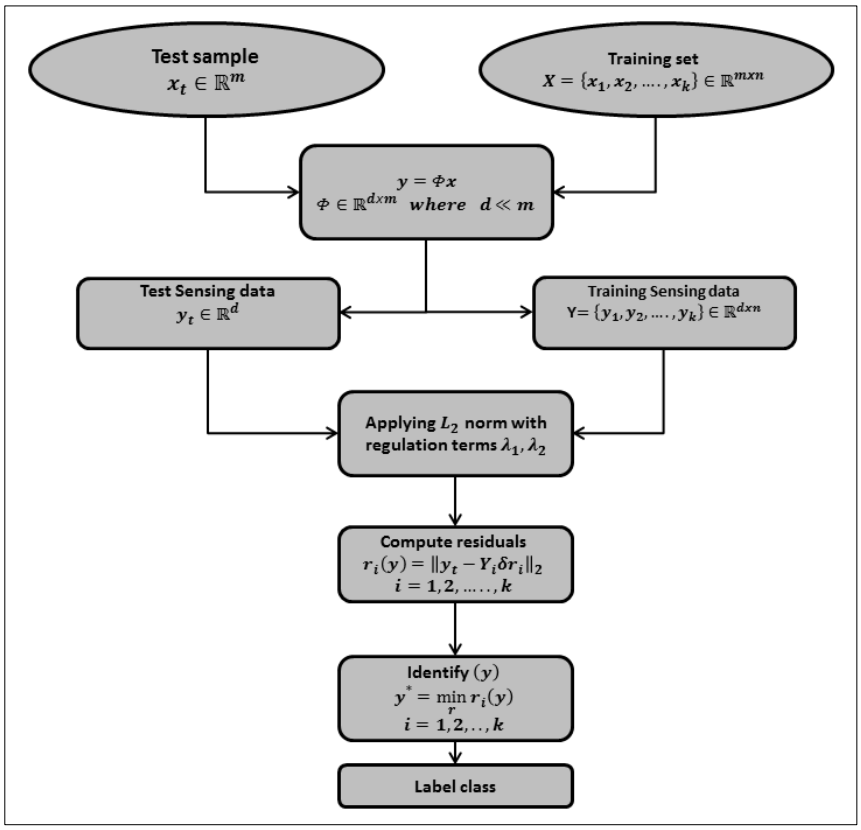


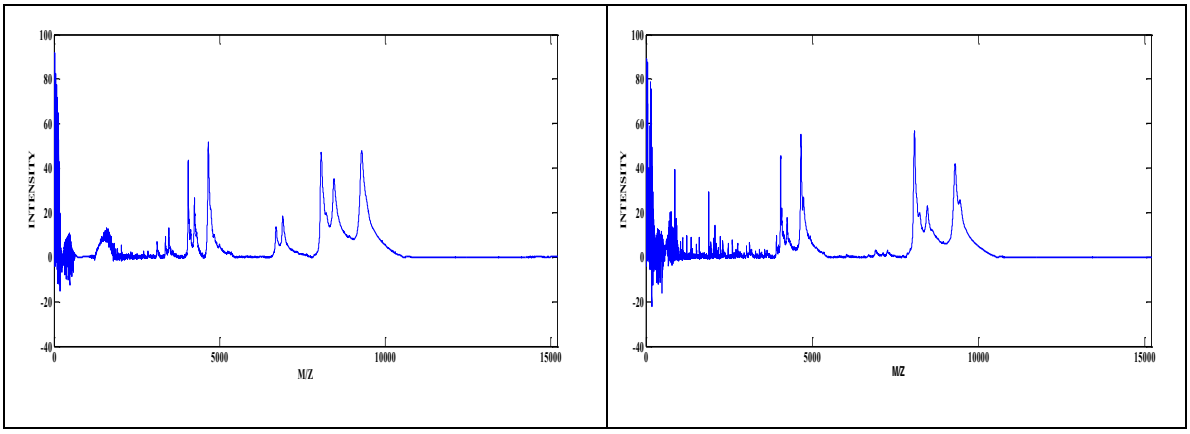
Figure 21. The general steps to apply the orthonormal L2 algorithm for MS data classification

4.1.2 Significance of Combining L2 with a Regularization Algorithm

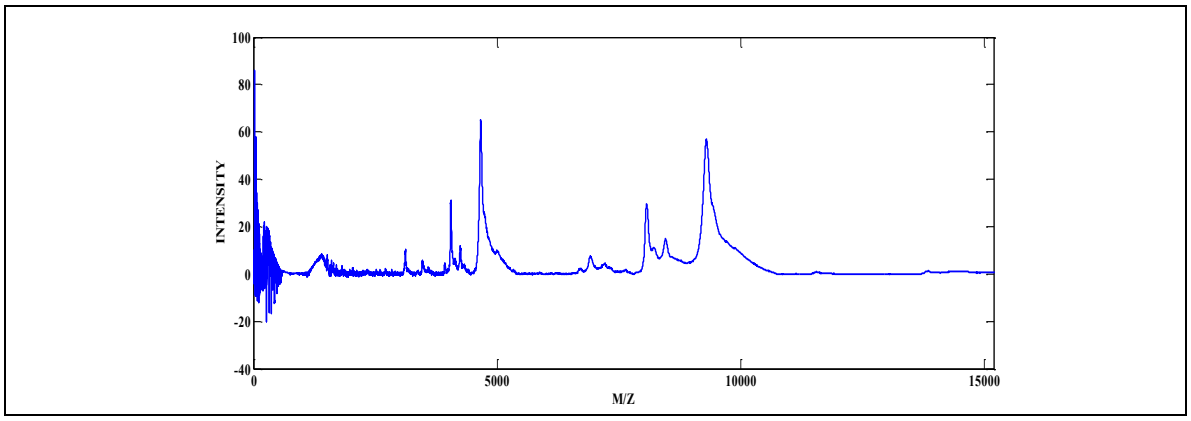
The two MS classification categories, namely the complete and partial set, have several disadvantages. The full set technique is computationally intensive due to the dimensionality of MS data despite the fact that all intensity peaks are considered in the decision for classification. The subset technique excludes some intensity peaks using the preprocessing procedure, or it selects just specific features for classification. This is a problem since all peaks are very important especially in clinical applications, therefore removing peaks that may have very important details will result in misclassification. However, the proposed algorithm in this dissertation has an advantage in that the whole data set is considered in the low dimension set. By projecting the data in low dimension space, which is a normal step in compressive sensing, the CS frame work uses the L1-norm as the sparse projection. The MS data provides a significantly poor performance because the produced data (MS sample) is high dimensional where the number of samples in the database is relatively small. Furthermore, adding the regularization terms depending on L1 and L2 will result in enhancing classification performance.

4.2 Reconstruction Data

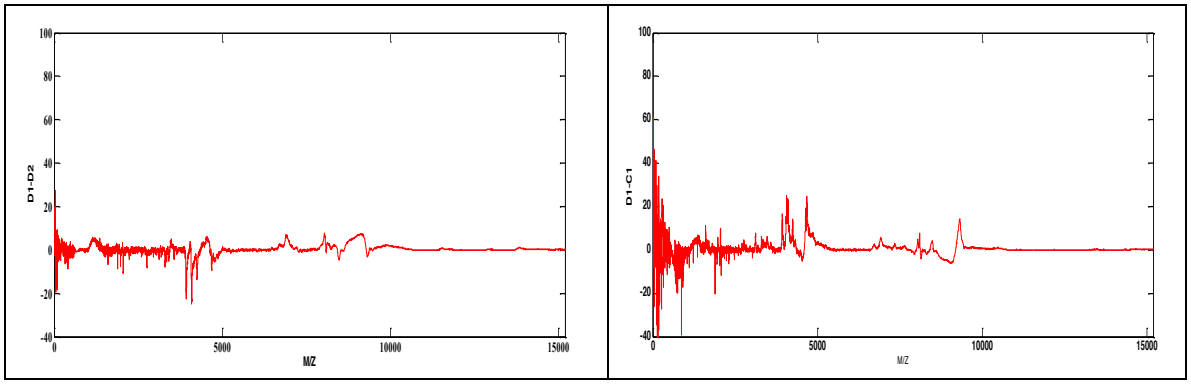
Another advantage of the L2 orthonormal algorithm is that, by selecting enough features (number of samples of m rows in the ϕ matrix) in sensing data used for classification, these features can be used to recover the original MS data.



Prostate cancer sample (D1) Healthy sample (C1)



Prostate cancer sample (D1)



The difference between (D1-D2) The difference between (D1-C1)

Figure 22. Difference between two disease samples and a disease sample with a healthy sample in prostate cancer MS dataset [66]

As a result of L2 classification, which has been derived according to a low dimension training set, we are able to identify the nearest sample $y^{\star}$ to a test sample. Since all information of MS has been preserved by projected data in low dimension, this study can still use this data as a fingerprint:

$$\boldsymbol{y_{FP} = y_t - y^{\star}} \tag{58}$$

Equation (50) finds x_{FP} which is considered sparse data in a high dimension space as shown in Figure(17). The difference between two disease samples (D1-D2) is considered as sparse. Many CS algorithms have been proposed to recover original data from compressed data. Each one can provide good performance depending on the original data. The quality performance includes the cost function, the run time, recovery error, etc. The L1 minimization [6] can be used to achieve this task.

$$min\|x_{FP}\|_1 \quad s.t \ \|y_{FP} - \phi x_{FP}\| \leq e \tag{59}$$

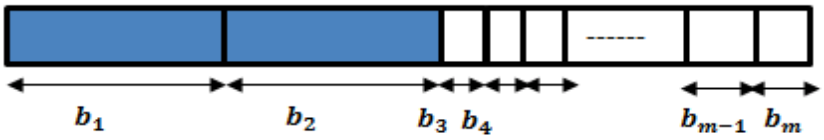
In fact, due to the MS instrument error, which causes a shift of feature around 0.2*M/Z [45], the number of sparse coefficients in the disease fingerprint will be large; just a few elements have zero entries. Recent algorithms such as **Block Sparse Bayesian Learning (BSBL)** [67] were proposed as new methods to solve the CS of non-sparse data problem. The BSBL family has been applied for non- sparse signals such as EEG, ECG by exploiting the intra-correlation of a block itself [68] [69].

4.2.1 Block Sparse Framework

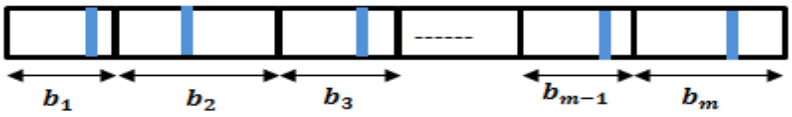
The block sparse framework has been used as a new sparse recovery for signals where the non-zero coefficients occur in cluster. The block structure can be found in many applications such as the multi-band signal [70], clustering of data on multiple subspaces [71] [72], and the multiple measurement vectors (MMV) problem [73]. In each application, there is a specific technique to recover the block representation as a new framework for CS. In this section, the block sparse recovery techniques will not be mentioned because it is out of the scope of this dissertation, but the general structure of block sparse will be explained.

$$x = [\underbrace{x_1, x_2, \dots . x_b}_{x^T(1)}, \underbrace{x_{b+1}, x_{b+2}, \dots . x_{2b}}_{x^T(2)}, \dots \dots \dots, \underbrace{x_{N-b-1}, \dots \dots . x_N}_{x^T(M)}]^T \quad (60)$$

In general, for a given vector, $x \in \mathbb{R}^N$. To define block sparsity, the vector x is viewed as concatenation of blocks with length $\boldsymbol{b}$, so $x(\ell)$ is denoting to the ℓ^{th} block. Figure(23) shows the block sparse concept. In (a) two blocks only are non- zero entries; in this case the vector is block sparse, but is not a sparse vector. In (b), the sparse vector and non-sparse block vector are shown; most blocks have non- zero coefficients revised from [74].



(a)



(b)

Figure 23. The vector $\boldsymbol{x} \in \mathbb{R}^{N}$ divided into $\boldsymbol{m-block}$ where the number of blocks <<N. The color space indicates non- zero coefficients. In (a) block sparse vector. (b) sparse vector

4.2.2 Sparse Bayesian Learning(SBL)

This section reviews the methodology of sparse Bayesian learning. Initially, the SBL has been proposed [75] [76] to find a robust sparse solution of a linear model system in context of regression and classification. The key feature of this methodology is seeking to find weight vectors with a few non-zero elements using the optimization of prior parameters. For a given set of input and target data pairs $\{x_n, t_n\}_{n=1}^{N}$, typically the target data prediction is represented based on a scalar-valued function$f(x)$. This function will be modelled by a linear-weight vector with a sum of the M term as:

$$\hat{f}(x) = \sum_{m=1}^{M} w_m \phi_m(x) = \phi w \tag{61}$$

where $\phi = [\phi_1, \phi_2, \dots\dots, \phi_M]^T$ is a general MxN design matrix with N columns, and $w = [w_1, w_2, \dots\dots, w_M]$ is a vector of weights that to be estimated. The sparsity property

will arise if some coefficients of w are set to zero. In general, the target samples can be written from the model with additive noise as:

$$t_n = f(x_n) + \varepsilon_n \tag{62}$$

where ε_n are independent samples from the noise process which is assumed, for sparse Bayesian framework, to be mean-zero Gaussian with a variance σ^2 [75]. This assumption is giving a multivariate Gaussian likelihood: a complete target set t which can be written as:

$$p(t|w, \sigma^2) = (2\pi)^{-N/2}\sigma^{-N}\, exp\left\{-\frac{1}{2\sigma^2}\|t - \phi w\|^2\right\} \tag{63}$$

One advantage of the multivariate normal distribution stems from the fact that it is mathematically tractable, and quality results can be obtained. The maximum likelihood estimation of w and σ^2 lead to over-fit. One study [75] encodes a preference for smoother (less complex) functions by making the popular choice of a zero-mean Gaussian prior distribution over *w*:

$$p(w|\alpha) = \prod_{i=0}^{M} \mathcal{N}(w_i|0, \alpha_i^{-1}) \tag{64}$$

where the key to the model sparsity is the use of *M* independent hyperparameters, $\alpha = (\alpha_1,\ \alpha_2, \ldots\ldots, \alpha_M)^T$ one per weight (or basis vector), which moderate the strength of the prior.

Having defined the prior, Bayesian inference proceeds by computing, from Bayes' rule, the posterior overall unknowns given the data:

$$p(w, \alpha, \sigma^2|t) = \frac{p(t|w, \alpha, \sigma^2)p(w, \alpha, \sigma^2)}{p(t)} \tag{65}$$

Due to complexity of computing the posterior directly from this form, in [75] [77] the postior is instead decomposed as:

$$p(w, \alpha, \sigma^2|t) = p(w|t, \alpha, \sigma^2)p(\alpha, \sigma^2|t) \tag{66}$$

The postior distribution over the weights is given by:

$$p(w|t, \alpha, \sigma^2) = \frac{p(t|w, \sigma^2)p(w|\alpha)}{p(t|\alpha, \sigma^2)} \tag{67}$$

Given α, the posterior parameter distribution is Gaussian; given via Bayes' rule, it is $p(w|t, \alpha) = \mathcal{N}(w|\mu, \boldsymbol{\Sigma})$ with

$$\Sigma = (\sigma^{-2}\phi^T\phi + A)^{-1} \tag{68}$$

$$\mu = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\phi}^{\boldsymbol{T}}\boldsymbol{t} \tag{69}$$

$$\mu = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\phi}^{T}\boldsymbol{t}$$

A is defined as $diag(\alpha_1, \alpha_2, \ldots\ldots, \alpha_M)$. In sparse Bayesian learning, the estimating solution w is given by the Maximum-A-Postior (MAP). The hyperparameters are estimated from the data by marginalizing over the weight and then performing a Maximum Likelihood (ML) optimization.

4.2.3 Block Sparse Bayesian Learning (BSBL)

This section briefly describes the BSBL framework [67] .The BSBL framework exploits the temporal correlation to improve the performance of SBL. Under this assumption the signal has a block structure. The basic model is given as:

$$y = \phi x + \boldsymbol{\epsilon} \tag{70}$$

where $y \in \mathbb{R}^{mx1}$ is a known measurement vector and $\phi \in \mathbb{R}^{mxn}$ is a sensing matrix, where columns represent an overcomplete basis. ϵ is a noise process that is assumed to be mean-zero with a variance of σ^2 Gaussian distribution $p(\epsilon, \sigma^2) = \mathcal{N}(0, \sigma^2)$. $x \in \mathbb{R}^n$ is a weight vector to recover and can be viewed with non-overlapping blocks; most blocks/groups are zero entries, and the signal is called block sparse.

$$x = [\underbrace{x_{1,}x_2, \ldots\ldots x_{bi}}_{x_1^T} \ldots\ldots\ldots, \underbrace{x_{g-1} + 1, \ldots\ldots\ldots x_{bg}}_{x_g^T}]^T \tag{71}$$

where b is the size of blocks, $b_i(\forall\, i)$ are not necessarily identical. The locations of non-zero blocks are unknown.

In this framework, to apply the Bayesian interference proceeds, each block $x_i \in \mathbb{R}^{dix1}$ is assumed to satisfy a parameterized multivariate Gaussian [77] [75] given by:

$$p(x_i;\ \alpha_i, B_i) \sim \mathcal{N}(0, \alpha_i B_i) \quad i = 1,2, \ldots . g \tag{72}$$

where α_i $(\forall\, i)$ are hyperparameters controlling the block-sparsity of x. Few blocks have a nonzero, and $B_i \in \mathbb{R}^{dixdi}(\forall\, i)$ are a positive defined matrix, which captures the correlation structure of blocks. The $\boldsymbol{\alpha B's}$ terms are represented in the diagonal matrix Σ_0.

$$\Sigma_0 = \begin{bmatrix} \alpha_i B_i & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha_g B_g \end{bmatrix} \tag{73}$$

Having defined the prior, using Bayes' rule, the posterior distribution of all unknowns over weight is thus given by:

$$p(x|y, \sigma^2, \{\alpha_i B_i\}_{i=1}^{g}) = \mathcal{N}(m_x\ , \Sigma_x) \tag{74}$$

where the posterior covariance and mean are respectively:

$$\Sigma_x = \left(\frac{1}{\sigma^2}\phi^T\phi + \Sigma_0^{-1}\right)^{-1} \tag{75}$$

$$m_x = \sigma^{-2}\Sigma_x\phi^T y \tag{76}$$

Once the hyperparameters $\sigma^2, \{\alpha_i B_i\}_{i=1}^{g}$ are estimated using ML, we apply the MAP estimate of x. The estimated data $\hat{x}$ can be reconstructed directly from the posterior mean:

$$\hat{x} \triangleq \sigma^{-2}\Sigma_x\phi^T y = \Sigma_0\phi^T(\sigma^2 I + \phi\Sigma_0\phi^T)^{-1} \tag{77}$$

The block sparsity is controlled by two parameters$(\alpha_i's, \Sigma_0)$; by setting $\alpha_k = 0$ for k^{th} , the block k will prune. Due to the presence of noise, $\alpha_k$will never be zero. Thus a threshold δ will be used to prune out small$\alpha_k(\forall i)$. The smaller value of threshold means fewer $\alpha_i's$ are pruned out, and thus few blocks will be zeros [78].

4.2.4 Applying BSBL for the Reconstruction Procedure

The MS data is not sparse, but the difference between the samples, which belong to the same category, can be assumed to be sparse. More precisely, it can be divided into groups, and most groups are considered null. In some cases, the pruned threshold is assumed δ). Once x_{FP} has been estimated, it can be used to reconstruct the original data as:

$$x = x^{\star} + x_{FP} \tag{78}$$

where $x^{\star}$ is the corresponding training sample of recognition result $y^{\star}$. This is summarized by applying the proposed recovery algorithm in the following steps:

- According to the classification procedure, $y^{\star}$ has been defined according to:

$$y^{\star} = \min_{i} c_i(y) = \|y - Y\delta r_i\|_2 \qquad i = 1,2,\ldots..K \tag{79}$$

- The data fingerprint y_{FP} is calculated as an equation (58).
- The linear representation of the system in equation (70) will be :

$$y_{FP} = \phi x_{FP} + \boldsymbol{\epsilon} \tag{80}$$

1. By substituting $y_{FP} = y$ and $x = x_{FP}$, this dissertation uses the BSBL algorithm through equations (71) to (76) to fit with the data:

$$x_{FP} = [\underbrace{x_{1,}x_2,\ldots..x_{bi}}_{x_1^T} \ldots\ldots\ldots., \underbrace{x_{g-1}+1,\ldots\ldots..x_{bg}}_{x_g^T}]^T \tag{81}$$

The posterior distribution of all unknowns over weight is thus given by:

$$p(x_{FP}/y_{FP}, \sigma^2, \{\alpha_i B_i\}_{i=1}^{g} = \mathcal{N}\left(m_{x_{FP}}, \Sigma_{x_{FP}}\right) \tag{82}$$

where

$$\Sigma_{x_{FP}} = \left(\frac{1}{\sigma^2}\phi^T\phi + \Sigma_0^{-1}\right)^{-1} \tag{83}$$

$$m_{x_{FP}} = \sigma^{-2}\Sigma_{x_{FP}}\phi^T y_{FP} \tag{84}$$

and Σ_0 is defined in equation(55).

2. Moreover, the output of the BSBL algorithm is

$$\hat{x}_{FP} \triangleq \sigma^{-2}\Sigma_x\phi^T y_{FP} = \Sigma_0\phi^T(\sigma^2 I + \phi\Sigma_0\phi^T)^{-1} \tag{85}$$

- Once $\hat{x}_{FP}$ is defined, the original MS test sample can be estimated as in equation (78).

CHAPTER V

EXPERIMENTAL RESULTS

In this section, the proposed algorithm has been applied for MS data. The classification of MS data is the first purpose this dissertation strives to achieve. The MS data is classified into specific known categories, and the proposed classifier is applied to classify a specific test sample to one known category. The performance of this suggested scheme will be compared with PCA/LDA and L_1 classification algorithms in the second target. In addition, in the last section, the recovery of the MS data from the classification data will be considered.

This study has been able to make use of prostate cancer SELD–TOF mass spectra data sets from the NIH and FDA Clinical Proteomic Program (http://home.ccr.cancer.gov/ncifdapromics- /ppatterns.asp) [66]. Table (3), shows this data listed according to a Prostate Specific Antigen (PSA) level. PSA is a protein produced by cells of the prostate gland. The PSA test measures the level of PSA in a man's blood. The results are usually reported as nanograms of PSA per milliliter (ng/mL) of blood. The PSA test leads into grouping the dataset into 4 classes: No evidence of disease, benign, presence of prostate cancer with a PSA level of 4-10 and prostate cancer with a PSA level of >10 ng/mL . The total number of samples in the database is 237 samples classified as shown in Table 2. Each spectrum is composed of 15,200 peaks defined by a corresponding m/z value. The goal of this study is evaluating the ability of

the orthonormal L2 norm method to discriminate prostate cancer according to the PSA[§] level from a benign condition.

Table **3**. The database of prostate cancer according to PSA level

Disease status	N
No evidence of disease and PSA level < 1 ng/mL [CLASS A]	**60**
Benign and PSA level ≥ 4 ng/mL [CLASS B]	**120**
Prostate cancer with PSA level 4-10 ng/mL [CLASS C]	**23**
Prostate cancer with PSA level >10 ng/mL [CLASS D]	**34**
TOTAL	**237**

Due to a high variation of amplitude peaks, the normalization process is used to set the intensities into new values in range with [0,1]. To normalize, equation 80 is used,

$$x_i^{normal} = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{86}$$

where $\max(x_i)$ and $\min(x_i)$ are the maximum and minimum intensity peaks.

5.1 Performance Comparison

For the assessment of classification performance, this study uses Overall Success Rate (OSR), sensitivity (Sen.), Positive Predictive Value (PPV) and Specificity (Spec.).

[§] Prostate Specific Antigen

Since there are four classes, the confusion matrix provides an adequate tool to correctly assess the model's ability to predict classes. For A, B, C and D classes, the confusion matrix is given as:

Known class		A	B	C	D
Predicted Class	A	t_{pA}	e_{AB}	e_{AC}	e_{AD}
	B	e_{BA}	t_{pB}	e_{BC}	e_{BD}
	C	e_{CA}	e_{CB}	t_{pC}	e_{CD}
	D	e_{DA}	e_{DB}	e_{DC}	t_{pD}

where, t_{pH} is the number of true (**correct)** predictions that belong to *H.* (H is one of four classes *A, B, C,* and *D.*) Whereas e_{VZ} is the number of error predictions between two classes: *V, Z* which are any of two available classes in the database.

I. OSR or accuracy is the most popular measure for classification and is defined as the trace of the confusion matrix, divided by the total number n of all classified samples. The total number of samples is 237 in our simulations:

$$\frac{1}{n}\left(t_{pA} + t_{pA} + t_{pC} + t_{pD}\right) \tag{87}$$

II. Sensitivity is estimated by measuring the ability of the classifier to recognize the positive labeled instances that are correctly identified. For example, Sen A is the sensitivity of class A given by:

$$Sen_A = \frac{t_{pA}}{t_{pA} + f_{nA}} \tag{88}$$

where $f_{nA} = e_{AB} + e_{AC} + e_{AD}$

III. The specificity would correspond to the true-negative that is predicted correctly. When the true-negative is not predicted to be a member of class A, class A would be calculated as:

$$Spec_A = \frac{t_{nA}}{t_{nA} + e_{BA} + e_{CA} + e_{DA}} \tag{89}$$

where $t_{nA} = t_{pB} + e_{BC} + e_{CB} + t_{pC} + e_{DB} + t_{pD} + e_{CD} + e_{DC} + e_{BD}$

IV. PPV or precision is defined as the total number of correctly predicted positive outcomes from the test, divided by the total number of predicted positive outcomes (correctly predicted or not). It is defined by:

$$PPV_A = \frac{\mathrm{tpA}}{\mathrm{tpA} + \mathrm{fpA}} \tag{90}$$

$where\ f_{PA} = e_{BA} + e_{CA} + e_{DA}$

Equations (87) to (90) will be applied for each class. The range of all assessment parameters are between 0 (complete misclassification) and 1 (perfect correct classification).

The database matrix has been arranged with the dimension of (15200x237) and contains four categories. To assess the L2-norm method, one sample has been selected randomly (removed from database) as a test sample. The dimensionality reduction step has been done using the sensing matrix ϕ with the number of rows (M=0.2*N where N is number of features). Each column contained random (0.125*M) entries of 1's, while other entries were zeros [79]. Figure 24 shows the sparse representation of the test sample using a linear combination of all training samples. Since the test belongs to class 2, the maximum coefficient in this sample belongs to this category. In figure 25, the test sample chosen belongs to subject 2, so one can assign the test sample to the category, which can give the best approximation $min\ r_i(y)$.

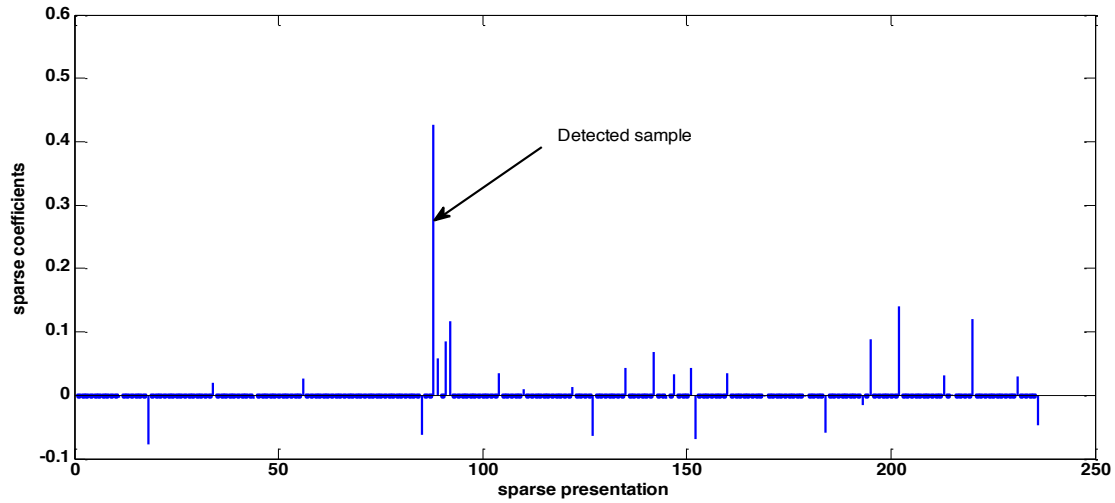


Figure 24. The sparse coefficients representation of the test sample from a linear combination of all training samples of 4 categories

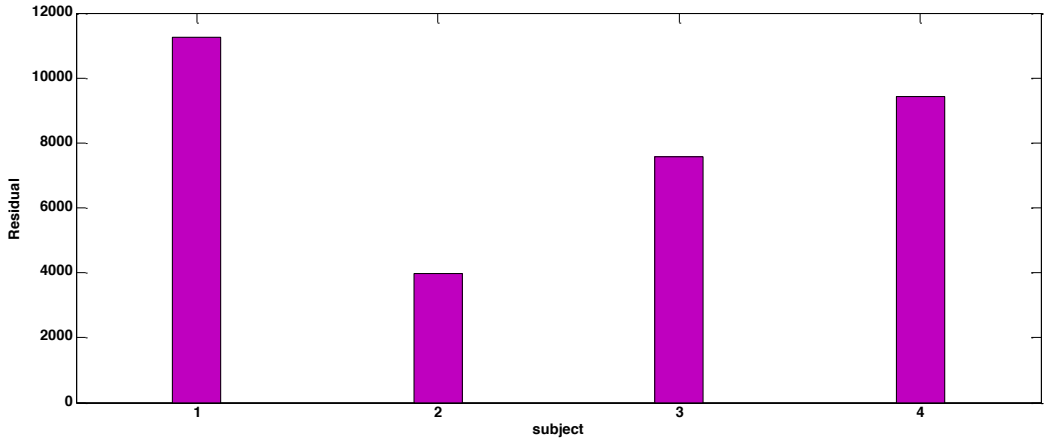


Figure 25. Histogram showing residuals $\boldsymbol{r_i(y)}$ of the test sample with respect to the projection of sparse representation computed $\boldsymbol{\delta r_i}$ by L2-norm

This study has divided the database into training and test sets and applied the L2 algorithm to the data. It also compared L2, the L2-Regulator and PCA/LDA (Q5) [49]. For the Q5 algorithm, the dimension of projected data has to be at most $(n - K)$ where n is the number of samples, and K is the number of classes to avoid the degenerate solution [49]. Since there are four classes, we use $n - 4$ largest components for PCA. Subsequently, the accuracy, PPV, sensitivity and specificity are computed for each algorithm. The results have been recorded as follows:

5.1.1 Confusion Matrix

Most measures are not processed directly from the raw classifier outputs, but from the confusion matrix built from these results. This matrix represents how the instances are distributed over estimated (rows) and true (columns) classes. Table shows the result of applying the three algorithms for classification. Each experiment has been averaged ten times with the same training/test sample sets. For L2- Regularization, let $\eta = \frac{\lambda_2}{(\lambda_1+\lambda_2)}$ and put $0 \leq \eta \leq 1$ as a condition to select those parameters. The regularization parameters ($\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$) are selected randomly 100, 0.5 ($\eta = 0.004975$) respectively. The results show the advantage of using regularization parameters to improve the classification performance.

Table 4. The confusion matrix for four classes in Q5, L2, and L2-regularization algorithms

Known Class	PCA/LDA A	B	C	D	L2- algorithm				L2- algorithm with Regularization $\lambda 1=100, \lambda 2=0.5$			
CLASS A	20	1	3	0	24	0	0	0	24	0	0	0
CLASS B	0	11	2	0	0	10	3	0	0	11	2	0
CLASS C	0	3	42	3	0	1	45	3	0	0	48	1
CLASS D	1	0	2	5	0	0	1	7	0	0	2	6

Note: M/N is the same ratio for all algorithms (0.2 of total features)

5.1.2 Classification Performance

The performance evaluation has been estimated according to confusion matrix values. The Spec, PPV and Sen, parameters are estimated for all classes. However, the OSR is found as the average value for all categories. Table shows the comparison of classification L2 performance according to four parameters (OSR, Sen, Spec and PPV). Since, the choice of test/training sets is randomly selected; each experiment has been repeated ten times. Each time, the test set is selected randomly from any class, and the algorithm must categorize which class this sample belongs to. Using this, the performance of the algorithm results will be more concrete and not just for a specific selection of sets but also the scheme can be more generalized. The average accuracies are 90.6%, 91.49% and 94.68% for the Q5 and L2- algorithms and the L2-regularization

respectively. However, the L2-regularization gives the best performance among Q5 for PPV. This is true for Sen as well; however, using Spec only for class C, Q5 is better than L2-regularization. The reason for this is that these values reflect the confusion matrix entries from the Spec formula $t_n = (1 + 3 + 11 + 2 + 2 + 5)$. Moreover, the value of t_{pA} for Q5 is less than L2 and L2-regularization $(20, 24, \text{and } 24)$ respectively. Overall, these results imply that the proposed method has more practicability than Q5. By adding regularization terms, the recognition data performance is more accurate and believable for the application of MS analysis

Table 5. The performance evaluation has been estimated according to confusion matrix

Known class	PCA/LDA			L2- algorithm			L2- algorithm with Regularization $\lambda 1=100, \lambda 2=0.5$		
	Sen	Spec	PPV						
CLASS A	1	1	1	1	1	1	1	1	1
CLASS B	0.98	0.93	0.87	0.97	0.83	0.77	1	1	0.84
CLASS C	0.97	0.64	0.84	0.93	0.93	0.92	0.91	0.92	0.98
CLASS D	0.98	0.57	0.5	0.96	0.7	0.87	0.98	0.85	0.75
	OSR_=0.906			OSR_AVG=0.9149			OSR_AVG=0.9468		

5.1.3 Comparison with the L1 Algorithm

Since the L1- norm [6] is applied to the underdetermined matrix, the number or sensing matrix (ϕ) will be very small. To see the performance of L2 with regularization

under a very small features selection, the same number of features (M) has been selected for L1, LDA and Q5. Table shows that the L1-algorithm has a better performance than Q5; however, overall, the L2- regularization has the best performance (i.e OSR 0.9321, 0.5106 and 0.9588).

Table 6. Results of classification comparison between L1- algorithm, PCA/LDA, L2-algorithm, and L2-regularization

ALGORITHM	M	TRAINING SET	TEST SET	OSR	SEN	PPV	SEPEC
L1-algorithm	140	143	94	0.9321	0.9775	0.705	0.6868
PCA/LDA	140	143	94	0.5106	0.8295	0.4401	0.4717
L2- norm	140	143	94	0.9149	0.9579	0.840	0.750
L2-REGULARIZATION	140	143	94	0.9621	0.9588	0.841	0.761

5.1.4 Training Percentage

This percentage is determined in order to make sure that the L2- norm algorithm is robust under a variety of conditions such as the size of testing samples. Table shows that the L2- regularization is still able to achieve a high performance for all assessment parameters.

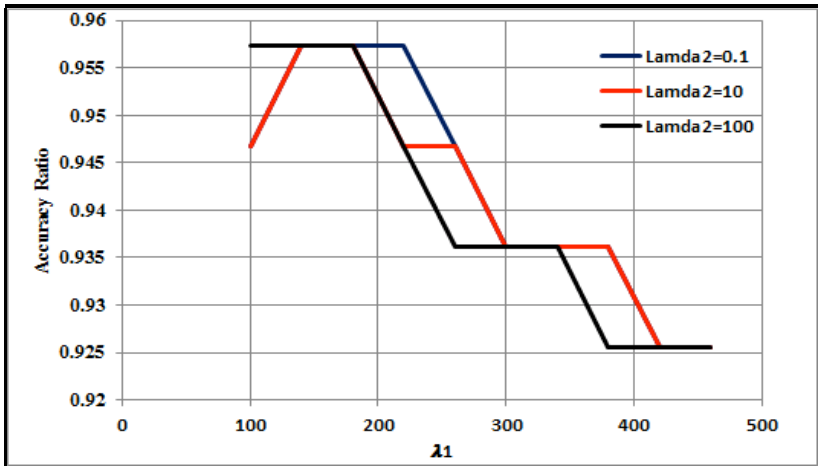
Table 4. Results of comparison between the three algorithms in terms of OSR, SPEC, PPV and SEN with different training set percentages

Training %	PCA/LDA				L2- NORM				L2- REGULARIZATION $\lambda 1=100$ $\lambda 2=0.5$			
	OSR	SPEC	PPV	SEN	OSR	SPEC	PPV	SEN	OSR	SPEC	PPV	SEN
30%	0.82	0.95	0.73	0.707	0.89	0.964	0.816	0.847	0.93	0.978	0.86	0.865
60%	0.92	0.98	0.87	0.805	0.92	0.968	0.787	0.890	0.95	0.974	0.95	0.893
75%	0.85	0.95	0.75	0.788	0.88	0.956	0.820	0.837	0.95	0.974	0.95	0.893

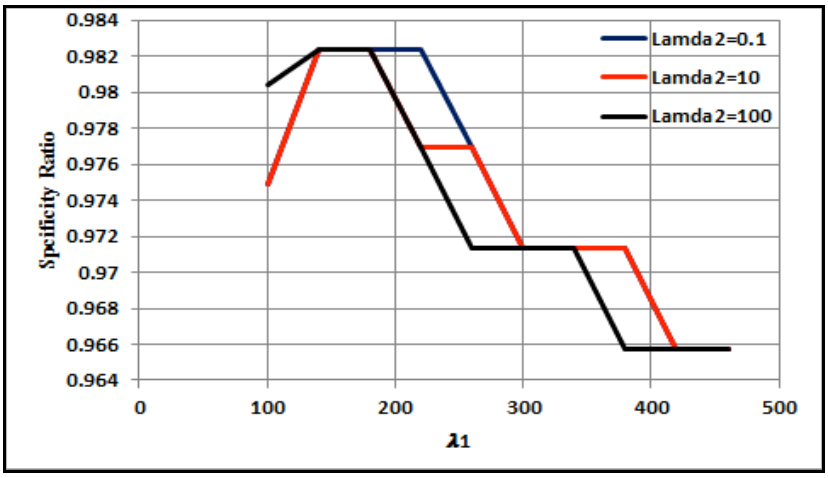
5.1.5 Setting the Regularization Parameters

To show the validity of *L2*-regularization, we implemented the algorithm using different values of η based on λ_1 and λ_2 for the range of $0 \leq \eta \leq 1$. For $\eta = 0$, the *L2* penalty term was removed while the *L1* term is null for $\eta = 1$. Figure (4) shows the effect of the regularized parameters of *L1* and *L2* in the all the assessed performance parameters. All parameters have been affected by $\boldsymbol{\lambda}_1$ term with different values of $\boldsymbol{\lambda}_2$. However, the best performance for the all parameters based on parameter $\boldsymbol{\lambda}_1$ was in the range $140 \leq \lambda_1 \leq 180$. Improved algorithm performance, as indicated by increased measures of OSR, SPEC, SEN and PPV, will allow for higher data throughput and

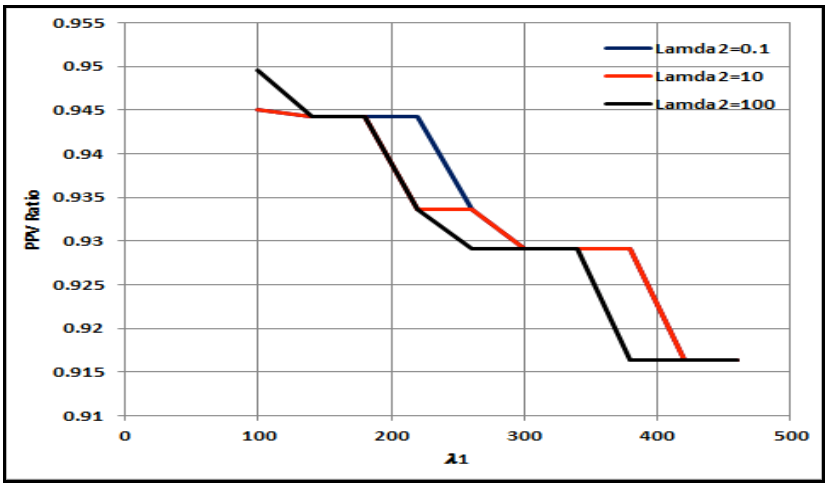
accurate measurements in applications such as disease biomarker detection in a laboratory or clinical setting.



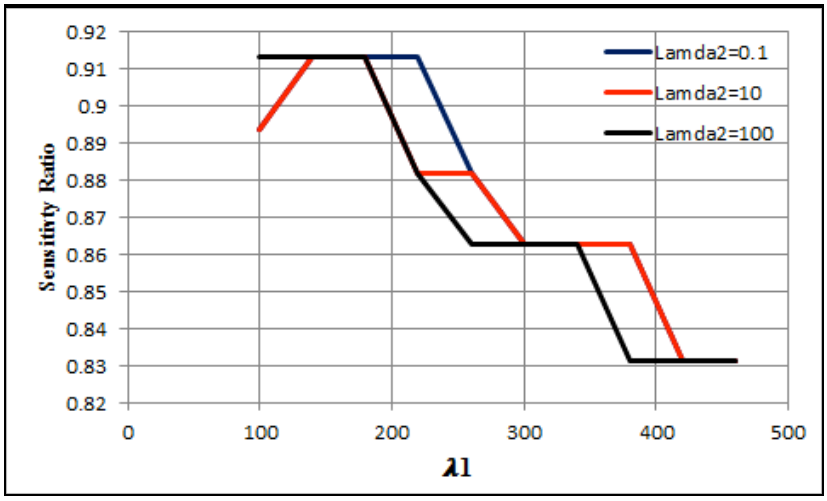
(a)



(b)



(c)



(d)

Figure 26. The regularization parameters versus the performance parameters (a) accuracy, (b) Specificity, (c) PPV and (d) Sensitivity

5.1.6 Relationship to Nearest Neighbor and Nearest Subspace

Unlike the conventional classification poplar methods; such as Nearest Neighbor (NN) and Nearest Subspace (NS) [80] [81], sparse representation uses all training

samples of all subjects to represent a test sample. In a NN classifier, the test sample y_t assigned to the subject i according the smallest distance from y_t to the nearest training sample of subject i:

$$d_i(\mathrm{y}) = \min_{j=1,2,\ldots n_i} \|y_t - y_{i,j}\|_2 \tag{91}$$

Figure 27 shows the distances between a sensing test sample that has been selected from subject 4 and all training samples (317 MS prostate cancer sample). Although the smallest distance is correctly associated to class 4, the variation of distances for other subjects is quite large. Due to the noise and the variation of MS data samples, the NN classifier may perform poorly.

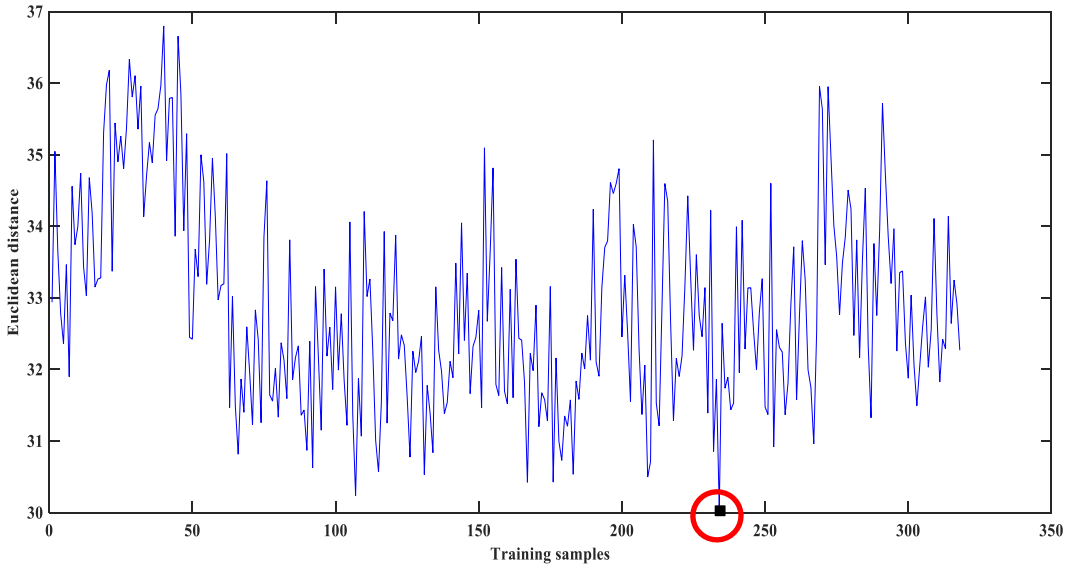


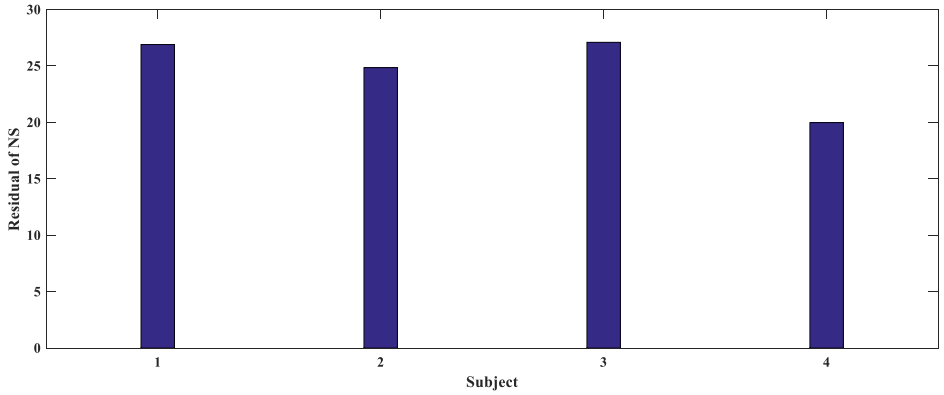
Figure 27. The Euclidean distances between the test sample and 317 MS prostate cancer sample

In an NS classifier, unlabeled sample y_t is classified to subject i if L2- norm distance from y_t to subspace spanned by all samples $y_i = [y_{i,1}, \dots\dots, y_{i,n_i}]$ is nearest among all subjects.

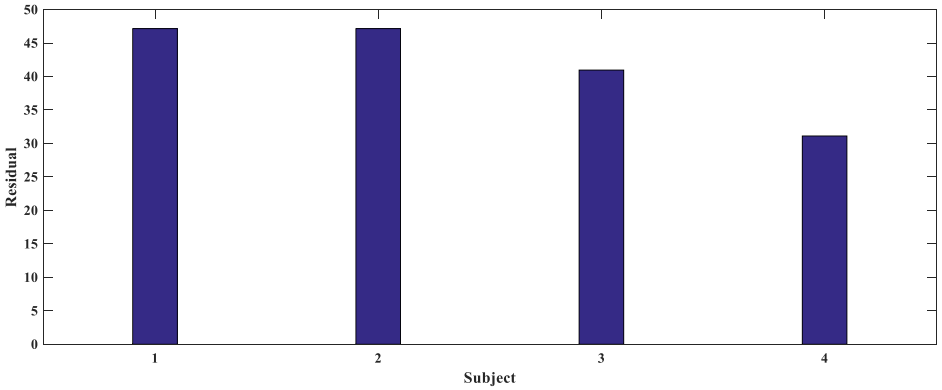
$$y_t = Y_i r_i + z_i \ , i = 1,2,..,k \tag{92}$$

$$d_i(y) = \min_{r_i \in \mathbb{R}^{n_i}} \|y_t - Y_i r_i\|_2 \tag{93}$$

where r_i is a weighing vector in class i and $\|z_i\|_2 \leq \varepsilon$ for a small $\varepsilon > 0$. Each r_i is an optimal representation of y_t in terms of subject i. In L2-regularization algorithm, only one of $\{Y\delta r_i\}_{i=1}^{k}$ is optimal and the rest have small norm. Figure 28 shows the same sample (from subject 4) as the last example. The results show that NS and L2-regularization classified the sample belonging to class 4, but the variation between two smallest residuals are 4.8714 and 9.8469 respectively. In other words, the L2-regularization algorithm is more discriminative than the NS classifier.



(a)



(b)

Figure 28. The residuals of the test sample from subject 4. (a) for NS classifier and (b) for L-regularization

Table 7 shows that the comparison of performance parameters in NN, NS and L2-regularization. The MS data was in sensing form with a compression ratio of 0.25 and 40% in the testing sample and the rest are training samples. The results confirmed that the proposed algorithm has more robust than both the NN and NS classifiers.

Table 5. Comparison of L2_regularization algorithm performance based on sparse representation with NN, NS methods

Classifier	ORS	Spec	PPV	Sen
NN	0.5433	0.8173	0.3704	0.3742
NS	0.8819	0.9455	0.8308	0.7668
L2-regularization	0.9449	0.9822	0.8954	0.9312

5.2. Ovarian Cancer Database

To verify our proposed framework, we used ovarian cancer WCX2 SELD–TOF mass spectra datasets from the NIH and FDA Clinical Proteomic programs (http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp). Two hundred and seventeen serum samples were used, including 17 cases of benign cancer, 50 cases of cancer A disease, 50 cases of cancer B disease, 50 cases of healthy control C and 50 cases of healthy control D, were examined by SELDI-TOF-MS with WCX2 protein-chips. Each spectra contained the 15154 feature.

To assess the L2-norm method, one sample was selected randomly as a test sample from normalized samples. The dimensionality reduction step was applied using the sensing matrix Φ with the number of rows (M=0.25*m where m is number of features). Each column contained random (0.125*M) entries of 1's, while other entries were zeros. In Figure 29, the test sample chosen belongs to subject 4, so one can assign the test sample to the category, which gives the best approximation $\min\ r_i(y)$.

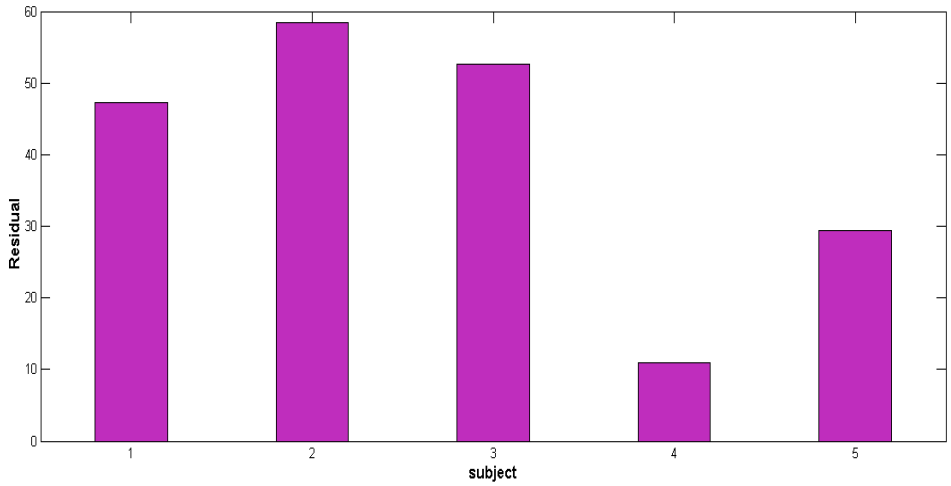


Figure 29. Histogram of residuals $r_i(y)$of the test sample with respect to the projection of sparse representation computed δr_i by L2 $- norm$

Since there are five classes, the confusion matrix provides an adequate tool to correctly assess the model's ability to predict classes. For the comparison, the accuracy, Sen, PPV and Spec of our method and PCA/LDA were obtained from the average of 10-fold cross-validation. The data set was divided into a 70% as a training set and the rest was reserved as a testing set. **Error! Reference source not found.** shows the confusion matrix for PCA/LDA and the least square method and L2 with mixing L1, L2 norms as regularization terms.

Table 9. Confusion matrix used to assess the performance of the classification algorithms

	PCA/LDA					L2-algorithm					L2- regularization				
Known class	Begin	A	B	C	D	Begin	A	B	C	D	Begin	A	B	C	D
Begin	4	0	0	2	0	4	2	0	0	0	5	1	0	0	0
cancer A	1	16	3	0	0	1	18	1	0	0	1	18	0	1	0
cancer B	0	4	15	1	0	0	2	18	0	0	0	2	18	0	0
control C	0	0	0	19	1	0	0	0	19	1	0	0	0	19	1
control D	4	0	0	0	16	0	0	2	1	17	0	0	0	1	19

Based on the results in the confusion matrix, the Sensitivity, Specificity and PPV will be calculated as a performance assessment of each classifier. The results are recorded in table 10. The regularization parameters are assigned the values of respectively. L2 algorithm performance had been improved by adding the regularization parameters as it is expected.

Table 10. 10- fold validation of PCA/LDA, L2- algorithm and L2- regularization

	PCA/LDA			L2-algorithm			L2- regularization		
classifier	Spec	PPV	Sen	Spec	PPV	Sen	Spec	PPV	Sen
Begin	0.937	0.444	0.667	0.987	0.800	0.667	0.987	0.833	0.833
cancer A	0.939	0.800	0.800	0.939	0.818	0.900	0.954	0.857	0.900
cancer B	0.954	0.833	0.750	0.954	0.857	0.900	1.000	1.000	0.900
control C	0954	0.863	0.950	0.985	0.950	0.950	0.960	0.905	0.950
control D	0.984	0.941	0.800	0.985	0.944	0.850	0.985	0.950	0.950

However, the accuracy for each algorithm calculated at different number of features and those features were selected randomly. For L2-rgularization, the features represent the sensing ratio(M/m). Figure 30 shows that PCA/LDA performed poorly at all features selections.

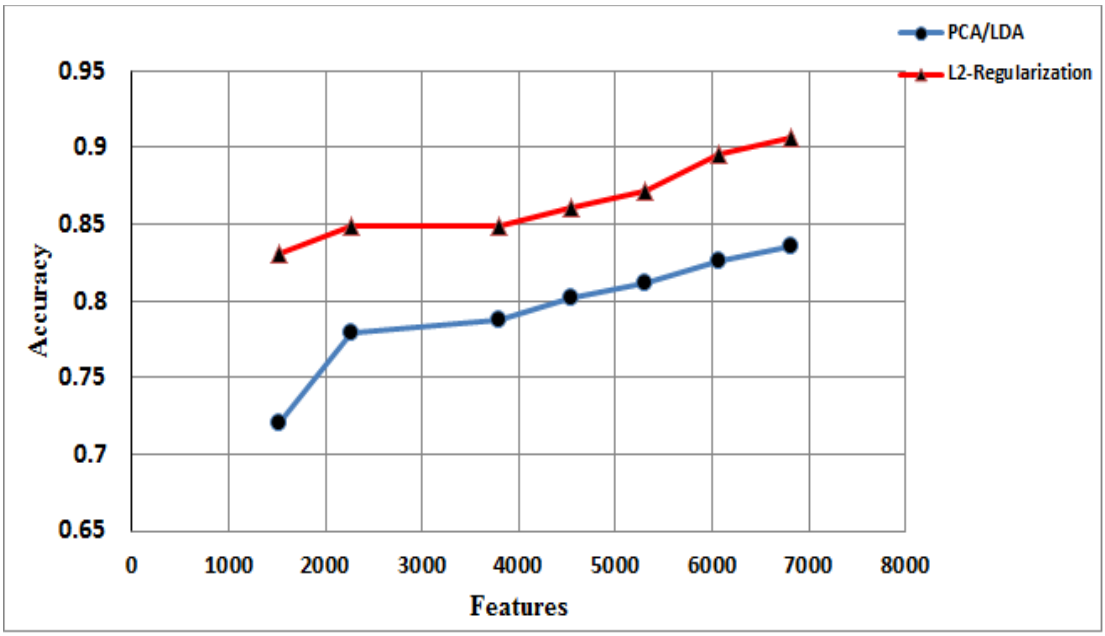
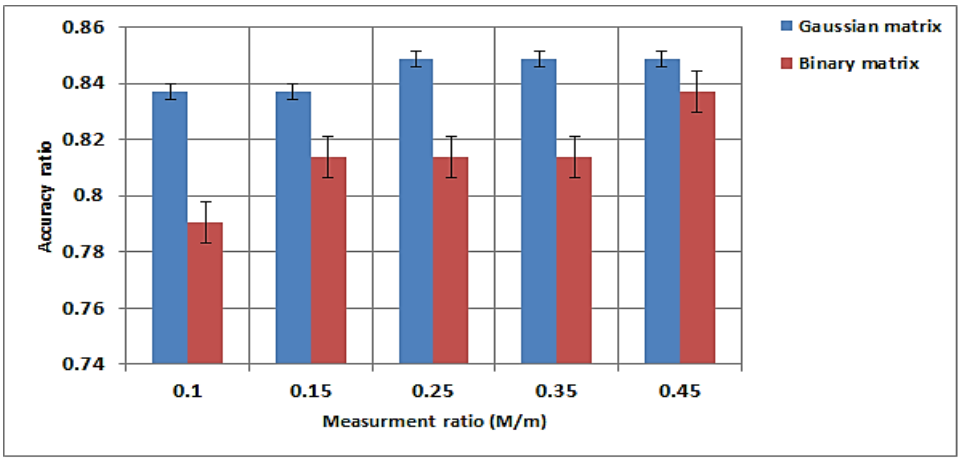
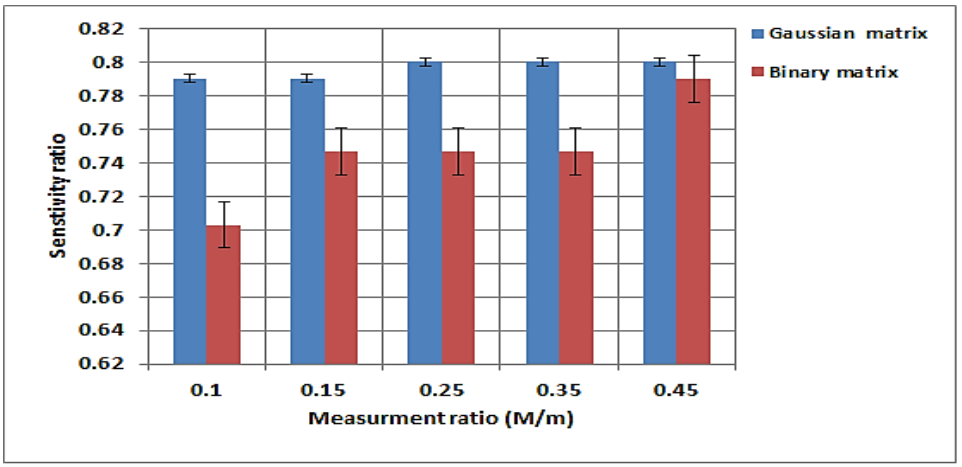


Figure 30. Classifier accuracy versus the features for PCA/LDA and L2-regulation (λ1=10,λ2=1.)

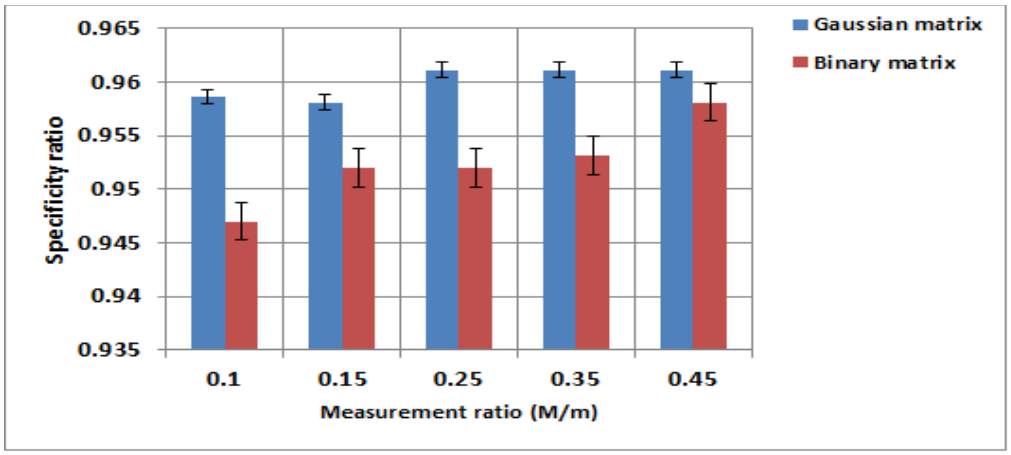
Since, the sensing data has been selected by applying the sensing matrix($\phi \in \mathbb{R}^{Mxm}$), in this section, two sensing matrixes Binary and Gaussian sensing matrix will be applied at different sensing ratios. At each ratio, the Sen, Spec and PPV are recorded as an average values for entire classes. The remaining plots in Figure 31 show the performance of L2-regularization on two sensing matrices under several compression ratios. In general, we can conclude that the achievement under a random Gaussian matrix outperforms the Binary sensing matrix under all performance parameters.


0.86
0.84
0.82
0.8
0.78
0.76
0.74
Accuracy ratio
0.1
0.15
0.25
0.35
0.45
Measurment ratio (M/m)
Gaussian matrix
Binary matrix

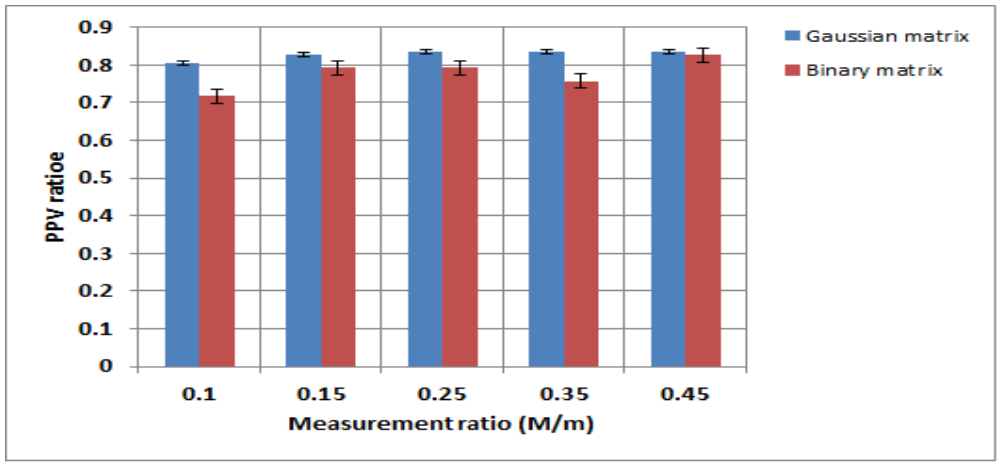
(a)


0.82
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
Senstivity ratio
0.1
0.15
0.25
0.35
0.45
Measurment ratio (M/m)
Gaussian matrix
Binary matrix

(b)



(c)



(d)

Figure 31. The sensing matrix effect on (a) the accuracy, (b) the sensitivity, (c) the specificity and (d) PPV

5.3. Recovery Results

The second goal of this study is the reconstruction of original data with high accuracy in relation to prior sensing data knowledge. This step needs to be a continuity of the recognition procedure. In other words, the features, which are used for recognition, will be used to reconstruct the original MS data by taking advantage of the MS sensing data that was used to recover the signal fingerprint. The data fingerprint is considered as sparse representation which has data preserved in a new space of MS. The L1-norm method using sparse difference (fingerprint) has been proposed in [6] with better performance than the regular sparse recovery (taking the entire data instead of the signal difference). Since the data fingerprint is not very accurate to be sparse, we deal with it as block sparse by dividing the signal difference into groups/blocks where just a few have non-zero elements. (We will take 10^{-3} as pruning threshold). For this purpose, the package BSBL [82] has been used. For evaluation purpose, the recovery error has been calculated as $\|x - x_R\|_2$ where x is the original data, and x_R is a reconstruction data. The BSBL-BO (the Bound-Optimization Method) [83] was applied to recover the data fingerprint. We defined the block partition; for simplicity, the block sizes are selected to be the same length $b_1 = b_2 = \cdots \ldots \ldots b_g$. To select the size of each block, this study changed the range of the size and estimated the Mean Square Error (MSE): defined as $\|x - x_R\|_2^2 / \|x\|_2^2$. The sensing matrix was a sparse binary matrix with a size of $(0.25 * N)$ which is the same for all experiments. In Figure 32 we can see that the resulting

quality of reconstruction data was almost the same for all range of blocks:

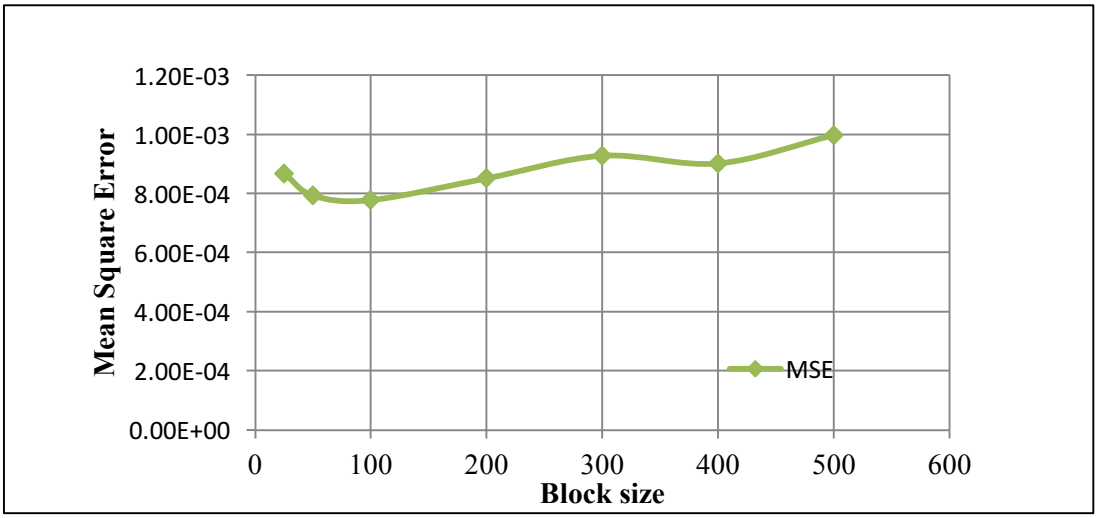


Figure 32. The effect of the block size regarding to MSE of recovery of MS data

Consequently, this study selected the block size to be $b_1 = b_2 = \cdots \ldots \ldots b_g = 100$, and the number of measurements ratio M/N ratio is from $0.01\ to\ 0.5$. The reconstruction performance has been compared with the spectral projected gradient (SPGL1) [84] [85]. Figure 33 shows that the BSBL-BO algorithm has a lower recovery error in all ranges of compressing ratios. In addition, the error rate is very small even when the compressing ratio is small, and that gives the BSBL-BO an advantage to recover the original data using a small number of high-quality measurements. Figure 28 provides an example of recovering data using L1- minimization and BSBL-BO using the same number of measurements and sensing matrix. However, in opposition to the L1-

minminization, the BSBL algorithms are time consuming (i.e the run time for BSBL is 1601.2s at the same conditions and it is 557.61s for the L1 algorithm.

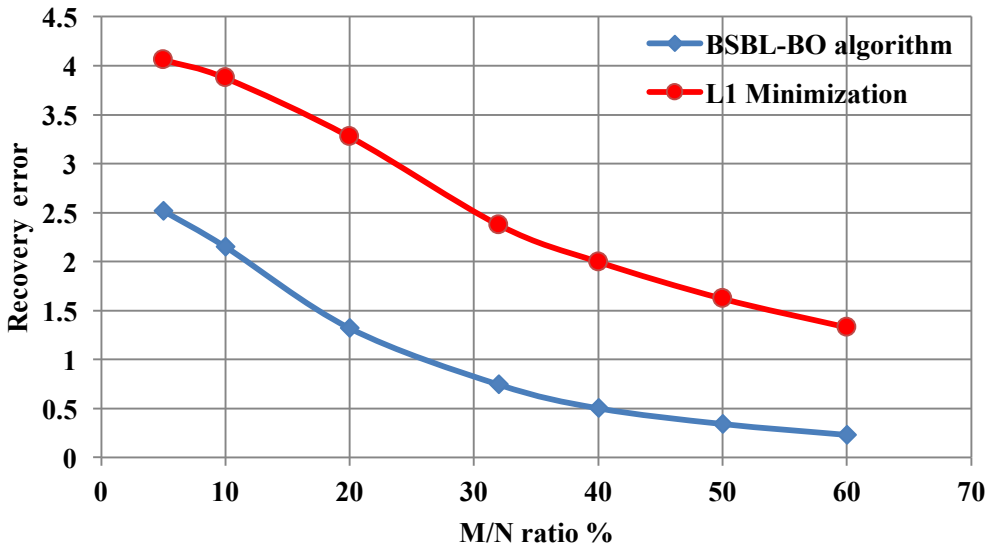
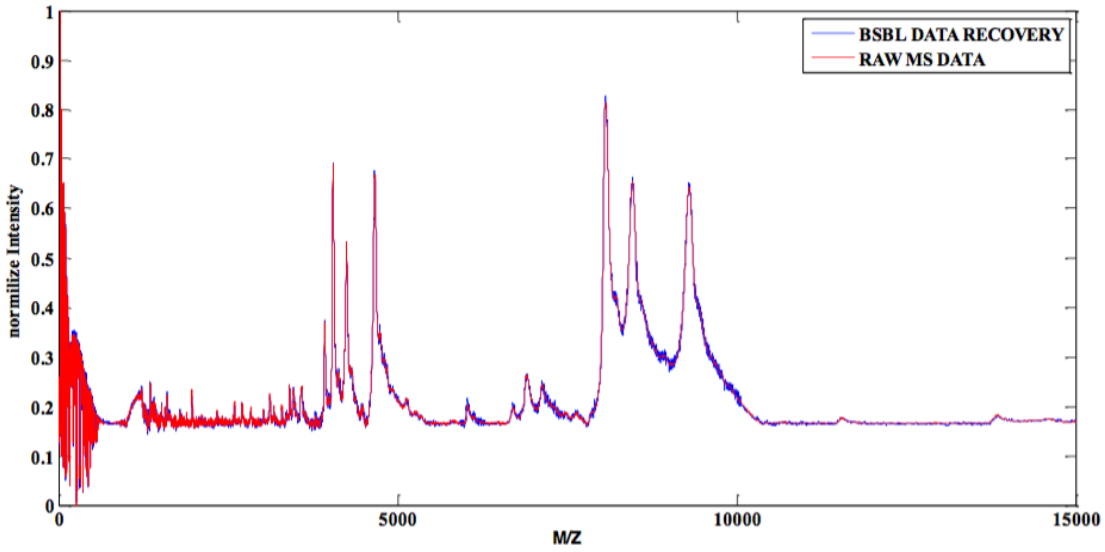
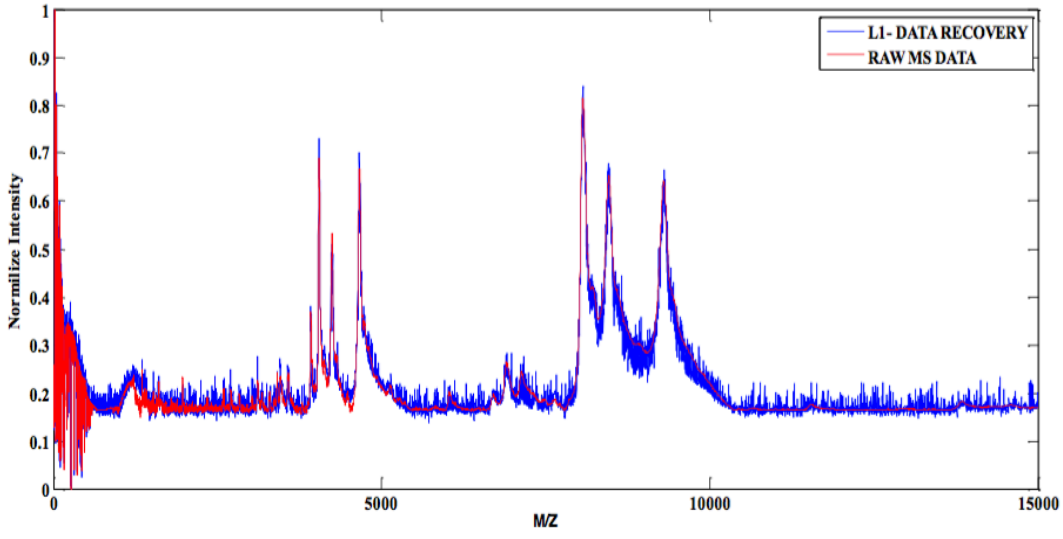


Figure 33. The average recovery error of L-minimization and BSBL-BO for the prostate cancer sample under different measurement rates. Each experiment rate has been repeated 10 times



(a)



(b)

Figure 34. An example of recovering an MS data sample using two scenarios: (a) BSBL-BO technique and (b) L1-minminization

CHAPTER VI

CONCLUSION AND FUTURE WORK

The high dimensionality of Mass Spectrometry (MS) data and noisy spectra are two of the main challenges facing the achievement of high accuracy recognition. The goal of this dissertation was to develop an accurate classification tool by employing compressive sensing (CS). Not only can CS significantly reduce MS data dimensionality, but also will allow for the full reconstruction of original data. The classification framework is capable of solving an overdetermined system but with a significant dimensionality reduction attribute and without any loss of accuracy. Classification is established using the reduced dimensional MS data and using a L2 norm and a mixed L2-L1-norms regularization terms.

To validate the proposed method, two cancer databases have been used (prostate, ovarian). The results show that L2-algorithm with regularization performed better than both the L1-algorithm and Q5 under all applicable assumptions. Regularization terms were used as design parameters and by selecting $0 \leq \eta \leq 1$, the algorithm resulted in an improved performance. Selecting the sensing matrix structure is also as vital for high performance, reconstruction, and computational complexity. In addition, we used signal difference to sparsify MS signals and implemented a reconstruction scheme for the identified disease signal from its low dimension feature space. Specifically, L1-minimization and BSBL algorithms were used to reconstruct MS data, and it was found

that BSBL outperformed L1. This multiple dissertation work, applied to proteomic MS data to accurately and more efficiently assess patient risk for prostate cancer or ovarian cancer constitutes a general framework for use in many other areas such the production a higher throughput, accurate detection of disease biomarkers and it may be utilized in the future for improved personalized medicine.

Recommendations for future work may be summarized as follows:

- An L2-algortithm with regularization can be used to identify the protein peptides by applying compressive sensing in a protein database. Since the peptides spectra have been reduced, the identification process will be faster and new spectra can thus be added to the database.
- By clustering the whole database, the Multiple Measurement Vector (MMV) algorithm can be applied. In addition, conducting a comparative study between the MMV and L2-algorithm with regularization will prove to be valuable.
- Further investigations are needed to consider MS samples of different lengths.
- Other classification techniques may be tested under the general proposed framework in this dissertation and of specific interest, a technique that will allow correlation between the samples to be utilized.

REFERENCES

[1] C. Dass, Fundamentals of contemporary mass spectrometry., vol. Vol. 16. , John Wiley & Sons,, 2007.

[2] Podwojski, K.& Fritsch, A.& Chamrad, D.& Paul, W.& Sitek, B.& Mutzel, P. & Rahnenführer, J., "Retention time alignment algorithms for LC/MS data must consider nonlinear shifts.," *Bioinformatics,* Vols. vol. 25, no. 6, p. 758–764, 2009.

[3] Alves, P., Arnold, R. J., Novotny, M. V., Radivojac, P., Reilly, J. P., & Tang, H., "Advancement in protein inference from shotgun proteomics using peptide detectability.," *Pacific Symposium on Biocomputing.,* vol. 12., pp. 409-420, 2007.

[4] Conrad T., Genzel M., Cvetkovic N., Wulkow N., Vybiral J., Kutyniok G. & Schütte C., "Sparse Proteomics Analysis-a compressed sensing-based approach for feature selection and classification of high-dimensional proteomics mass spectrometry data.," *eprint arXiv:1506.03620,* 2015.

[5] Taşkın, Vedat, Berat Doğan, and Tamer Ölmez, "Prostate Cancer Classification from Mass Spectrometry Data by Using Wavelet Analysis and Kernel Partial Least Squares Algorithm".

[6] Liu, Ji-xin, and Quan-Sen Sun., "Mass spectrum data processing based on compressed sensing recognition and sparse difference recovery," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, IEEE,2012.

[7] Liu, Ji-xin, and Quan-sen Sun., "Compressive sensing via sparse difference and fractal and entropy recognition for mass spectrometry sensing data.," *IET Signal Processing,* vol. 7, no. 3, pp. 201--209, 2013.

[8] E. Hoffmann, Mass spectrometry, Wiley Online Library,, 1996.

[9] Gross, Jürgen H., Mass spectrometry: a textbook, Springer Science \& Business Media, 2004.

[10] Zhang, Jianqiu and Gonzalez, Elias and Hestilow, Travis and Haskins, William and Huang, Yufei, "Review of peak detection algorithms in liquid-chromatography-mass spectrometry," *Current genomics,* vol. 10, 2009.

[11] Hernandez, Patricia, Markus Müller, and Ron D. Appel., "Automated protein identification by tandem mass spectrometry: issues and strategies," *Mass spectrometry reviews,* vol. 25.2, pp. 235--254, 2006.

[12] Mckee, Trudy and Mckee, James R, "Amino Acids and Proteins," in *Biochemistry: The Molecular Basis of Life*, OXFORD UNIVERSITY PRESS, 2011, pp. 1-60.

[13] Ma, Bin, "Challenges in computational analysis of mass spectrometry data for proteomics," *Journal of Computer Science and Technology,* vol. 25, no. 1, pp. 107--123, 2010.

[14] Standing, Kenneth G, "Peptide and protein de novo sequencing by mass spectrometry," *Current opinion in structural biology,* vol. 13, no. 5, pp. 595--601, 2003.

[15] Pan, Chongle and Park, Byung H and McDonald, William H and Carey, Patricia A and Banfield, Jillian F and VerBerkmoes, Nathan C and Hettich, Robert L and Samatova, Nagiza F, "A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry," *BMC bioinformatics,* vol. 11, no. 1, p. 118, 2010.

[16] adygov, Rovshan G and Cociorva, Daniel and Yates, John R, "Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book," *Nature methods,* vol. 1, no. 13, pp. 195--202, 2004.

[17] Serang, Oliver and Noble, William, "A review of statistical methods for protein identification using tandem mass spectrometry," *Statistics and its Interface,* vol. 5, no. 1, p. 3, 2012.

[18] Mann, Matthias and Hendrickson, Ronald C and Pandey, Akhilesh, "Analysis of proteins and proteomes by mass spectrometry," *Annual review of biochemistry,* vol. 70, no. 1, pp. 437--473, 2001.

[19] Mark A. Davenport, Marco F. Duarte, Yonina C. Eldar and Gitta Kutyniok, "Introduction to compressed sensing," *Preprint,* vol. 93, pp. 1-64, 2011.

[20] Berg, Ewout van den, "Convex optimization for generalized sparse recovery.," (Vancouver), 2009.

[21] Davenport, Mark A and Duarte, Marco F and Eldar, Yonina C and Kutyniok, Gitta, "Introduction to compressed sensing," *Preprint,* vol. 93, pp. 1-64, 2011.

[22] Mallat, Stephane, A Wavelet Tour of Signal Processing The Sparse Way, Academic press, 2009.

[23] Schulz, Adriana, Eduardo Antônio Barros Da Silva, and Luiz Velho., Compressive sensing, IMPA, 2009.

[24] Chen, Scott Shaobing and Donoho, David L and Saunders, Michael A, "Atomic decomposition by basis pursuit," *SIAM journal on scientific computing,* vol. 20, no. 1, pp. 33--61, 1998.

[25] Mallat, Stéphane G., and Zhifeng Zhang., "Matching pursuits with time-frequency dictionaries.," *Signal Processing, IEEE Transactions on,* vol. 41, no. 12, pp. 3397--3415, 1993.

[26] Rath, Gagan, and Arabinda Sahoo, "A comparative study of some greedy pursuit algorithms for sparse approximation," in *Signal Processing Conference, 2009 17th European. IEEE, 2009.*.

[27] Randall, Paige Alicia., "Sparse recovery via convex optimization," 2009.

[28] Duarte-Carvajalino, Julio Martin, and Guillermo Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *Image Processing, IEEE Transactions on,* vol. 18, no. 7, pp. 1395--1408, 2009.

[29] Köse, Kıvanç, "Signal and Image Processing Algorithms Using Interval Convex Programming and Sparsity," bilkent university, 2012.

[30] Kutyniok, Gitta, "Theory and applications of compressed sensing," *arXiv preprint*

arXiv:1203.3815, 2012.

[31] Baraniuk, Richard G, "Compressive sensing," *IEEE signal processing magazine,* vol. 24, no. 4, 2007.

[32] Stephen Boyd, Lieven Vandenberghe., Convex optimization., 7 ed., Cambridge university press, 2007.

[33] Duarte, Marco F., and Yonina C. Eldar., "Structured compressed sensing: From theory to applications," *Signal Processing, IEEE Transactions on,* vol. 59, no. 9, pp. 4053--4085, 2011.

[34] Candè, Emmanuel J., and Michael B. Wakin., "An introduction to compressive sampling.," *Signal Processing Magazine, IEEE,* vol. 25, no. 2, pp. 21--30, 2008.

[35] Foucart, Simon and Rauhut, Holger, A mathematical introduction to compressive sensing, Springer, 2013.

[36] Baraniuk, Richard, et al., "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation,* vol. 28, no. 3, pp. 253--263, 2008.

[37] Adamczak, Radoslaw, et al, "Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling," *Constructive Approximation,* vol. 34, no. 1, pp. 61--88, 2011.

[38] Grant, Michael, Stephen Boyd, and Yinyu Ye., "cvx users' guide," *2011-12-15 [2013-09-01]. http://cvxr, com/cvx/doc,* 2009.

[39] Candes, Emmanuel J., Justin K. Romberg, and Terence Tao., "Stable signal recovery from incomplete and inaccurate measurements.," *Communications on pure and applied mathematics,* vol. 59, no. 8, pp. 1207-1223., 2006.

[40] Duarte, Marco F., et al., "Recovery of compressible signals in unions of subspaces.," *Information Sciences and Systems,,* pp. 175-180, 2009.

[41] Lu, Hongtao and Long, Xianzhong and Lv, Jingyuan, "A fast algorithm for recovery of jointly sparse vectors based on the alternating direction methods," in *International*

Conference on Artificial Intelligence and Statistics, 2011.

[42] Davies, Mike E and Eldar, Yonina C, "Rank awareness in joint sparse recovery," *Information Theory, IEEE Transactions on,* vol. 58, no. 2, pp. 1135--1146, 2012.

[43] Deng, Wei and Yin, Wotao and Zhang, Yin, "Group sparse optimization by alternating direction method," in *SPIE Optical Engineering+ Applications*, International Society for Optics and Photonics, 2013, pp. 88580R--88580R.

[44] van den Berg, Ewout and Schmidt, Mark and Friedlander, Michael P and Murphy, Kevin, "Group sparsity via linear-time projection," *Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada,* 2008.

[45] Hong, Yan-jun, et al., "Discrimination analysis of mass spectrometry proteomics for ovarian cancer detection.," *Acta Pharmacologica Sinica,* vol. 29.10, pp. 1240-1246., 2008.

[46] Watson, Nathanial E., et al., "Classification of high-speed gas chromatography–mass spectrometry data by principal component analysis coupled with piecewise alignment and feature selection," *Journal of Chromatography A,* pp. 111-118, 2006.

[47] Adam, Bao-Ling, et al., "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men.," *Cancer research,* vol. 62, no. 13, pp. 3609--3614, 2002.

[48] Du, Pan, Warren A. Kibbe, and Simon M. Lin., "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics,* vol. 22, no. 17, pp. 2059--2065, 2006.

[49] Lilien, Ryan H., Hany Farid, and Bruce R. Donald, "Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum," *Journal of computational biology,* vol. 10.6, pp. 925-946., 2003.

[50] Zhang, Xuegong, et al, "Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data," *BMC bioinformatics 7.1,* p. 197, 2006.

[51] Guyon, Isabelle, et al., "Gene selection for cancer classification using support vector

machines," *Machine learning 46.1-3,* pp. 389-422..

[52] Cristianini, Nello, and John Shawe-Taylor., An introduction to support vector machines and other kernel-based learning methods, Cambridge university press, 2000.

[53] Karegowda, Asha Gowda, A. S. Manjunath, and M. A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection.," *International Journal of Information Technology and Knowledge Management,* vol. 2.2, pp. 271-277, 2010.

[54] Quinlan, J. R., San Mateo, C4.5 Programs for Machine Learning: Morgan KaufmANN., 1993, pp. 235-240.

[55] Quinlan, J. R., "Bagging, Boosting, and C4.5," *In Proceedings of the Thirteenth National Conference on Artificial Intelligence,* pp. 725--730, 1996.

[56] Wright, John, et al., "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on 31.2,* pp. 210-227, 2009.

[57] Ke, Jiqing, et al., "Sparse representation based feature selection for mass spectrometry data.," *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on,* pp. 57--62, 2010.

[58] Bartels, A., Dülk, P., Trede, D., Alexandrov, T., & Maaß, P., "Compressed sensing in imaging mass spectrometry.," *Inverse Problems,* vol. 29, no. 12, p. 125015, 2013.

[59] M. a. P. V. Golbabaee, "Joint trace/TV norm minimization: A new efficient approach for spectral compressive imaging.," in *Image Processing (ICIP), 2012 19th IEEE International Conference on.*, 2012.

[60] Duarte, Marco F., and Richard G. Baraniuk., "Kronecker compressive sensing.," *Image Processing, IEEE Transactions on,* vol. 21, no. 2, pp. 494--504, 2012.

[61] August, Yitzhak, et al., "Compressive hyperspectral imaging by random separable projections in both the spatial and the spectral domains," *Applied optics,* vol. 52, no. 10, pp. D46--D54, 2012.

[62] Liu, Ji-xin, and Quan-Sen Sun., "Mass spectrum data processing based on compressed sensing recognition and sparse difference recovery," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, IEEE,2012.

[63] Bruckstein, Alfred M and Donoho, David L and Elad, Michael, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review,* vol. 51, no. 1, pp. 34--81, 2009.

[64] Simonis, Joseph P., "Inexact Newton Methods Applied to Under--Determined Systems," Sandia National Laboratories, 2006.

[65] Zou H., and Trevor H., "Regularization and variable selection via the elastic net.," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 67.2, pp. 301-320., 2005.

[66] Petricoin, Emanuel F., et al., "Serum proteomic patterns for detection of prostate cancer.," *Journal of the National Cancer Institute 94.20,* pp. 1576-1578, 2002.

[67] Zhang, Zhilin, and Bhaskar D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning," *Selected Topics in Signal Processing, IEEE Journal of 5.5,* pp. 912-926, 2011.

[68] Zhang, Zhilin, et al., "Compressed sensing of EEG for wireless telemonitoring with low energy consumption and inexpensive hardware.," *Biomedical Engineering, IEEE Transactions on 60.1,* pp. 221-224., 2013.

[69] Zhang, Zhilin, et al., "Compressed sensing for energy-efficient wireless telemonitoring of noninvasive fetal ECG via block sparse Bayesian learning.," *Biomedical Engineering, IEEE Transactions on 60.2,* pp. 300-309, 2013.

[70] Mishali, Moshe and Eldar, Yonina C, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *Signal Processing, IEEE Transactions on,* vol. 57, no. 3, pp. 993--1009, 2009.

[71] Elhamifar, Ehsan, and René Vidal., "Sparse subspace clustering," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009.

[72] Elhamifar, Ehsan, and René Vidal, "Clustering disjoint subspaces via sparse representation," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010.

[73] Lu, Hongtao and Long, Xianzhong and Lv, Jingyuan, "A fast algorithm for recovery of jointly sparse vectors based on the alternating direction methods," in *International Conference on Artificial Intelligence and Statistics*, 2011.

[74] Elhamifar, Ehsan, and René Vidal., "Block-sparse recovery via convex optimization," *Signal Processing, IEEE Transactions on,* vol. 60, no. 8, pp. 4094--4107, 2012.

[75] Tipping, Michael E., "Sparse Bayesian learning and the relevance vector machine.," *The journal of machine learning research 1,* pp. 211-244, 2001.

[76] Wipf, David P., and Bhaskar D. Rao., "Sparse Bayesian learning for basis selection.," *Signal Processing, IEEE Transactions on 52.8,* pp. 2153-2164., 2004.

[77] Tipping, ACFME and Faul, A, "Analysis of sparse Bayesian learning," *Advances in neural information processing systems,* vol. 14, pp. 383--389, 2002.

[78] Wipf, David P and Rao, Bhaskar D and Nagarajan, Srikantan, "Latent variable Bayesian models for promoting sparsity," *Information Theory, IEEE Transactions on,* vol. 57, no. 9, pp. 6236--6255, 2011.

[79] Amini, Arash, and Farokh Marvasti, "Deterministic construction of binary, bipolar, and ternary compressed sensing matrices.," *Information Theory, IEEE Transactions on 57.4,* pp. 2360-2370, 2011.

[80] Kuncheva, Ludmila I and Jain, Lakhmi C, "Nearest neighbor classifier: simultaneous editing and feature selection," *Pattern Recognition Letters,* vol. 20, no. 11, pp. 1149--1156, 1999.

[81] Tsuda, Koji, "Subspace classifier in the Hilbert space," *Pattern Recognition Letters,* vol. 20, no. 5, pp. 513--519, 1999.

[82] Z. Zhang, "https://sites.google.com/site/researchbyzhang/bsbl," 2012. [Online].

[83] Z. a. B. R. Zhang, "Extension of SBL algorithms for the recovery of block sparse signals with intra-block correlation," *Signal Processing, IEEE Transactions on 61.8,* pp. 2009-2015, 2013.

[84] Friedlander, E. van den Berg and M. P., "Probing the Pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing,* vol. 31, pp. 890-912, 2008.

[85] Friedlander, E. van den Berg and M. P., "{SPGL1}: A solver for large-scale sparse reconstruction," June 2007. [Online].