



Hochschule Neubrandenburg
University of Applied Sciences

Fachbereich Landschaftswissenschaften und Geomatik

Faculty of Landscape Sciences and Geomatics

**Efficient sparse signal recovery of remote sensing
data: a classification method for hyperspectral image
data**

Supervisor: Prof. Dr. Gerd Teschke

Department of Mathematics, Geometry and Applied Computer Sciences and the rector of NB
University of applied Science

Co-supervisor: Hon-Prof. Dr. Erik Borg

German Air Space Center (DLR)

Morteza Abdipourchenarestansofla

A thesis submitted to the Faculty of Landscape Sciences and Geomatics, Neubrandenburg
University of Applied Science, in fulfilment of the requirements for the master degree in
Geodesy and Geoinformatics.

Urn:nbn:de:gbv:519 - thesis 2018 – 0846 - 1

Winter 2019

Declaration

I, Morteza Abdipourchenarestansofla, declare that this thesis is the outcome of my research studies and it is being submitted only for the degree of master in Geodesy and Geoinformatics at Neubrandenburg University of Applied Sciences.

Date and signatures

To those who are scrambling but not giving up

Acknowledgements

I would like to extend my intimate gratitude to my supervisor, Prof. Gerd Teschke. Prof. Teschke has probably been the best supervisor I could ask for. Thank you for being a great motivation for me to explore the new research ideas in data mining and analysis domain. He was always open to me to discuss my progress in the thesis and he has a high discipline character that makes it easy to get along with. I have learned many things from him that help me to make my Master studies as a worthwhile experience.

I would also like to thank you to Hon-Prof. Dr. Erik Borg as my co-supervisor. During the course called remote sensing, he motivates me to do my master thesis in remote sensing domain and turn the outcomes of the analytical algorithms to be a solution in real world problems. I appreciate to Dr. Sadegh Jamali and Hon-Prof. Dr. Borg as my examiners for taking time to examine my thesis and provide helpful remarks and encouraging comments.

Abstract

Nowadays the concern of finding an efficient algorithm that can answer some of the open questions in big data analysis and mining has been gradually arose. Such questions can be regarded by the question of representing the data in a meaningful way in which the most useful information highlighted. Therefore, the motivation of answering these questions encourage this thesis to develop a principle classification algorithm called Efficient sparse signal recovery for big data representation for a classification task. In this thesis, we develop a classification principle algorithm that is based on the sparse coding for the classification of given test pixel from a hyperspectral image. Hyperspectral imagery in remote sensing domain has the characteristic of big data in terms of velocity, verity and volume. This data is a set of non-homogenous system that expose the ill-posed problem. Thus, a robust and efficient algorithm must be developed to treat such data effectively. Sparse representation draws a great attention in hyperspectral image representation and analysis. Employing sparsity-based model involved two main problems. Firstly, the problem of the representation of an informative dictionary, and secondly the issue of implementing a proper optimization problem that can effectively solve the objective function. This thesis focuses on the latter aspect while the dictionary issue is also tackled by proposing a Geometric dictionary. There have been many algorithms for finding the optimized minimum of the well-known objective functionals “least square ” with l_1 -norm regularization parameter (in statistic is called Lasso and in sparse coding it is known as Basis Pursuit) that lead to the sparsity measurement. The minimization of such functionals have some barriers, such as being non-convex (non-smooth function). Hence, current algorithms such as greedy algorithms like Orthogonal Matching Pursuit (OMP) and even Iterative Reweighted Least Square (IRLS), and Basis Pursuit take many iteration and computation to convergent, which is not efficient for computing high dimensional dataset. Recently, an adequate numerical solution has been gradually built for addressing such optimization problem very effectively. This effective numerical solution is called Iterative Shrinkage algorithm motivated by classical Donoho-Johnston shrinkage method. Hence, we develop the so-called Iterative Shrinkage algorithm in three phases and apply the developed algorithm on four different classes of a hyperspectral image for the classification task. The first phase begins with implementing the soft shrinkage thresholding algorithm and follow this in the second phase of the development that we inject the steepest descent iteration which can effectively deal with the large coefficients and lead to the acceleration of the iterative soft shrinkage. Lastly we present an optimization function called Joint sparse measurement comprising of the two previous phases which can uniquely represent the relevant dictionary for the given test pixel. The experimental results indicate that the developed version of the shrinkage algorithm can effectively minimize the objective functional with a fast convergence in terms of iteration steps. In addition, the problem of the representation of an informative dictionary is solved by proposing a geometric dictionary inspired by the Singular Value Decomposition (SVD) that leads to a lower amount of atoms to be presented in each sub-dictionary. The resulting output from the classification of four given classes verifies the performance of our proposed efficient signal recovery algorithm.

Zusammenfassung

Heutzutage ist es ein Anliegen, einen effizienten Algorithmus zu finden, der einige der offenen Fragen in der Analyse großer Datensätze beantworten kann. Beispielsweise die Darstellung der Daten in einer sinnvollen Art und Weise in der die nützlichsten Informationen hervorgehoben wurden. Dieses Anliegen erbrachte den Ansatz dieser Arbeit, ein prinzipielles Klassifizierungspaket zu entwickeln für die Darstellung großer Datensätze für eine Klassifizierungsaufgabe. In dieser Arbeit wird ein Klassifizierungs-Algorithmus basierend auf der sparsamen Kodierung für die Klassifizierung eines gegebenen Testpixels aus einem hyperspektralen Bild entwickelt. Hyperspektrale Bilder im Fernerkundungsbereich haben die Charakteristik von Big Data in Bezug auf Geschwindigkeit, Richtigkeit und Volumen. Bei diesen Daten handelt es sich um ein nicht homogenes System, das das ungünstig gestellte Problem aufdeckt. Daher muss ein robuster und effizienter Algorithmus entwickelt werden, um solche Daten effektiv zu behandeln. Die spärliche Darstellung zieht große Aufmerksamkeit bei der Darstellung und Analyse von hyperspektralen Bildern auf sich. Der Einsatz der sparsamen Kodierung beinhaltet zwei Hauptaspekte. Einerseits das Problem der Darstellung eines informativen Wörterbuchs, andererseits das Problem der Suche nach einer geeigneten Optimierung zur Lösung des Optimierungsproblems. Der Fokus dieser Arbeit liegt auf dem zweiten Problem, während das erste Problem auch mit einem geometrischen Wörterbuch angegangen wird. Es gibt viele Algorithmen für die Optimierung des bekannten Problems der kleinsten Quadrate mit dem Regularisierungsterm der l_1 -Norm (in der Statistik, Lasso und in der spärlichen Codierung als "Basis Pursuit" bekannt), die zur spärlichen Messung führen. Die Minimierung einer solchen Funktion weist einige Barrieren auf, so dass sie nicht konvex sind. Vorgeschlagene Algorithmen wie Greedy-Algorithmen so wie Orthogonal Matching Pursuit (OMP), Iterative Reweighted Least Square (IRLS) und Basis Pursuit benötigen daher viele Iterationen und Berechnungen zum konvergieren, was für die Berechnung von hochdimensionalen Datenmengen nicht effizient ist. In letzter Zeit wurde schrittweise eine adäquate numerische Lösung entwickelt, um dieses Optimierungsproblem sehr effektiv anzugehen. Diese effektive numerische Lösung ist der iterative Shrinkage-Algorithmus, der durch die klassische Donoho-Johnston-Schrumpfungsmethode motiviert ist. Daher befasst sich diese Arbeit mit der Entwicklung des sogenannten iterativen Shrinkage-Algorithmus in drei Stufen. Dabei wird der entwickelte Algorithmus auf vier verschiedene Klassen eines hyperspektralen Bildes für die Klassifizierungsaufgabe angewendet. Im ersten Stadium beginnen wir mit der Implementierung des Soft-Shrinkage-Thresholding-Algorithmus. In der zweiten Stufe der Entwicklung führen wir die Iteration mit dem steilsten Abstieg ein, die effektiv mit den großen Koeffizienten umgehen kann und die iterative weiche Schrumpfung beschleunigt. Abschließend wird eine Optimierungsfunktion, bekannt als Joint-Sparse-Messung, vorgestellt, welche die beiden vorherigen Schritte umfasst, die das relevante Wörterbuch für das gegebene Testpixel eindeutig darstellen können. Die experimentellen Ergebnisse zeigen, dass die entwickelte Version des Shrinkage-Algorithmus das Optimierungsproblem mit einer schnellen Konvergenz effektiv minimieren kann. Zusätzlich wird das Problem der Darstellung eines informativen Wörterbuchs gelöst, indem ein geometrisches Wörterbuch vorgeschlagen wird, das von *Singular Value Decomposition* (SVD) inspiriert ist. Dies führt dazu, dass in jedem Teilwörterbuch weniger Atome

vorhanden sind. Die Ergebnisse der Klassifizierung von vier Klassen belegen die Leistung des vorgeschlagenen Optimierungsproblems.

Appendix of the efficient sparse signal recovery

Plots and graphs are presented in appendix. Appendix presents the output of the model and promotes some insights of the algorithm behind the proposed sparsity based algorithm. Furthermore, the visualization of the dictionary depicted in plots and graphs that are also available in the Appendix.

Contents

Contents	v
Chapter 1	1
1.1. Introduction.....	1
1.2. Problem statement and Motivation	3
1.3. Contributions.....	4
1.4. Summary of the chapters.....	5
Chapter 2	7
2.1. Background	7
2.2. Linear dependency	9
2.3. Sparse Approximation	10
2.4. Geometric View of Norms and Sparsity	13
2.5. Optimization problem.	15
2.5.1. Overdetermined system.....	15
2.5.2. Underdetermined System	16
2.5.3. Constrained optimization strategy	17
2.5.4. Steepest descent projection.	17
2.5.5. Proximity optimization strategy	19
2.5.6. Iterative soft shrinkage algorithm	20
2.7. The Quest for Dictionary	20
Chapter 3	23
3.1. Hyperspectral Imagery	23
3.2. Hyperspectral Image processing	24
3.3. Spectral Unmixing and Endmember Extraction	25
3.4. Dimensionality reduction for Hyperspectral Images (HSI).	27
3.5. Hyperspectral Imagery classification.....	31
3.6. Pixel-Wise Image Classification.....	31

Chapter 4	33
4.1. Efficient sparse signal recovery for Hyperspectral Imagery data classification	33
4.2. Classification Problem, a Prior-knowledge	33
4.3. Data Model and Classification Principle	33
4.4. Sparse Recovery Principle as a Classification Problem.....	34
4.4.1. l_1 Sparse recovery via Soft-Shrinkage Iteration.	35
4.4.2. l_1 Constrained Recovery via Projected Steepest Descent Iteration.	36
4.4.3. Joint Sparsity Measure Recovery using Projected Steepest Descent iteration.....	36
4.5. Condensation of the a-prior given dictionaries.	37
4.5.1. Geometric base dictionary construction.....	37
Chapter 5	42
5.1. Experimental design.....	42
5.2. Background and relevant work.	42
5.3. Experiments	44
5.3.1. Data Set Description	45
5.3.2. Experimental Design.....	47
5.4. Experimental Result.....	51
5.4.1. l_1 Sparse recovery via Soft-Shrinkage Iteration.	52
5.4.2. l_1 Constrained Recovery via Projected Steepest Descent Iteration.	53
5.4.3. Joint Sparsity Measure Recovery via Projected Steepest Descent iteration.	55
5.5. Result and Discussion	57
Chapter 6	58
6.1. Summary	58
6.2. Conclusion	59
6.3. Future Direction	60
Appendix:.....	61
Intuition and output of the proposed algorithm.....	61
References.....	66

Chapter 1

1.1. Introduction

Nowadays, the ongoing advancement in Remote Sensing technology provides dailies information of the Earth in a complex and huge manner. The proliferation of remote sensing data leads to a term called big data. In this digital era, the main focuses in both research and industry is on improving our ability to extract knowledge from large and complex collections of digital data. Hyperspectral data pose a challenge due to its high dimensionality. Hyperspectral imagery (HSI) data contains the more distinguishable information of the objects compare to multispectral imagery data. A hyperspectral image has higher spectral resolution than a multispectral image. Hyperspectral data due to the higher dimensionality and velocity are prone to be considered as big data. Therefore, finding an appropriate model, which can touch every point in data efficiently, is the heart of the problem in big data. Sparsity based model have been recently investigated for hyperspectral images classification and several improvements have been made in different aspect. Indeed, the simplicity and flexibility implementation of sparsity based model make a scalable algorithm for parallel processing specially for big data in distributed platforms. In sparse representation, pixels can sparsely be represented throughout liner transformation. The assumption of sparsity model is that the given test pixel can be represented by a linear combination of a scaler multiplication and its subspace, where each subspace is spanned by a few elements from a set of basis vectors. Sparsity based model has been applied in many applications, particularly for hyperspectral image processing, such as image compression, signal recovery, image classification, sparse unmixing (Huang A, Zhang H , Pižurica A., 2017; Iordache, M-D., Bioucas-Dias, J., Plaza, A., 2011; Chen C., Chen N., Peng J., 2016; Ülkü, i., Kizgut E., 2018). Sparsity is a very powerful prior for identification of the real signal out of the indirect corrupted/noisy signal measurement. When the goal is to find a close approximation of the real measurement, then one tries to recover the noisy signal by posing a penalty term called regularization frame. This keeps the approximation in a reasonable manner. This procedure is also called sparse representation and when the main objective is classification, one tries to find the closest feature vector (pixel) to the given feature vector (test pixel) which then by some meaning represent the corresponding class of given vector. The representation of the given pixel can be performed by sufficient linearly constrained optimization problems or proximity optimization strategy. Generally, transforming an image within the linear concept is based on a generative sparsity model introduced by (Olshausen, B.A., Field, D.J., 1997). It is built upon learning a dictionary D using a set of training feature dataset. The learning dictionary can be employed for sparse representing of the given signal/pixel. This type of sparsity typically mentioned in the literature as sparse representation (Razaviyayn, M., Tseng, H-W., Luo Z-Q., 2014). Using sparse representation a pixel $x \in R^n$ can be modeled as a linear combination of a set of vectors $\{d_1, d_2, d_3, \dots, d_m\}$ called atoms in dictionary. A sparsity based model is given by,

$$\min_{\alpha \in R^m} \|\alpha\|_0 \text{ Subject to } x = D\alpha(1.1).$$

Here, $D \in R^{n \times m}$ is an underdetermined ($n < m$) system in which n is the number of equation and m is the number of unknown. Due to the underdetermined nature of D , the linear system admits infinitely many solutions of which we are seeking for the one with the fewest known zero elements. Hence $\|\alpha\|_0 = \{i: \alpha_i \neq 0, i = 1, 2, \dots, m\}$. The atoms in the dictionary D are corresponding to the training set constructed by a prior-knowledge that can be explained in classification task as a supervise learning problem, and $\alpha \in R^m$ is the coefficient vector that scales the atoms to the corresponding direction as close as given test feature vector. In sparsity based model, the vectors in D are playing the main role in a better representation of the given image. It should be noted that sparsity based model in terms of dictionary has two main routes. In the first route, the model can learn from given training dataset presented in the dictionary, by this means, in every iteration of the algorithm the represented training set in the dictionary will be updated along with the coefficients until a reasonable choose of atoms acquired, i.e. convergence. This is called dictionary learning, and it is frequently an over-complete dictionary where the number of samples are higher than the number of dimension space. Second route is concerned about constructing the dictionary in prior to the objective function which is also called pre-define dictionary. Both approaches can be solved by linear programming or greedy pursuit algorithms such as Basis Pursuit (Basis) (Chen, S.S., Donoho, D.L., Saunders, M.A., 2001) or Orthogonal Matching Pursuit (OMP) (Pati, Y.C., Rezaiifar, R., Krishnaprasad PS, 1993). For sparsity, dictionary-learning approach many optimization methods have been proposed in cooperation with the aforementioned algorithms, such as Method of Optimal Direction (MOD), K-SVD, Stochastic Gradient Descent, Lagrange Dual Method, and Lasso can be mentioned. The MOD and K-SVD are sharing the same weaknesses, being efficient only for lower dimensional dataset (due to the cost of matrix inversion and computing pseudoinverse in higher dimensional case) and having the possibility of being stuck in local minimum. However, sparse coding can be also done with constructing a dictionary that has the most informative representation and then the focus in the problem of finding the optimal and sparsest solution is to find a fast and accurate optimization strategy with l_p -norm where $0 \leq p \leq 1$. The optimization problem can be treated by the Greedy Pursuit algorithms, and the other general optimization framework are inefficient and normally required too many iterations and computations to reach their destination (Elad, 2013). This is especially the case for higher dimensional problems like in hyperspectral image processing. In recent years a new family of numerical algorithms have been developed that can address the issues mentioned above (Elad, 2013). This family is called/named the Iterative-Shrinkage algorithms motivated by optimality condition. These methods can be also applied to a constrained optimization problem (sub-gradient in non-convex chose for p) to accelerate the convergence and take the global minimum. Moreover, the elements in dictionary can be orthogonal bases in which the redundant information represents the fundamental directions in each sub-space (i.e. full rank matrix), and require custom algorithms mentioned above to find the coefficients α . An image typically is represented by pixels, which are in a cell grids called entries, containing intensity value. The dataset associated with an image can be understood as an array of all pixels. Images that are captured by a normal camera can only cover the visible light that is the comfort zone of human eyes (Geladi, L.M.P., Grahn, H.F., Burger J.E.,

2007). The optical spectrum covers three channels/bands of electromagnetic waves that in digital camera called RGB. Nowadays there are a lot of cameras that can capture images behind the ability of human eyes. In remote sensing, sensors are mainly specified based on spectral and spatial resolution (e.g. spectral and hyperspectral). Spectral sensor conveys electromagnetic wavelength in different portion presented as bands. The only difference that makes the hyperspectral images more feasible is their ability in object discrimination, which has variety of applications such as precision agriculture, man-made and land cover classification, object detection and more generally earth observation and environmental modeling. Hyperspectral imaging in remote sensing is the technology of obtaining environmental information by imaging geographical location via airborne and space born platforms. Hyperspectral images typically acquire information in hundreds contiguous spectral bands ranging from infrared to ultraviolet spectrum. For example, the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) provides spectral radiance in 244 contiguous spectral bands with 10 nm spectral and 20 m spatial resolution in the range of $0.4 - 2.5\mu m$ (Chang, 2013). The image data from hyperspectral image is considered as three-dimensional data cube. Sparsity based models are great tools for processing and analyzing such big data and can tackle a significant amount of problems for satellite imagery. Such that, denoising, pixel-unmixing, classification, data fusion and even more possible potential that can be mentioned which depends on the designing of the sparsity based model.

1.2. Problem statement and Motivation

With recent advent of very high-spectral resolution, hyperspectral imagery contributes to the discovery of many material substances, which could not be discovered by multispectral imagery (Chang, 2013). This property of hyperspectral data attracts many applications in real world problem, such as land management, environmental modeling, geology, urban planning, agriculture, ecology and conservation, hazard mapping, and energy management. Therefore, there is a significant need rising up to deal with such a complex data, which have the property of big data such as, high volume, variety and velocity. Nowadays this data can be archived by increasing volume, from Petabytes to Exabyte, because of huge number of bands are taken by continuously using airborne/space borne sensor. Thus, hyperspectral image analysis is falling under big data characteristic in which hundreds of bands are taken by continuously using hyperspectral spectrometer (Anand, R., Veni, S., Aravinth, J., 2017). In addition, hyperspectral images are commonly associated with the pixel mixing problems (Dias, J.M.B., Plaza, A., Valls, G.C., Scheunders, P., Nasrabadi, N., Chanussot, J., 2013). Due to these special characteristics of hyperspectral images, they are not good for daily operations (decision-making). Hence, advanced and efficient algorithms must be developed to touch every information in an efficient manner and can operate faster (in a streaming manner) than traditional algorithms. Many learning algorithms have been proposed for hyperspectral image classification. Supervised and unsupervised classification methods of which supervised learning algorithms use a set of observation to train the machine and find the best separating hyperplane (logistic regression, support vector machine), and unsupervised learning algorithms use a clustering algorithm and is based on the proposed cluster

to classify the new given pixel. Nevertheless, processing such big data, especially in streaming applications for real world problems, needs fast and simpler algorithm that is scalable on distributed platform for parallel computing and perform well in terms of speed and accuracy. Therefore, sparsity based models proposed effective algorithms. It turns out that in sparse representation many coefficients are not needed (Qazi Sami ul Haq, et al, 2010) and can be reduced by restricting them via a regularization parameter $l_{0 \leq p \leq 1}$ to keep them small and set to zero that also lead to avoid overfitting. Hyperspectral data can be considered as a dynamic system in which “one can mathematically prove that for dynamic system, sparse controls can always stabilize the system, showing once gain the powerful machinery of sparse representations” (Fornasier, M., Peter S., 2015). Unlike conventional images with Hyperspectral resolution, hyperspectral images are limited by relatively lower spatial resolution. Therefore, the problem of unmixing arises and sparsity model proved as a good-based model for pixel unmixing which leads to the state of the art endmember extraction. Sparse encoding intrinsically does several tasks such as pixel unmixing, denoting and classification. Moreover, sparsity based model solved the problem of feature selection in variety of application for both regression and classification tasks (Yan, Hand., Yang, J., 2015; Yan, 2013). Hence, the mentioned advantages of sparse representation has motivated us to develop a classification principle in the context of sparsity that can solve the mentioned problems for hyperspectral images with the focus on classification task. In addition, we propose a geometric base dictionary that represent the training data in an informative manner.

1.3. Contributions

The contributions of the thesis are as follows:

We develop a classification principle for high dimensional spectral images called hyperspectral imagery in remote sensing domain. The general idea is to model a high spectral feature dimension pixel as a column vector, which is represented by some dictionary. The assumption is that, for different groups of pixels we have by a-prior knowledge different dictionaries available. The classification process results in sparse recovery algorithms, where the recovered sparse vector contains basic information for the membership to the one of the classes.

- We proposed a geometric base dictionary for sparse representation that has the ability to sparsify the recover vector at most and contributes to the performance of the proposed sparse signal recovery algorithm in this thesis.
- We start with implementing an iterative procedure called Iterative Shrinkage algorithm to solve the optimization problem in sparsity-based model specifically designed for classification task.
- We develop the Iterative Shrinkage algorithm by reformulating the unconstrained optimization problem to a constrained optimization problem, which also leads to the acceleration of the convergence using the steepest descent iteration. It is important to

mention that, mapping inverse problems can be formulated as a minimization problem that can be solved by forward backward or iterative shrinkage/thresholding in which non-smooth functions with sparsity constraints can be minimized effectively. Furthermore, the soft shrinkage operator cannot deal with the biased estimation of the large coefficients. Hence, we injecting a stepwise operator (steepest descent) on the approximation allows reducing the bias in practice. Inverse problems can equivalently be formulated as constrained/unconstrained minimization problems. Then, optimization theory gets involved to deal with these minimization problems (Engl, Heinz Werner, Hanke, Martin, Neubauer, A., 2000).

- Eventually, we have proposed a joint sparsity optimization problem which is comprised of the two previous steps and the ability to provide block sparsity measurement of the coefficient that leads to a unique way of identifying of that dictionary which is more relevant for representation of given pixel.

1.4. Summary of the chapters

The thesis organized in six chapter.

Chapter 2 discusses the mathematical concept and understanding of the sparsity based models. We walk through some relevant background of linear algebra and explain the sparse approximation and different norms. Following up this we cut a glimpse at the optimization problems, and next we move forward with two-optimization strategy that are used in this thesis.

Eventually, we discuss about the dictionary with its importance and about the two main approaches for the presentation of the dictionary for the dictionary of sparse coding.

In **chapter 3**, an introduction of hyperspectral images, their characteristics and application are given. We move forward with the common processing task for hyperspectral images such as pixel-unmixing, dimensionality reduction. This chapter is also concerned about the classification of problem for hyperspectral images and review some approaches such as sparse representation and machine learning algorithms to perform the classification task.

Chapter 4 presents the proposed efficient sparse signal recovery for sparse representation classification task. In this chapter, we introduce an efficient sparse signal recovery containing the developed version of iterative soft shrinkage algorithm via steepest descent. The proposed efficient sparse signal recovery has also the ability of finding the most relevant dictionary for the given test sample. Furthermore, we propose a geometric dictionary inspired by singular value decomposition concept that contributes to the performance of the proposed sparsity based algorithm in this work. The ultimate goal of our proposed sparsity based algorithm is to advance the data analysis and mining task for high dimensional data. In other words, the focus is on accelerating a principle sparse approximation while preserving the accuracy. Thus, the ultimate goal is a fast convergence.

One may acquire a high degree of accuracy with low number of iteration, which shows the power of the sparse representation among the most presented algorithms in real world problems such as image processing, video processing, and signal processing. Overall, we promise an algorithm that needs less iteration to minimize the objective function in sparse representation. We implement an Iterative soft shrinkage scheme from scratch, and then designed this sparsity-based model via injecting a steepest descent iteration to control the threshold with a parameter called step length. Eventually an optimization function proposed, comprising the scheme of iterative soft-shrinkage, steepest descent, and a unique property that advance the sparsity model in a block wise manner, which leads to the identification of the relevant sub-dictionary for the given test sample.

Chapter 5 contains the details of applying the proposed efficient sparse signal recovery for hyperspectral image classification. We apply the proposed algorithm on a hyperspectral scene from Indian Pines. This dataset is from AVIRIS sensor freely available in 200 spectral dimension. For the experimental design, we choose four classes including corn, grass-pasture, woods, and stone-steel-towers. After extracting the corresponding spectral signature of each class based on the ground truth, we randomly separate them to 70 percent training-set and 30 percent test-set. The total sample size is 2078 in 200 dimension. The training set used for the constructing the dictionary and the test set is used to check the performance of the designed efficient sparse signal recovery in terms of accuracy and computation time. The algorithms run in each step of its development on the given dictionary and test set. Dictionary presented in each step of the development of our scheme is in two form. One time it is present as an over complete dictionary (low rank matrix) with all dataset and another time with our proposed geometric dictionary. The performance of the proposed algorithm in this experimental design meets the promise of promoting a fast convergence and significant accuracy in the classification task. Furthermore, the ideal Geometric dictionary contributes in the general performance of our developed algorithm. The result shows a significant enhancement in the convergence of the optimization function after developing the implemented shrinkage function for the sparse representation classification. Moreover, the accuracy verifies the reliability of the developed scheme. The general result on the classification is as follows:

- 93 % accuracy in the first stage with 150 iterations.
- 93 % accuracy in the second stage with 120 iterations.
- 98 % accuracy with 90 iterations in the complete scheme (efficient sparse signal recovery) comprising of iterative soft-shrinkage, injected steepest descent, and block sparsity measurement operation.
- In addition, the result on geometric dictionary demonstrate much higher performance for the proposed efficient sparse signal recovery rather over complete dictionary.

Chapter 6 we provide a summary of the objective and the instruction of the whole procedure, and we discussed about the potential of the proposed schema and its future direction. Lastly, the conclusion is provided.

Chapter 2

2.1. Background

In this thesis the vector denoted by lowercase letter, e.g. x while matrices denoted by uppercase letter, e.g. D and their elements are presented with indexes such as D_i . Assume a dimension of a sample being R^n where n is the dimension space and thus all the samples are concatenated in a form of matrix called dictionary $D \in R^{n \times m}$ where m is the number of sample. Suppose the number of feature dimension is less than the future vector (samples) ($n < m$), then the dictionary D referred as an over-complete dictionary which is refer to an under-determined system, since number of unknown is less than number of equation. Sparsity of a vector means that some elements of a vector are zero. By using the linear combination of a basis matrix D we can represent the given feature vector $x \in R^{n \times 1}$. Such that can be given as;

$$x = D\alpha \quad (2.1)$$

Where α denotes the coefficient vector that scales the atoms until finding the corresponding span for the given test sample with condition of if only $k \ll m$ elements of α are nonzero and the rest are zero. Then we call this k -sparse solution for given signal x .

Recall the equation (2.1), the inner product of two vectors, $v \in R^n$ and $v \in R^n$ can be computed as (2.2), and the inner product of two matrixes $U \in R^{n \times m}$, and $V \in R^{n \times m}$ can be given by (2.3).

$$\langle v, v \rangle = v^T v = v_1 v_1 + v_2 v_2 + \dots + v_n v_n \quad (2.2)$$

$$\langle U, V \rangle = \text{tr}(U^T V) = \sum_{i=1}^m \sum_{j=1}^n X_{ij} Y_{ij} \quad (2.3)$$

Where the operator $\text{tr}(A)$ denotes the sum of diagonal entries of the matrix A , that is called the trace of matrix A .

Norm of vector v (2.4) can be represented in n dimensional feature vector in Euclidian space (2.5).

$$v = [v_1, v_2, v_3, \dots, v_n,] \quad v \in R^n \quad (2.4)$$

Thus

$$\|v\|_p = (\sum_{i=1}^n |v_i|^p)^{1/p} \quad (2.5)$$

Which is the p -norm or the l_p -norm ($1 \leq p \leq \infty$) of vector v . Furthermore, p can be represented by 1 which is called l_1 -norm, which is the sum of absolute values of elements in vector v . Moreover the l_p -norm of a vector can be restricted by $p = 2$ that is Euclidian norm and represented as l_2 -norm (2.6). Figure 2.1 represent the different types of l_p -norms in 2D.

$$\|v\|_2 = \sqrt{(v_1^2 + v_2^2 + \dots + v_n^2)} \quad (2.6).$$

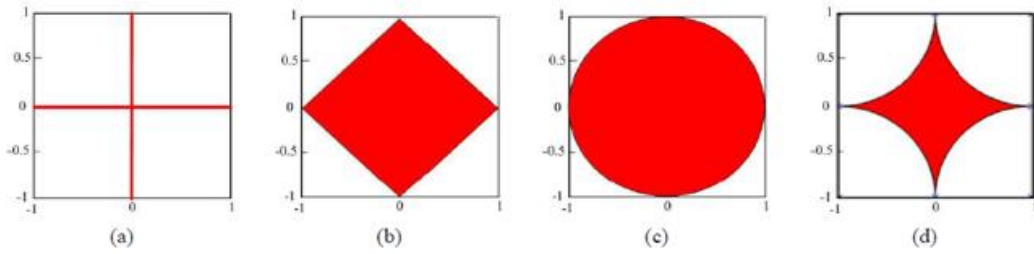


Figure 2.1. Geometric interpretation of different norms in 2-D (Zhang Z., Xu Y., Yang J., Li X., Zhang D., 2016). (a), (b), (c), (d) are the unit ball of l_0 -norm, l_1 -norm, l_2 -norm, and $l_{0 < p < 1}$ -norm in 2D space respectively.

The sparsity of a vector v is normally represented as $\|v\|_0$. This notation is regard to the number of nonzero element of vector v that is given by (2.7) (Bruckstein, A.M., Donoho D. L., and Elad M., 2009).

$$\|v\|_0 = \lim_{p \rightarrow 0} \|v\|_p^p = \lim_{p \rightarrow 0} \sum_{i=1}^n |v_i|^2 \quad (2.7)$$

As shown in (2.7) the notation intuitively stands for sparse representation problem. The relation between various form of l_p -norm can be found in figure 2.2 in which represents the shape of the function $|\alpha|^p$ with various value of p . Indeed, the summation of all nonzero interies is accrued by count of the nonzero location entries of vector v . The property of the l_p -norms can be assessed in terms of smoothness and convexity. As shwon in figure 2.2 basically l_p -norm ($0 < p < 1$) function is nonconvex, nonsmooth, and global nondifferentioable function. In contrast the l_1 -norm is convex, nonsmooth, global nondifferentiable function, and l_2 -norm is convex, smooth, global differentiable function (Zhang Z., Xu Y., Yang J., Li X., Zhang D., 2016).

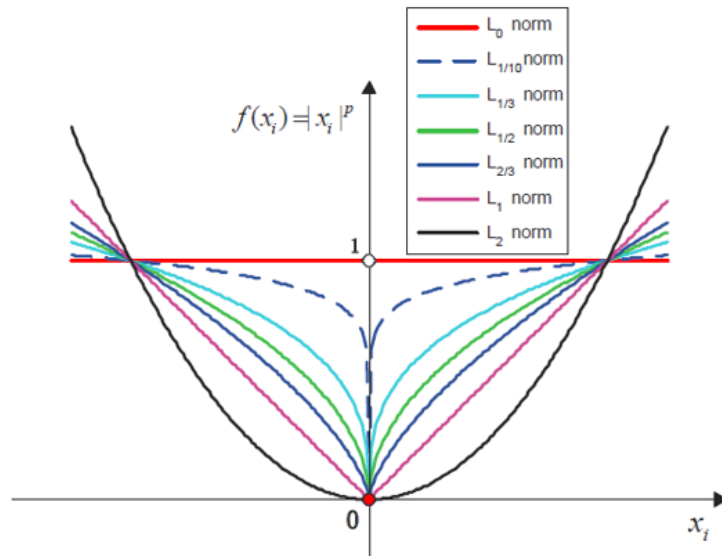


Figure 2.2. The behavior of $|\alpha|^p$ for $p = \{0, \dots, 2\}$. As p tends to zero, $|\alpha|^p$ approaches the indicator function, which is 0 for $\alpha = 0$, and 1 for $\alpha \neq 0$.

2.2. Linear dependency

A concrete example of the linear dependency and dimensionality reduction can be given as follows, Consider \vec{v} and \vec{u} numerically given by (I) as 2D vectors, the goal is to find a way to reduce the number of vector. This idea comes from linear system of equation that takes the advantages of linear dependency concepts according to the basis vector in linear algebra.

$$\vec{v} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \vec{u} = \begin{pmatrix} 4 \\ 8 \end{pmatrix} \quad (I)$$

Based on the linear combination concept when a vector is in span of the other (basis vector) then we can factorize that vector by finding the scalar multiplication that sufficiently satisfy the equality concept.

Hence, solving (I) is as follows:

$$\lambda \vec{v} = \vec{u} \quad (3.3) \lambda = 2$$

Therefore 2 is a scalar multiplication of vector v that can expand the \vec{v} to \vec{u} . Indeed, we are able to get rid of redundant direction/dimension by factorizing vector \vec{v} and eventually address the problem of dimensionality. This concept called linear dependency.

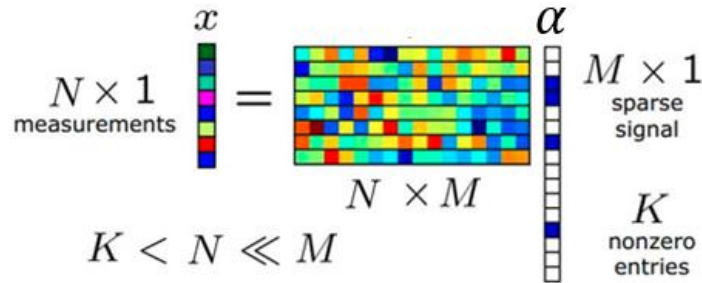
2.3. Sparse Approximation

Sparse approximation also called sparse representation (SR) is a mathematical concept for sparse solutions of a linear system of equation. In mathematics, a linear system of equation comprises a set of linear equations that have the same variables. Sparse approximation has gained much attention in image processing, signal processing and machine learning. Sparse representation is inspired by compressed sensing (CS) (Donoho D. L., 2006). CS theory suggests that if a signal is sparse or compressive the original signal can be recovered by a few measurements, which are remarkably less than suggested methods such as Shannon's sampling theorem (Zhang Z., Xu Y., Yang J., Li X., Zhang D., 2016). Sparse representation has many applications in image processing such as image denoising, deblurring, compression, super resolution, and image classification (Baraniuk, R.G., Candes, E., Elad, M., and Ma, Y., 2010). The assumption in SR is that, the unknown pixel or signal of interest is modeled as a sparse combination of few atoms represented in given dictionary and the approximation is controlled by a regularization term, which is the energy (norm) of the function. Sparsity is a very powerful prior for identification of the real signal out of the indirect measurement corrupted/noisy signal. When the goal is to find a close approximation of the real measurement of the given pixel then we are trying to recover the real signal approximately based on the given noisy signal and a regularization frame that keeps the approximation in a reasonable manner. This procedure is also called sparse representation and when the main objective is classification one tries to find the closest feature vectors (groups) to the given feature vector (test pixel) which then by some meaning represent the corresponding class of given vector. The representation of the given pixel can be performed by several approaches such as linearly constrained optimizations, and proximity optimization problem. Generally, transforming an image within the linear concept is based on a generative sparsity model introduced by (Olshausen, B.A., Field, D.J., 1997). It is based on learning a dictionary D using a set of training feature datasets. The learned dictionary can be employed for sparse representing of the given signal/pixel. This type of sparsity is typically mentioned in the literature as sparse representation (Razaviyayn, M., Tseng, H-W., Luo Z-Q., 2014). Using sparse representation an image x can be modeled as a linear superposition of a set of vectors $\{d_1, d_2, d_3, \dots, d_m\}$ called atoms in dictionary D , given by:

$$x = D\alpha, (2.1)$$

Where the atoms in dictionary D are corresponding to the training set constructed by a priori-knowledge that can be explained as supervised learning, and $\alpha \in R^m$ is the coefficient vector that scales the atoms (columns in the dictionary) to the corresponding direction as close as given test feature vector. The construction of this dictionary is an active field of research that scientists and engineers are dealing with. Designing of the dictionary has an effect in both accuracy and computational time complexity. Such that choosing the dictionary that sparsifies the signals can be done via two approaches (i) dictionary learning approach that is based on some mathematical model and (ii) building a sparsifying dictionary which is based on the mathematical structure of the data (Rubinstein, R., Bruckstein, A.M., Elad, M., 2010). Sparse representation establishes a meticulous mathematical framework to study high dimensional data and ways to decode the

structure of the data in a sufficient manner (Baraniuk, R.G., Candes, E., Elad, M., and Ma, Y., 2010). The simple representation and good scalability of the sparse representation is one of the best advantage of this algorithm that can be reliably implemented on distributed and parallel computing platform. The sparsity algorithm based on the presented linear system of equations (1.1) can be explained in this way that $D \in R^{n \times m}$ where n is the number of equation (feature dimension) and m is the number of unknown (sample dataset) is undetermined since the number of unknowns is less than the number of equations ($n < m$). Therefore, due to the underdetermine nature of D the linear system admits infinitely many solutions α in which we seek for the one with fewest nonzero (2.2) elements that satisfy $x = D\alpha$ condition.



$$\min_{\alpha} \|\alpha\|_0 \text{ Subject to } x = D\alpha \quad (2.2)$$

Where $\min_{\alpha} \|\alpha\|_0 = \{i: \alpha_i \neq 0, i = 1, 2, \dots, m\}$ is the l_0 pseudo-norm which counts the number of non-zero entry of coefficient vector α . This property is well known as NP-Hard, which is an exhaustive search for finding the minimum of the given function. Ultimately, sparse approximation/representation implies that only a few elements with non-zero entry are able to approximate the solution such that (2.3)

$$\alpha_k \neq \alpha_i \ll n < m. \quad (2.3)$$

This motivation allows us to decode the given x by a combination of a few atoms in dictionary that span the space to find the given vector. Since this problem is NP-Hard (Amaldi, E., and Kann, V., 1998), the solution can be found in an approximation manner using l_1 (2.4), such that using a convex relaxation of the problem, obtained by employing l_1 -norm instead of l_0 where $\|\alpha\|_1$ simply sums the absolute values of nonzero entries of α .

$$\min_{\alpha} \|\alpha\|_1 \text{ Subject to, } x \approx D\alpha \quad (2.4)$$

There have been many algorithms to solve the problem in 2.4. Indeed, one needs to clarify, which algorithm is the proper method for the posed problem. The main component of the sparsity-based models is the dictionary. The dictionary is the collection of training set which acquired by a given dataset. Dictionary can be constructed in various ways. There are tons of literature about dictionary learning or construction of a dictionary in prior to the objective function (Hao S., Wang, W.,

Bruzzone, L., 2017; Liu W., Wen, Y., Li, H., Zhu, B., 2014). Furthermore, the optimization problem is an important aspect that should be considered. Various optimization algorithms developed for solving the problem of sparse approximation. Sparse representation theory can be categorized from different point of views. Since different method, have their particular motivations there have been different prospective for categorization. For instance, in terms of atoms the available sparsity based models can be divided in two groups (i) dictionary construction base model, (ii) dictionary learning based method. Based on the literature (H. Cheng., Z. Liu., L. Yang., and X. Chen, 2013) sparse representation algorithms considered in three classes, (i) convex relaxation, greedy algorithms, and combinational methods. In addition, sparse representation in terms of optimization are consider in four optimization problems, (i) the smooth convex problem, (i) non-smooth convex problem (ii) smooth non-convex problem, and (ii) non-smooth non-convex problem (J. A. Tropp, A. C. Gilbert, and M. J. Strauss, 2006; Tropp, 2006). In addition, a review paper by (Zhang Z., Xu Y., Yang J., Li X., Zhang D., 2016) categorized the available sparsity based algorithms with respect to the analytical solution and optimization viewpoints into four groups. (i) The greedy strategy approximation, (ii) constrained optimization strategy, (iii) proximity algorithm based optimization strategy, and homotopy algorithm based sparse representation. One of the famous algorithm for solving the problem in 2.4 is known as the Basis Pursuit (BP) algorithm (2.5) (Gill, P.R., Wang A., Molnar, A., 2010)

$$\min_x \frac{1}{2} \|x - D\alpha\|_1 + \lambda \|x\|_1, \quad (2.5)$$

This is an instance of convex optimization, which is the least square solution with a penalty term. λ , denotes the parameter that controls the trade-off between sparsity and reconstruction fidelity also called regularization parameter and the rest are as before. The problem of basis pursuit can be handle using linear programming solver or alternatively using the approximation method such as matching pursuit (MP). MP is a greedy technique that finds none zeros locations of the coefficients one at the time. The sparse representation problem can be solved perfectly under the mild conditions via BP and MP that guaranty the unique solution (Donoho D. , 2006). Nevertheless, in the noisy case where x associated with some noise the solution is approximated via (2.5). Indeed, the best projection of multi-dimensional data into the span of a dictionary, which has special properties, can be approximated by BP Denoting, and similarly via matching pursuit. Constrained optimization strategy motivated from the idea of finding a suitable way to transfer a non-differentiable optimization problem to a differentiable optimization problem by replacing l_1 -norm penalty term by an equal constraint condition in a minimization problem. Indeed, by constrained optimization problem we make the minimization problem feasible by solve the problem of being convex but non-smooth function. The proximal algorithms can be efficiently represented as a powerful algorithm for solving constrained, non-smooth, large scale, or distributed version of optimization problem (Parikh, N., and Boyd S. , 2013). The main objective of proximal algorithm based optimization is to separate the objective function into two-piece. Meaning that the optimization function can be separated by removing the regularization term and solve the problem like convex function, such Iterative Shrinkage thresholding algorithm.

2.4. Geometric View of Norms and Sparsity

This section gives a summary on sparse representation and based sparsity based model into different categories in terms of norm.

As discussed in 2.1 an over-complete dictionary has infinitely many solutions in which the sparse representation seeks for the k -sparse solution (non-zero elements) (2.2). Let assume $D = [d_1, d_2, \dots, d_m] \in \mathbb{R}^{n \times m}$ where n and m denote the number of equations (feature dimension) and the number of unknown (sample) respectively where ($n < m$). Matrix D is the basis dictionary that constructed by the measurement data called over-complete dictionary. Each column of D is a sample that is called atom and the test feature dataset can be given by $x \in \mathbb{R}^n$. Let us generally assume we want to approximate the give test sample using all of the unknowns. Thus we can represent it as (2.6),

$$x = d_1\alpha_1 + d_2\alpha_2 + \dots + d_m\alpha_m \quad (2.6)$$

Can be written as $x = D\alpha$ (2.6)

In which $\alpha_i \in \mathbb{R}^m$ represents the coefficients associated with their sample. The given problem is an ill-posed problem if there is not any prior knowledge or constrained to the solution of α . Indeed, there is not exist a unique solution to the (2.6) that can present z . Thus, a regularization parameter needs to control the parameter α to be restricted by a bindery which is the concept of ℓ_p -norm which we discussed its principle at the beginning of this chapter. Assume a 2D vector for $y = D\alpha$ where $y \in \mathbb{R}^2$ then figure 2.3 gives an intuition of ℓ_1 -norm and ℓ_2 -norm in which ℓ_1 promote the solution whiten the intersection in horizontal axes and, thus the solution for two entries in the other axes will be zero. In contrast with ℓ_1 , ℓ_2 -norm is nor promoting sparse solution, since the equation $y = D\alpha$ intersects in two points within the ℓ_2 -norm (circle) shape. Therefore, two entries are non-zero (figure).

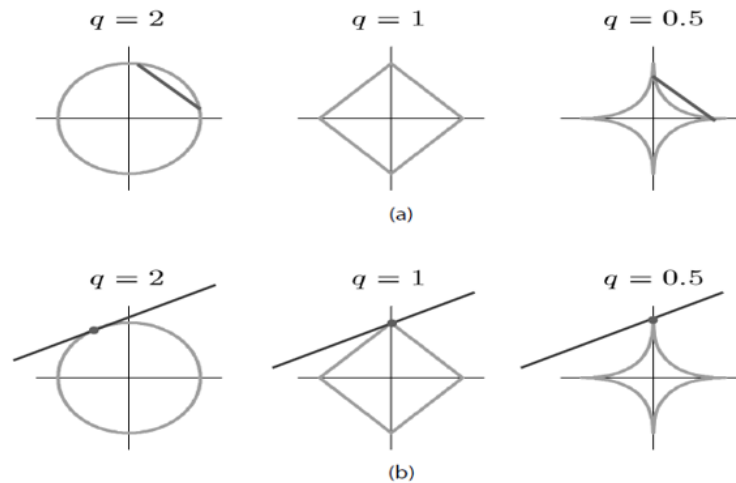


Figure 2.3. Depict the sparsity level in different l_p -nom. (a) Level sets $\|\alpha\|_q^q = 1$ for several values of q . (b) Optimization of (P_q) as inflation of the original-centered l_p -balls until they meet the set of feasible points as $D\alpha = x$ (Rish, I., Grabarnik, G., 2014).

Regarding the difficulty of solving the under-determined system of equations, one can relax the equation (2.6) via imposing a penalty term (one choose of p for l_p -norm). Depending on the choice of p we can sparsify the solution of the coefficient α . Furthermore, the real data are assuming to be associated with noise that affects the approximation. Therefore, the original model modified to the

$$x = D\alpha + \varepsilon \quad (2.7)$$

Where $\varepsilon \in R^n$ refer to the presentation noise in each dimension. Ultimately, the problem can be approximately obtained by minimizing the least square solution (2.8).

$$\hat{\alpha} = \arg \min \|\alpha\|_0 \quad \text{s.t} \quad \|y - D\alpha\|_2^2 \leq \varepsilon \quad (2.8)$$

This minimization problem (2.8) can be solved via various approaches. Indeed the question of using which optimization strategy guaranties the convergence to local or global minimum arises. Depending on the application, dataset and the posed problem the choice of optimization problem must be selected. Such optimization problems to solve the minimization problem for (2.8) can be mentioned as Lagrange multiplier, linear programming, quadratic programming, and convex optimization. The equation (2.9) is the Lagrange multiplier that introduce as a constrained optimization along with for l_0 -norm to solve (2.8).

$$L(\alpha, \lambda) = \arg \min \|x - D\alpha\|_2^2 + \lambda \|\alpha\|_0, \quad (2.9)$$

Since this problem with l_0 -norm is NP-hard, we used l_1 -norm. The origin of the l_1 -norm is Lasso problem (Tibshirani, 1996; R., 2011). l_1 -norm has been used in many application such machine learning, computer vision (Patel V. M., and Chellappa, R., 2014) etc. Therefor the problem in (2.9) can be approximated via l_1 -norm (2.10).

$$L(\alpha, \lambda) = \arg \min \|x - D\alpha\|_2^2 + \lambda \|\alpha\|_1, \quad (2.10)$$

Moreover, the problem in 2.9 can also be slaved by l_2 -norm (2.11).

$$L(\alpha, \lambda) = \arg \min \|x - D\alpha\|_2^2 + \lambda \|\alpha\|_2, \quad (2.11)$$

The problem in (2.10) is a convex but no differentiable. That can be solved by proximity optimization problem. In addition, this problem can be convert to a constrained strategy using by indicating a stepwise direction for the derivation. It should be mentioned that the problem in (2.11) is not prone to give a sparse solution.

2.5. Optimization problem.

In this section, some of the optimization function will be defined. Furthermore, two main optimization functions that used in this thesis will be explained. Further we discussed the under determined and over determined system of linear equations.

Optimization functions aims to minimize or maximize the objective function. Let us assume a cost function such as least square solution (2.12) for underdetermined and/or overdetermined system of equations. Given a problem of system of equations (2.6),

$$x = D\alpha \quad (2.6)$$

Then the cost function will be the output value in least square solution called residual given by,

$$J(\alpha) = r(\alpha) = \|x - D\alpha\|_2^2 \quad (2.12).$$

Thus, the optimization function also called objective function is given by two terms called, the cost function and regularization parameter, which is a weighted sum of least square solution (2.9). The goal in optimization function is to minimize the objective function respect to the coefficient vector α and the constrained λ .

$$\hat{\alpha} = L(\alpha, \lambda) = \arg \min \|x - D\alpha\|_2^2 + \lambda \|\alpha\|_0, \quad (2.9)$$

This problem can be solve based on its property in many ways. Indeed, choosing a specific number of P for l_p -norm requires a specific algorithm to minimize the objective function (2.9). Since we would like to have the sparsest solution, l_1 -norm provides sparse solutions rather than l_2 -norm (Schmidt, 2005). Although, the choice of l_1 -norm is a reasonable choice but finding the best minimization strategy for such problem (2.10) is challenging. Indeed, due to the property of l_1 -norm, the function become a non-differentiable that needs efficient optimization strategies. Therefore, the problem can be solve via different approaches, such that, proximity optimization strategy, and constrained optimization strategy can be mentioned (Zhang Z., Xu Y., Yang J., Li X., Zhang D., 2016). In a more general case of least square solution (cost function), the problem is convex. Hence, normal equation is the close form solution. First, let us consider two cases of linear system (underdetermined and overdetermined), and then look behind the minimization of the (2.12) for both system, and eventually solve the problem of optimization in (2.10).

2.5.1. Overdetermined system

Consider the linear system $x = D\alpha$ where there is no solution to this system in the case where D has more rows (equations) than columns (unknowns) where column are linearly independent. Therefore, this system called overdetermined system. One may seek for the solution by finding the coefficient vector α that minimizes the least square solution. In other words, the solutions that minimize the energy of the error (2.12), which also called cost function.

$$J(\alpha) = \min_{\alpha} \|x - D\alpha\|_2^2, \quad (2.12)$$

Expanding $J(\alpha)$ gives (2.13),

$$J(\alpha) = (x - D\alpha)^T(x - D\alpha)$$

$$x^T x - x^T D\alpha - D^T \alpha^T x + D^T \alpha^T D\alpha$$

$$\text{Since } D^T \alpha^T x = x^T D\alpha$$

$$\text{Then } x^T x - x^T D\alpha - x^T D\alpha + D^T \alpha^T D\alpha$$

$$x^T x - 2x^T D\alpha + D^T \alpha^T D\alpha \quad (2.13)$$

Eventually by taking the derivative of (2.13) we will get (2.14)

$$\frac{\partial}{\partial \alpha} J(\alpha) = -2D^T x + 2D^T D\alpha = 0$$

$$D^T D\alpha = D^T x \quad (2.14)$$

Now assume the D is invertible then the solution of (2.12) using norm equation can be analytically given by (2.15)

$$\alpha = (DD^T)^{-1}D^T x \quad (2.15)$$

2.5.2. Underdetermined System

Consider the linear system $x = D\alpha$ that the matrix D has less rows (equations) then columns (unknowns) in which, the rows are linearly independent, then this system has infinitely many solutions. This system called underdetermined system. In this case, the common procedure is to find a solution x with minimum norm. Which is solving for an optimization problem given by

$$\min_{\alpha} \|\alpha\|_2^2 \quad \text{Subject to, } x = D\alpha \quad (2.17).$$

In this case, the minimization preformed via Lagrange multipliers (2.18)

$$L(\alpha, L) = \|\alpha\|_2^2 + L^T(x - D\alpha) \quad (2.18)$$

Therefore, the derivation of Lagrange given by (2.19 and 2.20).

$$\frac{\partial}{\partial \alpha} L(\alpha) = 2\alpha - D^T L \quad (2.19)$$

$$\frac{\partial}{\partial L} L(L) = x - D\alpha \quad (2.20)$$

Set the derivations to zero we get (2.21 and 2.22).

$$\alpha = \frac{1}{2} D^T L \quad (2.21).$$

$$x = D\alpha \quad (2.22).$$

Simply plugging $\alpha(2.21)$, into (2.22) we get,

$$x = \frac{1}{2}DD^T L(2.23).$$

Now let assume DD^T is invertible, then the solution of Lagrange multiplier is given by,

$$L = 2(DD^T)^{-1}y \quad (2.24)$$

Eventually plugging (2.24) give the solution of (2.24) in (2.21) and then we get,

$$\alpha = D^T(DD^T)^{-1}y \quad (2.25)$$

Now, it is possible to verify that the solution α can satisfy the equation $x = D\alpha$ by plugging in,

$$D\alpha = D[D^T(DD^T)^{-1}y] = DD^T(DD^T)^{-1}y = y \quad (2.26)$$

Therefore, the approximation of solution for $\min_{\alpha} \|\alpha\|_2^2$ s.t $x = D\alpha$ given by,

$$\alpha = D^T(DD^T)^{-1}y \quad (2.25).$$

A common approach to approximate a linear system of equations is to minimize the objective function. So far the problem of minimization of objective function with the penalty term (2.9) depends on the choice of p-norm varies in terms of finding the best solution. Recall the problem in (2.10).

$$J(\alpha) = \min_{\alpha} \|x - D\alpha\|_2^2 + \lambda \|\alpha\|_1, \quad (2.10)$$

Where $\lambda > 0$. This is a convex and non-differentiable function. To solve such optimization function which promise a sparse solution many algorithms have been proved. We discussed the two main approaches that recently become very famous for solving an ill-posed problem in linear system of equations, including proximity optimization strategy and constrained optimization strategy.

2.5.3. Constrained optimization strategy

Constrained optimization strategy commonly utilized in order to obtain the solution of l_1 -norm regularization parameter. These methods treat the non-differentiable unconstrained problem by reformulating it as a smooth differentiable constrained optimization problem with an efficient convergence to obtain the squire solution (Schmidt, M., Fung, G., Rosales, R., 2009). There are different type of constrained optimization methods that solve the original unconstrained non-smooth problem, such as steepest descent direction, Gradient Projection Sparse Representation (GPSR), normal Sub-gradient strategy, coordinate-wise sub gradient strategy.

2.5.4. Steepest descent projection.

This method uses the gradient descent algorithm in order to solve the non-differentiable problem. Gradient descent is one of the thousand methods to solve the system of linear equations, by reformulating it to a quadratic minimization (QM) problem. Such QM problems, linear list squires

(2.12) can be mention. Thus, the solution of (2.6) in a general form is (2.26). The least squire method gives us a nice property (being convex) to implement a minimization problem (2.12).

$$x = D\alpha(2.6)$$

$$x - D\alpha = 0 \quad (2.26)$$

$$J(\alpha) = \min_{\alpha} \|x - D\alpha\|_2^2, \quad (2.12)$$

The minimization of (2.12) subject to α can be done via iterations (2.27).

$$\alpha^{n+1} = \alpha^n - \beta \nabla J(\alpha) \quad (2.27)$$

Where β is the learning parameter, in other words it scales the step of directional derivative, and $\nabla J(\alpha)$ is given by (2.28).

$$\nabla J(\alpha) = D^T(x - D\alpha) \quad (2.28)$$

Hence, the solution of least squire is given by iterating Gradient Projection Sparse Representation (GPSR) also called line search algorithm given by (2.29) as a negative gradient;

$$\alpha^{n+1} = \alpha^n - \beta D^T(x - D\alpha^n) \quad (2.29)$$

Recall the objective function in (2.10) the l_1 -norm is a not differentiable. As mentioned, we can reformulate this problem to an unconstrained problem via talking a directional derivative in a stepwise manner over cost function, which is a convex function. Hence, we are able to solve the first part of the optimization function (2.29), but the second part will be done by shrink the coefficients to zero based on a given optimality condition (2.30) to get the sparsest solution (Figueiredo, M.A.T., Nowak, R.D., Wright, S.J., 2007).

$$J(\alpha) = \min_{\alpha} \|x - D\alpha\|_2^2 + \lambda \|\alpha\|_1, \quad (2.10)$$

$$g_i^{(n)} = \begin{cases} (\nabla J(\alpha^n))_i, & \text{if } \wedge \alpha_i^{(n)} > 0 \vee (\nabla J(\alpha^n))_i < 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.30)$$

Therefore, choosing an initial guess for α with the optimality condition in (2.30), we would extend the (2.29) with the condition (2.30) given by (2.31).

$$\beta_0 = \min_{\beta} J(\alpha^n - \beta g^{(n)}) \quad (2.31)$$

In fact, we search by each iteration of α^n along the negative gradient $-\nabla J(z(n))$, projecting onto the non-negative orthant, and performing a backtracking line search until a sufficient decrease is achieved in J . Moreover, we shrink the coefficient to zero when its derivative is equal to the previous derivation conducted by an iteration operator. Furthermore, with the best starting point for α_0 we can guarantee a faster convergence with a proper step β which minimizes the algorithm in (2.31). Thereof an explicate computation for step length β_0 can be given by (2.32)

$$\beta_0 = \frac{(g^{(n)})^T g^{(n)}}{(g^{(n)})^T D g^{(n)}} \quad (2.32)$$

To avoid the value of α_0 to become very small or very large, we confine it within an interval of $(0 < \beta_{min} < \beta_{max})$ and to optimize the choice of best value in the interval we can define the *mid* (a, b, c) operation to define the middle value of its three scalar arguments (Figueiredo, M.A.T., Nowak, R.D., Wright, S.J., 2007).

One might consider the other sub Gradient strategies, for optimization function at non-differentiable points. In non-smooth optimization the local minimums achieved as a zero vector containing the elements of sub differential $\partial f(\alpha)$ CITATION Fle13 \l 1033 (Fletcher, 2013). The sub gradient of the absolute value function $|\alpha_i|$ given by the *signum* functions $gn(\alpha_i)$. The *signum* function takes on the sign of α_i whenever α_i is non-zero, and when α_i is zero then the *signum* function can take any value in range of $[-1, 1]$. Therefore the optimality condition transfer to the following (2.33):

$$g_i^{(n)} = \begin{cases} (\nabla J(\alpha^n))_i + \lambda \text{sign}(\alpha_i) = 0, |\alpha_i| > 0 \\ |(\nabla J(\alpha^n))_i| \leq \lambda, \alpha_i = 0 \end{cases} \quad (2.31)$$

The steepest descent projection for sparse solution achieved by a coordinate wise sub gradient method in which the optimality condition will be (2.32):

$$g_i^{(n)} = \begin{cases} (\nabla J(\alpha^n))_i + \lambda \text{sign}(\alpha_i), |\alpha_i| > 0 \\ (\nabla J(\alpha^n))_i + \lambda, \alpha_i = 0, (\nabla J(\alpha^n))_i < -\lambda \\ (\nabla J(\alpha^n))_i - \lambda, \alpha_i = 0, (\nabla J(\alpha^n))_i > \lambda \\ 0 \alpha_i = 0, -\lambda \leq (\nabla J(\alpha^n))_i \leq \lambda \end{cases} \quad (2.32)$$

This optimality condition, yield a descent direction for a sub-optimal α on the objective function,

2.5.5. Proximity optimization strategy

Proximity optimization strategy, aims to solve the problem of constrained convex optimization problems. The core idea in proximity algorithms motivated by employing proximal operator to solve the sub-problem in a iterative manner. This is more computationally efficient than the original problem. The proximity algorithm utilized in order to solve the non-smooth, constrained convex optimization problem (Parikh, N., and Boyd S. , 2013). In addition, the problem of sparse representation with l_1 -norm (2.10) is non-smooth convex optimization problem, which can efficiently tackled via employing proximal algorithm. Hence, the problem in (2.10) reformulated as (2.33).

$$\min P(\alpha) = \{\lambda \|\alpha\|_1 + \|x - D\alpha\|_2^2 | \alpha \in R^m\} \quad (2.32)$$

Which is consider as the constrained sparse representation of problem (2.10).

2.5.6. Iterative soft shrinkage algorithm

For solving the problem of non-convex ℓ_1 -norm sparse representation (2.32) a number of algorithm has been proposed, such as iteratively reweighted least squares (IRLS), iteratively thresholding method (ITM), and look up table (LUT) (Zuo, W., Meng D., Zhang, L., Feng, X., Zhang, D., 2014). To solve the problem of ℓ_1 -norm a soft thresholding operator (figure 2.4) given by (Donoho D., 1995).

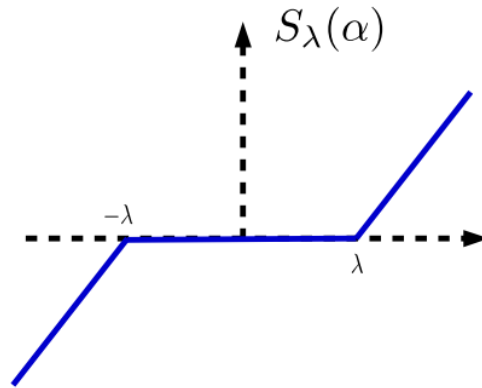


Figure 2.4. Generally, when the $|\alpha|$ is less or equal than the given threshold λ , the soft-thresholding operator uses the thresholding rule to assign $t_1(\alpha, \lambda)$ to 0. In contrast when $|\alpha|$ is bigger than given threshold then $(|\alpha| - \lambda)$ should be operated and along with sign function.

This thresholding method called Iterative soft shrinkage thresholding and can be used in many convex optimization strategy to solve the problem of ℓ_1 -norm. Indeed, mapping inverse problems can be formulated as an optimization problem and solved by forward backward or Iterative Shrinkage/Thresholding in which non-smooth functions with sparsity constraints can be minimized effectively. Furthermore, the soft shrinkage operator cannot deal with the biased estimation of the large coefficients. Hence injecting a step (steepest descent) on the approximation allows to reduce the bias in practice (Kowalski, M., 2015).

2.7. The Quest for Dictionary

Dictionary is one of the most important component of the sparsity base model (2.2). Dictionary is a set of training sample that used to recover the given signal/image. Sparse dictionary base model is a vast field and the entire details about this concept is not in the scope of this thesis. A dictionary must be properly designed in order to present the latent structure in the data.

$$x = D\alpha \text{ s.t. } \|\alpha\|_0 \leq k \quad (2.2)$$

Where $D \in R^{n \times m}$ is the dictionary, which is consider as the system of equations, n represent the number of equations and m denotes the number of unknowns in the system. Other words the rows are the data dimension and the columns are their corresponding observation that called atom (Elad, 2013). The system can be presented as either a linear system or non-linear system. However, the dictionary can be constructed in prior to the algorithm like basis pursuit (2.6). Hence, the minimization applied only on the coefficient vector α .

$$\min_{\alpha} \frac{1}{2} \|x - D\alpha\|_1 + \lambda \|\alpha\|_1, \quad (2.6)$$

Furthermore, the dictionary can be learned along with the coefficient vector α (2.24) (Mukherjee, S., Basu, R., Seelamantula, CS., 2016).

$$\min_{\alpha, D} \frac{1}{2} \|x - D\alpha\|_1 + \lambda \|\alpha\|_1, \quad (2.24)$$

There is also the case where the coefficients are fixed and only the atoms in the dictionary get update in each iteration. The main concern after defining a sufficient optimization algorithm is to answer the question of, how can we wisely choose D that performs well for the representation of the given signal and/or image. The following sections give a brief answer to this question. A various number of dictionary have been developed and proposed in response to the rising needs. These dictionaries emerge from two sources, (i) either mathematical model or (ii) realization of the data (Rubinstein, R., Bruckstein, A.M., Elad, M., 2010). Dictionaries formed by analytical formula refers to the earlier stage of transfer design such as Fats Fourier transformation, wavelets, wavelet packets, contourlets, and curvelets (Rubinstein, R., Bruckstein, A.M., Elad, M., 2010). However, the mentioned method is limited to lower dimensional signals and/or images. In the second approach, the fundamental goal of learn a dictionary is to preform best on the training set where the constructed dictionary represents the signal/image in informative presentation. Dictionary learning takes several routs. One can update the dictionary via minimizing the optimization function such that K-SVD can be mentioned. The other possibility is to construct a dictionary in prior to the optimization function such as Basis Pursuit (BP). That means the dictionary construct by some means, such as being orthonormal dictionary, which contains orthogonal column vectors. The goal of dictionary learning is to discover a set of base atoms (elements) that can describe the hidden pattern in the given data. In contrast, in dictionary learning, atoms in the dictionary are not require to be orthogonal. For dictionary learning algorithm, the dictionary can be an over-complete spanning-set, and has to be inferred form that input data. Forming a dictionary can be done via several algorithms. Such that, Recursive Least Square (RLS) which is a dictionary based algorithm, and continuously update the training atoms until convergence (Skretting, K., and Engan, K., 2010), Method of Optimized Directions (MOD) in this method, selection of atoms is done by frame design technique (Engan, K., Aase, S.O., Husoy, J.H., 1999), and many other method such as k-SVD can be mentioned. K-SVD method is a sparse base dictionary-learning algorithm, which motivated by k-mean algorithm and iteratively apply sparse coding on the obtained dictionary until it fists the data (Anaraki, F.B., Hughes S.M., 2013). Methods such as MOD, and K-SVD are not suitable for high dimensional dataset and they are prone to be stuck in local minimum (Rubinstein, R., Bruckstein, A.M., Elad, M., 2010). In addition, sparse dictionary learning is not considering the redundancy of the atoms and thus it has a high computational complexity (Zhu, Z., Qi, G., Chai, Y., Li, P., 2017). Moreover, the dictionary can be defined (Zhu, Z., Qi, G., Chai, Y., Li, P., 2017) before utilizing in a sparsity based model. This called dictionary construction and/or predefined dictionary (Rubinstein, R., Peleg, T., Elad, M., 2013; Vasanth Raj, P.T., and Hans W.J., 2015). Indeed, the dictionary can be mathematical

describe, having orthogonal columns that avoid redundancy in the dataset and reduce the number of samples to be presented in the dictionary. Hence the amount of computational time is significantly reducing. Such geometric dictionary has been proposed by (Zhu, Z., Qi, G., Chai, Y., Li, P., 2017) which is motivated by PCA that fits the high dimensional dataset very well.

Chapter 3

3.1. Hyperspectral Imagery

Earth observation via Remote Sensing imagery is gaining advancements in the era of hyperspectral imagery (HSI) figure 1. Image spectroscopy as a technique of acquiring information across electromagnetic spectrum allows us to capture images with contiguous hundreds spectral bands ranging from visible and solar infrared interval. “A hyperspectral image is captured as a three-dimensional data cube (figure 2) comprising two-dimensional spatial information and one-dimensional spectral information. The spectral signature of a pixel is a vector whose entries correspond to the spectral responses of an object in different bands” (Huang A, Zhang H , Pižurica A., 2017). Hyperspectral imagery data contains the more distinguishable information of the objects compare to multispectral imagery data. By the means that a hyperspectral image has higher spectral resolution than a multispectral image. With recent advent of very high-spectral resolution, hyperspectral imagery contributes to discover many material substances, which could not be discovered by multispectral imagery (Chang, 2013). Since HSI data are very sensitive to capture even small portion of electromagnetic range, numerous application arise including precision agriculture (Zhang, X., Sun, Y., Shang, K., Zhang, L., & Wang, S., 2016), environmental monitoring (Moroni, M., Lupo, E., Marra, E., & Cenedese, A., 2013) and urban planning (Weber, C., Briottet, Xavier, B., Aguejdad R., Aval, Josselin, A., 2018). Hyperspectral sensor provides hundreds spectral features which each feature called channel/band where each band covers a small portion of electromagnetic spectrum. Hyperspectral images cover a large area of surface via either space-borne or airborne platform.

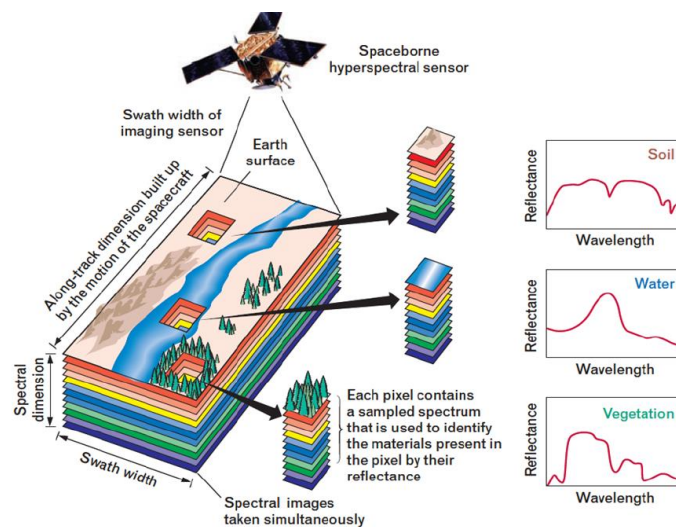


Figure 3.1. Illustrate the concept of a space-borne hyperspectral scene (Shaw, G.A. and Burke, H.K., 2003) capturing hundreds spectral information measured in each pixel as reflectance. The variation of spectrum over scene represent individual object. Spectral variation of three material shown in the left side of the figure.

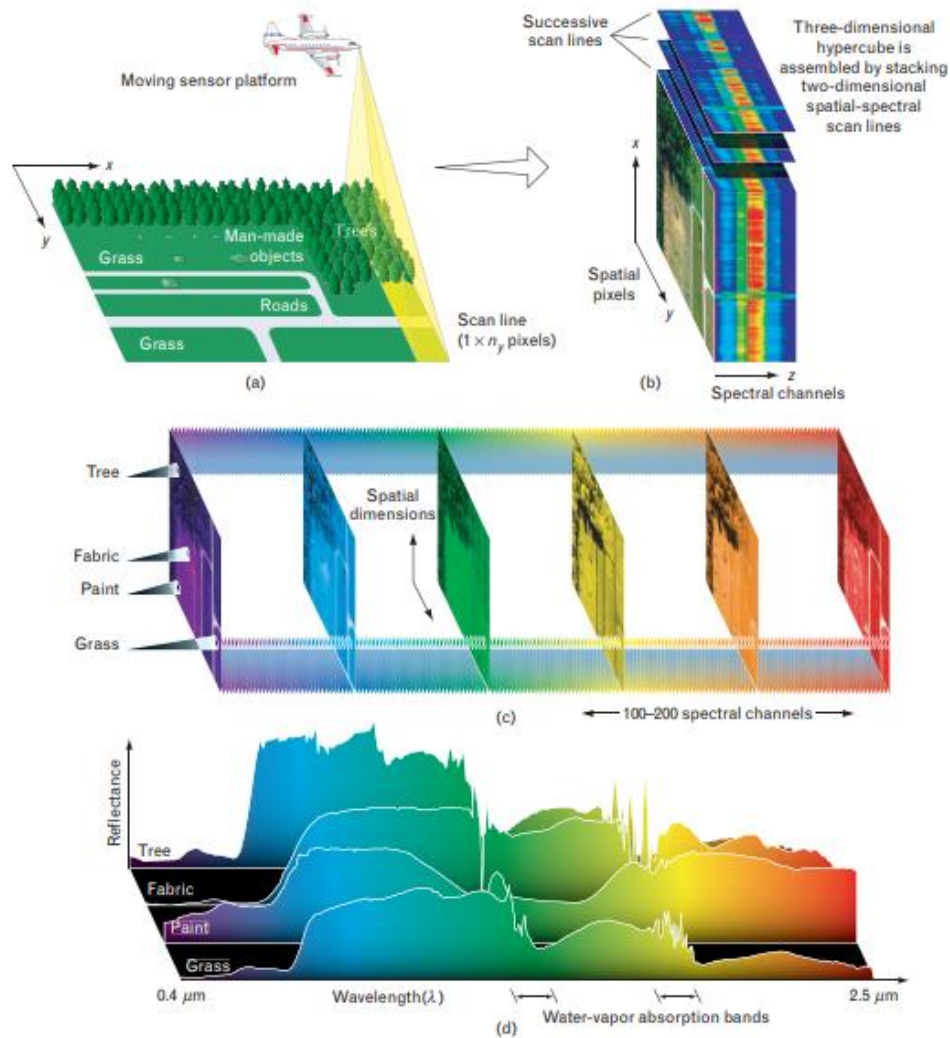


Figure 3.2. Structure of the hyperspectral data cube. (a) A push-broom sensor on an airborne or space borne platform collects spectral information for a one-dimensional row of cross-track pixels, called a scan line. (b) Successive scan lines comprised of the spectra for each row of cross-track pixels are stacked to obtain a three-dimensional hyperspectral data cube. In this illustration, the spatial information of scene presented by x and y dimensions of the cube, and the amplitude spectra of the pixels are projected into the z dimension. (c) The assembled three-dimensional hyperspectral data cube can be treated as a stack of two-dimensional spatial images, each corresponding to a particular narrow waveband. A hyperspectral data cube typically consists of hundreds of such stacked images. (d) Alternately, the spectral samples can be plotted for each pixel or for each class of material in the hyperspectral image. Distinguishing features in the spectra provide the primary mechanism for detection and classification of materials in a scene (Shaw, G.A. and Burke, H.K., 2003).

3.2. Hyperspectral Image processing

Retrieving information from hyperspectral data is a challenging task. Apart from hyperspectral image (HSI) preprocessing (e.g. atmospheric-correction, geo-correction), processing task is the

core of the feature analysis. Hyperspectral Image Processing (HSIP) due to the complexity and diversity and some number of limitations considered to be challenging. Hyperspectral images due to their capability of capturing narrow bands are prone to be a redundant set of spectral dimensions. Additionally, unlike conventional images with hyperspectral resolution, hyperspectral images are limited by relatively lower spatial resolution. Therefore, the problem of spectral unmixing (linear and nonlinear) arises which leads to the state of the art endmember extraction task (Dias, J.M.B., Plaza, A., Valls, G.C., Scheunders, P., Nasrabadi, N., Chanussot, J., 2013). Other important issues for processing hyperspectral data is their property regarding their higher dimensionality and temporal resolution. This demand of accelerating computational processing is to increasing the speed of interpretation of hyperspectral images in various applications. Considering one that wants to perform a time series analysis on crop agriculture via hyperspectral data (Eerens, H., Haesen, D., Rembold, F., Urbano F., Tote C., Bydekerke, L., 2014), then dealing with such big dataset needs efficient algorithms, which can minimize the computational time and maintain the accuracy. There are many type of processing approach based on the application and ultimate goal. Such image-processing task in hyperspectral image can be mentioned as, spectral unmixing, endmember extraction, target detection, change detection, edge detection and dimensionality reduction. In hyperspectral imagery, dimensionality reduction is generally following an approach of retrieving the original direction of information in which the remains spectral information are orthogonal bases vectors. Ultimately, in spectral dimensionality reduction the goal is to present the data in the lower and most relevant feature direction with retain the maximum variance in the spectral signature. This approach also leads to the elimination of the noise associated to the real signal via only separating noise from signal dimension space. For dimensionality reduction of hyperspectral images several algorithms being used such as Principle Component Analysis (PCA) (Li, Y., Wu, Zebin Wu., Wei J., Plaza, A., Li, J., Wei Z., 2015), Linear discriminate analysis (LDA) (Li, W., Prasad, S., Fowler, J.E., Bruce, L.M., 2011). A comprehensive description of hyperspectral images processing and analyzing is behind the scope of this thesis. The focus is on hyperspectral image classification via sparsity-based model. Some references regrading hyperspectral image processing and analysis in different topics including endmember extraction, unmixing and compression and son on, can be covered by (Chang, 2013; Valls, G-C., Tuia, D., Chova L-G., Jiménez, S., Malo J., 2012). In this thesis, the focus is on hyperspectral image classification but a few common and relevant tasks in hyperspectral image analysis will be briefly explain such as pixel unmixing/endmember extraction and dimensionality reduction.

3.3. Spectral Unmixing and Endmember Extraction

The fundamental goal in hyperspectral image processing in terms of spectral unmixing which lead to endmember extraction is to retrieve a group of pure spectrum of an individual object in the entire scene that called endmember. The basic approach for hyperspectral image processing is to match each individual spectral signature (pixel) to one of the spectral reference in spectral library. This

approach is only feasible when the entire pixels of the scene of interest have their pure representation in spectral library. Therefore, a prior measurement in the field is demanding for this basic approach. Hyperspectral images are limited by relatively lower spatial resolutions. Thus, most of the hyperspectral images are containing mix pixel. That means most of the individual pixels are more likely prone to be a mixture of different material. This is also may cause by the mixture of distinct material that are formed naturally. The resulting image of mixed-pixel spectrum may resemble multiple reference spectra. Endmember implies the original member/material of the pixel which represent the set of abundances at each pixel that indicates the percentage of each endmember that are presented in pixel (Bioucas-Dias, J-M., Plaza, A., Dobigeon, N, Parente, M., Du, Q., Gader, P., Chanussot, J., 2012).

There are many techniques for endmember extraction. These techniques can be expressed in two assumptions. The first assumption is based on existence of pure pixel and the second one is regarded by the assumption of absence of pure pixels for the endmember extraction (Plaza, J., Hendrix E.M.T., García I., Martín, G., Plaza, A., 2012). The first assumption lays on the existence of at least one pure pixel in hyperspectral dataset for each individual material on the scene. However, this is usually not a valid assumption due to spatial resolution, phenomenal mixing and other considerations (Plaza, J., Plaza, A., Perez, R., Martinez, P., 2009). Techniques among many others (Du, Q., Raksuntorn, N., Younan, N.H., King, R.L., 2008). Therefore, the focus is on the developing an algorithm for endmember identification that do not relies on the presence of pure pixel. Many unmixing algorithms have been developed (Bioucas-Dias, J-M., Plaza, A., Dobigeon, N, Parente, M., Du, Q., Gader, P., Chanussot, J., 2012; Ma, W-K., Bioucas-Dias J.M., Chan, T-H., Gillis N., Gader, P., Plaza, A-J., Ambikapathi, A., Chi C-H., 2014) using several approaches such as geometrical (Donoho, D-L., I-M-J Biometrika., 1994), statistical (Nascimento, José M. P., Bioucas-Dias, José M., 2012) and sparsity based model (Tang, W., Shi, Z., Wu, Y., Zhang C., 2014). Spectral unmixing algorithms are mostly use the invers computation of the spectral signature to retrieve the endmember. Two common such models are linear and non-linear unmixing model. In linear model a concept of linear combination of the pixel are used but non-linear models are more complex and use techniques such as kernel based models (Wang, W., Qian, Y., 2016) and machine Learning algorithms (Ahmed , A-M., Duran, O., Zweiri, Y., Smith, M., 2017). However, the liner-mixing model has been studied and presented be a well suited and standard technique. It is assumes that each spectral vector can be approximately recovered by a linear combination of endmembers weighted by their corresponding fractional abundance which in the scene (Figure 3.3). Given a spectral vector $x \in R^B$ where B is the number of bands then the model can be mathematically given as,

$$x \cong \sum_{j=1}^d \Phi_j \beta_j + \epsilon \quad \text{s.t } \beta_j \geq 0 \quad d = 1, 2, \dots, d \quad (3.1)$$

$$\sum_{j=1}^d \beta_j = 1.$$

Where $\Phi \in R^{B \times d}$ is a set of given endmember as columns in which $\Phi_j \in R^B$, $j = 1, 2, \dots, d$ as an endmember, $\beta = [\beta_1, \beta_2, \dots, \beta_p]^T \in R^d$ and $\epsilon \in R^B$ is the associated error which each band taken into the model.

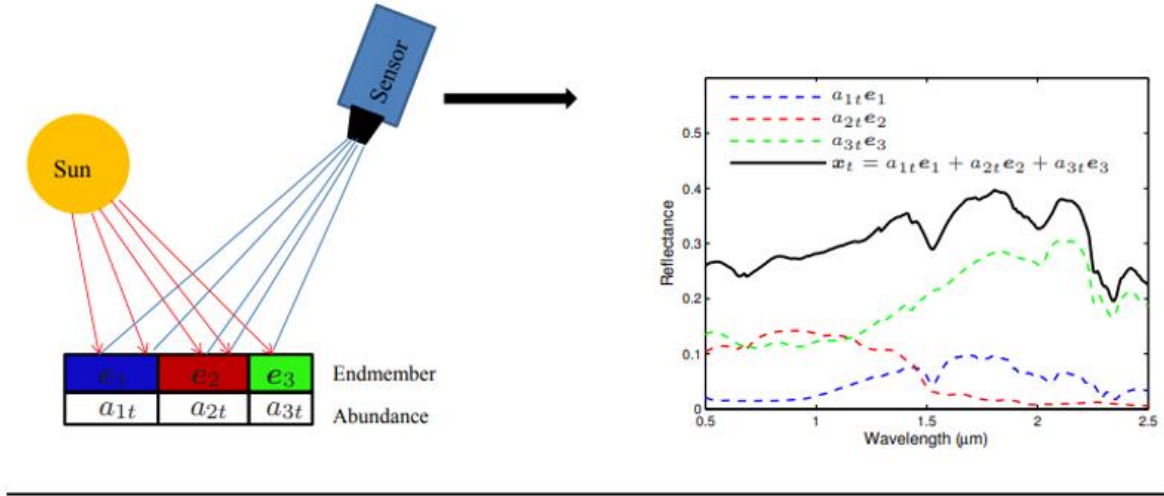


Figure 3.3. Demonstrates the linear mixing. The observed spectrum x_i is a combination of the endmembers e_1, e_2, e_3 , with the respective weights a_{1t}, a_{2t}, a_{3t} called abundances.

Spectral unmixing in practice calls for efficient sparse regression techniques (Iordache, M-D., Bioucas-Dias, J., Plaza, A., 2011). Under the process of spectral unmixing via linear mixture model several steps must be performed, namely, atmospheric correction that conduct transformation of radiance to reflectance, data reduction, unmixing and invers operation. Data reduction preforms dimensionality reduction on hyperspectral images that lead to a faster computation. This step discussed in section 3.2.2. During unmixing stages, the endmember along with their abundances at each pixel will be identified. Eventually the invers operation solved by an optimization problem given a spectral vector and endmember. The objective function aims to minimize the residual between given pixel and the combination of endmember and abundance (coefficient) vector.

3.4. Dimensionality reduction for Hyperspectral Images (HSI).

The high dimensionality of spectral features in a HSI may affect the classification result in terms of accuracy and computational time speed (Pal, M.; Foody, G., 2010; Tong, F., Tong, H., Jiang, J., Zhang, Y., 2017). Hyperspectral images acquiring information in hundred contiguous spectral bands, and thus the data volume to process are considered to be huge. In addition, hyperspectral sensors due to their narrow bands observation are expected to have a relatively strong linear dependency across contiguous spectral bands that convey almost the same information. Indeed, such a large number of dataset associated manipulating feature vector in different higher dimension spaces. In sophisticated algorithms, working with such large dimension of feature spaces is more challenging. Time computation and significant storage influence the accuracy of the statistical estimation of a fixed number of samples. Accordingly, this phenomenon known as "Hughes

phenomenon" in hyperspectral image processing (Schweizer, S.M, Moura, J.M.F., 2001; Bellman, 1956). Dimensionality reduction is used for various purposes in exploitation of hyperspectral images. Such as image compression, feature selection, denoising, classification. To address the Spectral dimensionality issue, two dimensionality reduction (DR) approaches are generally used, including DR based transformation (DRT), and DR based band selection (DRBS). For DRT approach, a general task is to use a component analysis (CA) algorithm that allows us to transfer the data into a lower dimension space that presents the fundamental direction of the dataset that is mainly based on some statistical assumption like maximum variance. As such CA algorithms, principle component analysis (PCA) (figure 1.4) is one of the widely used and famous transformation methods that is used in hyperspectral image processing (Rodarmel, C., Shan J., 2002). Singular value decomposition (SVD) as a PCA based algorithm, maximum noise fraction (MNF) (A. Green, M. Berman, P. Switzer and M. Craig., 1998) and hyperspectral signal identification by minimum error (HySime) (J. Bioucas-Dias and J. Nascimento., 2008) are the most well-known dimensionality reduction and denoising algorithms for hyperspectral datasets. The PCA based on eigenvalue decomposition (Rodarmel, C., Shan J., 2002) aims to find a lower dimensional space by presenting the most relevant eigenvectors. Indeed, PCA computes the eigenvalues and their corresponding eigenvectors, and on the basis of the magnitude of eigenvalue selects a set of PCs that are orthogonal bases and retain the most variation in the dataset. There are many methods for spectral dimensionality reduction. Such that, higher order statistics-based CA transformations like Independent Component Analysis (ICA) and PCA can be mentioned. Nevertheless, the problem of using ICA is that there is no assumption of significance of selected components while in PCA the corresponding principal components have the eigenvalues and in SVD, singular values that represent the significance of selected number of PCs (Chang, 2013). Generally, the purpose of DRT is to compact the data by reducing the dependency in the spectral feature dimension and present it in a lower dimension of spectral feature that provides the fundamental direction of the feature spaces. The DRBS aims to preserve the original dimension and use the advantage of the information gain (IG) and the spectral curve of the hyperspectral dataset (Xie ID Fuding., F.Li., Lei, C., Ke, L., 2018), to select the most relevant bands for a particular object and leave the other bands behind.

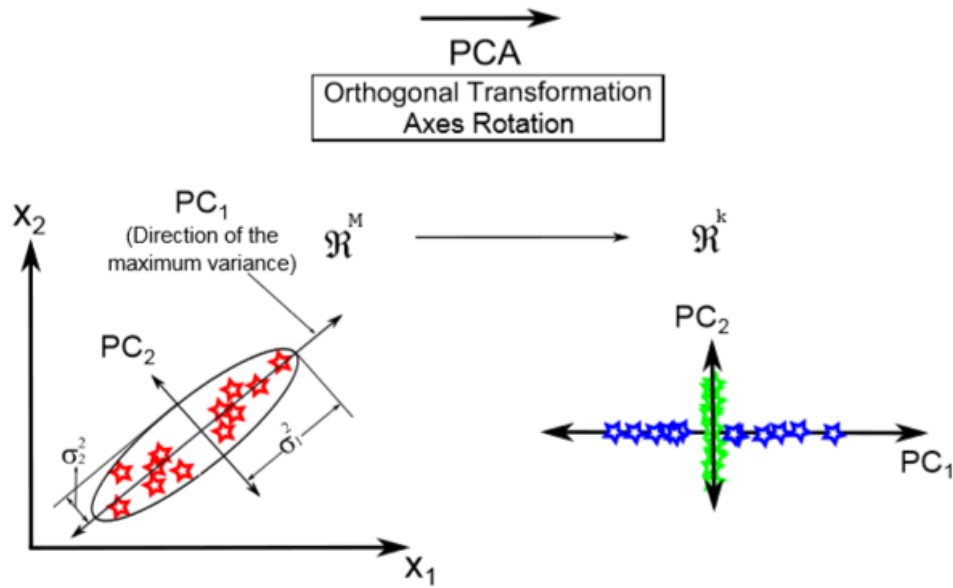


Figure 3.4. Illustrates orthogonal projection using Principle Component Analysis concept. Example of the two-dimensional data (x_1, x_2) . The original data are on the left that presented in their original coordinate, i.e. x_1 and x_2 , the variance of each variable is graphically represented and the direction of the maximum variance, i.e. the principle component PC_1 , is shown. On the right hand side the original data are projected (after shifting the mean center to the origin) on the first (blue stars) and second (green stars) principle components.

The spectral domination reduction of hyperspectral image (e.g. figure 1.5) can be described as follows:

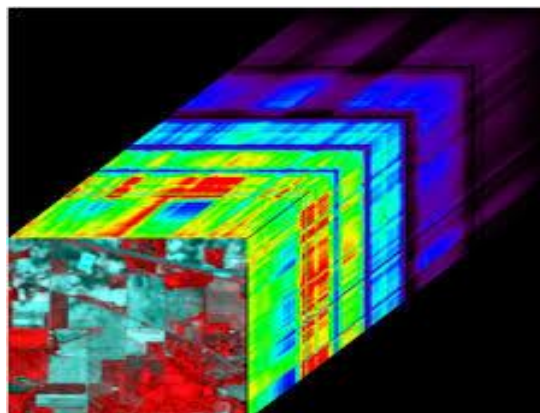


Figure 3.5. Indian Pines dataset, which has the size of $145 \times 145 \times 220$ (AVIRIS)

In order to compute the principle component such hyperspectral dataset can be represent as a matrix A instead of multiple arrays that eventually the matrix $A \in R^{d \times B}$ where d and B denote the number of pixels in the scene and spectral bands respectively. The fundamental goal of DRT in hyperspectral images is to de-correlate the neighboring contiguous bands by projecting them to

uncorrelated co-ordinate system in a lower dimensional k space which can be denoted as transforming data from $A \in R^{d \times B}$, to $A_{reduce} \in R^{d \times k}$. It has been shown by (Singh, A. and A. Harison, 1985) that in remote sensing, it may be more effective to work with data co-variances rather than data variances. That is implies to standardized principal components analysis (SPCA). Assume $S = \{d_i\}_{i=1}^d$ is a set of B dimensional pixel (feature) vectors and μ is the mean value of i -th B dimensional feature vectors in the sample pool S given by $\mu = \left(\frac{1}{d}\right) \sum_{i=1}^d d_i$. Let transpose the sample data matrix A to be $A^T \in R^{B \times d}$, and call it matrix $X = [d_1 d_2 \dots d_d]$. Then the sample covariance matrix of the S is obtained by,

$$C = \frac{1}{d} [XX^T] = \frac{1}{d} [\sum_{i=1}^B (d_i - \mu)(d_i - \mu)^T].$$

The covariance matrix is represented as follows,

$$\begin{pmatrix} Var(x_1, x_1) & Cov(x_1, x_2) & \dots & Cov(x_1, x_M) \\ Cov(x_2, x_1) & Var(x_2, x_2) & \dots & Cov(x_2, x_M) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(x_M, x_1) & Cov(x_M, x_2) & \dots & Var(x_M, x_M) \end{pmatrix}$$

The output of a covariance matrix is a square matrix of spectral bands where $C \in R^{B \times B}$ in which B is as before. The covariance matrix is a positive semi definite $C = C^T$ that the diagonal entries contain variance of each spectral band B_i , $i = 1, 2, \dots, B$ and the off diagonal entries represent the per-wise covariance between each two variables as shown in matrix above.

Covariance matrix used in order to find the PCs space by computing eigenvalues and their corresponding eigenvectors given by,

$$V\Sigma = \lambda V$$

Where V and λ denote the eigenvector and eigenvalue of covariance matrix respectively. The eigenvalues are the scalar values and eigenvector are the principle component vectors with non-zero entry. The eigenvectors represent the direction of PCA space and the eigenvalue are the scalar multiplication of eigenvector that represent the robustness of the eigenvector (Hyvärinen, 1970; Strang, G., & Aarikka, K., 1986).

3.5. Hyperspectral Imagery classification

Remote Sensing data are becoming more and more feasible in facilitating research on studying any object even system on the Earth. Such that researches that become very practical nowadays in the real world problems are including, urban planning, land management, urban management, environmental modeling, agricultural crop management and monitoring, landscape planning, environmental conservation, biodiversity monitoring, ecology, Energy management. Nowadays much governmental and non-profit organization produce a substantial amount of remote sensing data with high quality and accuracy. Producing such vast amount of data in terms of diversity, velocity, and volume needs efficient algorithms to treat them and extract actionable insight from this large and complex collection of digital data. Hence, these data are prone to be considered as big data. Processing and analyzing remote sensing data are varies based on the ultimate purpose. Some fundamental tasks need to be done beforehand in analyzing this type of data. This fundamental task comprises two main approach called preprocessing and processing. Preprocessing of a remote sensing data including, Geometric Correction, Atmospheric Calibration, missing data reconstruction and so on. Nevertheless, in the processing task the focus is on performing fundamental tasks for analyzing phenomena. Such this tasks, image enhancement, pixel-unmixing, image denoting, image classification can be mentioned. Image classification is one of the most important task in remote sensing data processing. Image classification has many applications in time series analysis, anomaly detection, change detection, target detection, habitat changes. Due to the mentioned complexity of remote sensing data, advanced algorithms have been proposed for classification tasks. Generally, several factors make the analysis of hyperspectral images complex and sophisticated. Indirect measurements like remote sensing always contaminated with noise and mixed pixel in which analyzing such data is a hard task calls the advance methods and algorithms. Many learning algorithms have been proposed for hyperspectral image classification, such as supervised and unsupervised classification in which supervised learning algorithms use a set of observation to train the machine and find the best separating hyperplane (logistic regression, support vector machine) and unsupervised learning algorithms use a clustering algorithm and based on the proposed cluster classify the new given pixel. Performing a classification task on remote sensing imagery data can be done in several approaches such as pixel-wise, Subpixel wise, and object- based image classification (Li, M., Zang, S., Zhang, B., Li, S., Wu, C., 2014).

3.6. Pixel-Wise Image Classification.

Pixel-wise classification method assume each pixel is pure and labeled as an endmember of land-cover/land-use (Xu M., Watanachaturaporn P., Varshney P., Arora M., 2005). With this method, remote sensing images are considering as a collection of pixels with spectral information that are used as input data set for pixel-bas classification. In general, pixel-wise classification algorithms can be divided into two groups: unsupervised classification and supervised classification. In supervised classification, pixels are represented in different groups according to the given labels (ground truth). The supervised classification algorithm, Support Vector machine (SVM), Logistic regression, Maximum Likelihood Classifier (MLC) (Shalaby A., Tateishi, R., 2007), Sparse

Representation can be mentioned. In unsupervised classification strategy the pixels are groups in different cluster based on the intensity value (Puletti N., Perria R., Storchi P., 2014). Moreover, there are several strategies that spatial information can be included. Such unsupervised classification algorithms, K-means, Iterative Self-Organizing Data Analysis Technique (ISODATA) (El_Rahman, 2016) can be mentioned. Among, these all algorithm machine learning techniques have shown better performance and result. Machine learning algorithms are developed to enhance the knowledge learning process. Artificial neural networks, Decision trees (Gislason, P.O., Benediktsson J.A., 2006), SVM (Mountrakis, G., Im, J., 2011) can be mentioned. Recently, advanced computer vision and signal processing algorithms have been applied on remote sensing data classification (Huang A, Zhang H , Pižurica A., 2017). Such this algorithm sparse coding has gained a great attention. But more generally, all these algorithms perform the classification task with the two main approaches that demonstrated in figure 3.6.

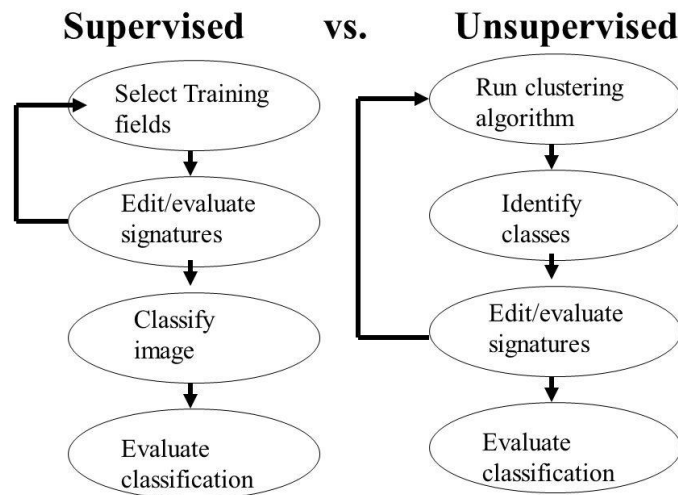


Figure 3.6. Illustrates the two different approaches for image classification.

Chapter 4

4.1. Efficient sparse signal recovery for Hyperspectral Imagery data classification

In this chapter, we develop classification principles for high dimensional spectral images called hyperspectral imagery in remote sensing domain. The general idea is to model a high spectral feature dimension pixel as a column vector, which is represented by some dictionary. The assumption is that, for different groups of pixels we have by a-prior knowledge different dictionaries are available. The classification process results in sparse recovery algorithms, where the recovered sparse vector contains basic information for the membership to the one of the classes.

4.2. Classification Problem, a Prior-knowledge

Assume we want to classify a B -dimensional pixel $x \in R^B$ into one of C preassigned classes, for which we have for the 1st class d_1 test samples that are stored as d_1 (column vector) in B dimension and represented in matrix $D_1 \in R^{d_1 \times B}$, for the 2nd class, we have d_2 test samples that are stored in D_2 and so on. Each group of pixels (prior-knowledge) individually presented in a matrix called sub-dictionary. Eventually all of these matrices concatenated in a unique matrix called dictionary that holds the properties of all of the given classes which is given by,

$$D = [D_1, D_2, D_3, \dots, D_C] \quad (4.1),$$

where $D_1 \in R^{B \times d_1}$ denotes the sub dictionary 1, and so on.

$$D = \begin{pmatrix} x_{1,d_1} & x_{1,d_2} & \cdot & \cdot & x_{1,d_C} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{B,d_1} & x_{B,d_2} & \cdot & \cdot & x_{B,d_C} \end{pmatrix} \quad (4.1)$$

$D \in R^{B \times d}$ where $d = d_1 + d_2 + d_3 + \dots + d_C$ which denotes the atoms (column vector) in the dictionary.

4.3. Data Model and Classification Principle

The classification idea goes as follows: Consider a given test pixel $x \in R^B$ to be classified. Hence, if x is a member of k -th class stored in the dictionary D , then it should be close to one of the atoms in D_k . In another words, a pixel $x \in R^B$ can be modeled as a linear combination of a set of vectors $d = \{d_1, d_2, d_3, \dots, d_n\}$ called atoms in dictionary. That mathematically is given by,

$$x \approx D_k \alpha_k \quad (4.2),$$

where $\alpha_k \in R^{d_k}$, and $\|\alpha\|_{R^{d_k}}^2$ is large enough.

Once the group C are geometrically well separated, it should not be possible to reasonably represent x by means of training samples from D_j from all $j = 1, 2, \dots, C$ but $j \neq k$, i.e. there exist no coefficient vector $\alpha_j \in R^{d_j}$ with $j \neq k$ such that

$$x \approx D_j \alpha_j (4.3)$$

Or even

$$x \approx D_1 \alpha_1 + \dots + D_{k-1} \alpha_{k-1} + D_{k+1} \alpha_{k+1} + \dots + D_C \alpha_C. (4.4)$$

In other words, if we consider the complete linear representation that involves all test samples,

$$x \approx D_1 \alpha_1 + D_2 \alpha_2 + \dots + D_C \alpha_C (4.5)$$

The task is to identify that dictionary for which the norm (or energy) $\|\alpha\|_{R^{d_k}}^2$ is significantly larger than the same quantity for the other class, i.e. x that belongs to the k -th class if,

$$\|\alpha_k\|_{R^{d_k}}^2 \gg \|\alpha_j\|_{R^{d_j}}^2 \text{ for all } j \neq k.$$

Furthermore, the dictionary should clearly describe the individual classes. That means the individual test samples d_j in each D_j from in almost all cases must be geometrically redundant set of vectors. This issue discussed and a construction approach proposed in section 4.4.1.

4.4. Sparse Recovery Principle as a Classification Problem

Sparse representation draws much attention in recent years and many application of sparse representation can be found that SR is reasonably a useful algorithm for them (X. Lu, H. Wu, Y. Yuan, P. Yan, and X. Li, 2013; Y. Yuan, X., Li, Y., Pang, X., Lu, and D., Tao., 2009), such that image classification can be mentioned, where the basic goal is to classify an image based on the predefined groups. The sparse representation based classification method dose not differ with the fundamental concepts of compress sensing (CS) theory which is including sparse representation, encoding measuring, and reconstruction algorithms (M. Elad., 2010). The sparse representation classifications (SRC) generally assumes that there exists a linear combination of a class sample that can approximately represent the given test sample x from the same class (4.2). SRC computes the sparse representation coefficients of the linear system of equations and eventually measures the reconstruction error called residuals for each individual class by employing their corresponding training sample and sparse coefficients that contributes to the approximation. Ultimately, the test sample x will be assigned to that class that expose minimum construction error (residual) given by,

$$class(x) = argmin_j (x) = arg \min \|x - D_j \alpha_j\|_2, j = 1, 2, 3, \dots, C (4.6).$$

Sparsity based method has shown great superiorities for image classification task since it deals very well with corrupted/noisy image (Z. Zhang, Z. Li, B. Xie, L. Wang, and Y. Chen., 2014). The sparse approximation for image classification can be expressed in two main categories in terms of the way of exploiting the atoms; (i) holistic representation based method and local representation based method. Holistic representation based method exploit the training sample of all classes to

represent the test sample, while local representation based method employs only the column vectors (atoms) of individual and/or several class at the same time (Y. Xu, D. Zhang, J. Yang, and J. Yang., 2011). On the other hand the approximation of x by means of dictionary D , the family of the representation α satisfying (2.4) is actually infinitely large with the degree of freedom identified with the null-space of D (Rubinstein, R., Bruckstein, A.M., Elad, M., 2010). Therefore, we must optimize the representation family to a set of informative representation coefficients that approximately satisfies the problem in (4.2). Hence, this problem can be reformulated as (4.7) with respect to some cost function.

$$\alpha = \arg \min_{\alpha} C(\alpha) \quad \text{Subject to } x = D_k \alpha_k \quad (4.7)$$

Practical choice of cost functional $C(\alpha)$ promotes sparse representation of the coefficients. Indeed, we want the sorted coefficients to decay quality. Hence, solving problem in (4.7) is referred to sparse recovery principle. We can define a cost function as some robust penalty function, which can be loosely defined as a function that is tolerant to large coefficients but strictly penalizes small non-zero coefficients. Sparse representation in terms of optimization are consider in four optimization problems, (i) the smooth convex problem, (i) non-smooth convex problem (ii) smooth non-convex problem, and (ii) non-smooth non-convex problem (J. A. Tropp, A. C. Gilbert, and M. J. Strauss, 2006; Tropp, 2006). Thus, the temptation of the convexity motivated us to use the least square solution. Although, the least square is a convex function but the choice of penalty l_p where $0 < p < 1$, for sparsity makes the problem hard (not differentiable) since the $p \leq 1$ is not a convex problem. First, let us consider our objective functional for the optimization problem given by,

$$\min_{\alpha} \|x - D\alpha\|^2 + \lambda \|\alpha\|_1 \quad (4.8).$$

There is a vast literature regarding solving the problem in (4.8), but it is very important to pick up the one and develop it that fits the data best, by means of accuracy and computational time load. However, for solving this unconstrained problem we start with a proximity optimization (Iterative soft-shrinkage) approach easy to compute and to implement schemes. This is followed by an efficient acceleration with proceeded descent method. Ultimately, we end up with optimization problems with joint sparsity measures that lead to sparse block-wise recoveries directly yielding the classification result. The following sections detail the steps mentioned above.

4.4.1. l_1 Sparse recovery via Soft-Shrinkage Iteration.

Consider the model

$$x^\delta = D\alpha + \epsilon \quad (4.9),$$

where $x^\delta \in R^B$ is the given pixel to be classified, $D \in R^{B \times d}$ is the dictionary in which B and d denote the number of spectral feature (band) and number of training sample respectively, and $\alpha = [\alpha_1^T \alpha_2^T \dots \alpha_C^T]$ represent the coefficient of the model. In order to recover α , we have to solve the minimization problem given by

$$\min_{\alpha} \|x^{\delta} - D\alpha\|^2 + \lambda \|\alpha\|_1 \quad (4.8).$$

Suppose $\|C\| < 1$ (otherwise rescale the system). Then an iterative computation is given by,

$$\alpha^{n+1} = S_{\lambda/2} \left(\alpha^n + D^T (x^{\delta} - D\alpha^n) \right) \quad (4.10)$$

4.4.2. l_1 Constrained Recovery via Projected Steepest Descent Iteration.

The method in (4.10) can be easily accelerated by applying the projected steepest descent method, which is described in Section 2.3.1.1 of Chapter 2. Indeed, the problem in (4.8) formulated as an unconstrained optimization problem. Consequently, the model in (4.10) will be associated with a constraint called step length, in order to compute the backward, forward minimization and find the steepest direction that leads to not only faster even global convergence. Starting point is to separate the cost function and the penalty term from each other that are concatenated as the optimization function (4.8). In other words, braking down the non-smooth function to local minimization problem by computing the gradient within each step length. Hence, all of the local gradient will give the representation of the global minimum. The procedure goes as follows: Consider the following optimization problem,

$$\min_{\alpha \in B_k(l_1)} \|x^{\delta} - D\alpha\|^2 \quad (4.11),$$

resulting in the projected iteration,

$$\alpha^{n+1} = P_{B_k(l_1)} \left(\alpha^n + \beta^n D^T (x^{\delta} - D\alpha^n) \right) \quad (4.12),$$

where $B_k l_1 = \{Y \in l_2: \|Y\|_1 \leq K\}$ and the step length control β^n .

Based on the derived α , we are now able to classify x . This approach delivers some sparse α but where the recovered non-zero coefficients are very likely distributed across different classes, i.e. an extra classifier must be applied to assign x to one class. This also can be done based on the introduced minimum residual classifier (2.6).

4.4.3. Joint Sparsity Measure Recovery using Projected Steepest Descent iteration.

The essential goal (and the possible advantages of this approach) is to identify in a unique way that dictionary that is most relevant for the sparse representation of x . The optimization problem is almost the same but with the joint sparsity or so-called block sparsity measure given by,

$$\min_{\alpha \in B_k(l_1)} \|x^{\delta} - D\alpha\|^2 + \lambda \sum_{k=1}^C \|\alpha_k\|_2 \quad (4.13),$$

or the constrained version leading to the projected steepest descent,

$$\min_{\alpha \in B_k(l_1)} \|x^{\delta} - D\alpha\|^2 \quad (4.14),$$

Where $\mathcal{B}_k l_1 = \{Y \in l_2: \sum_{K=1}^C ||Y_k||_2 \leq K\}$. The resulting method is again the projected steepest descent iteration,

$$\alpha^{n+1} = P_{\mathcal{B}_k(l_1)} \left(\alpha^n + \beta^n D^T (x^\delta - D\alpha^n) \right) \quad 4.15,$$

Where $\mathcal{B}_k l_1 = \{Y \in l_2: \sum_{K=1}^C ||Y_k||_2 \leq K\}$ and the step length control β^n . The essential difference is the structure of the projector $P_{\mathcal{B}_k(l_1)}$.

The Joint sparsity algorithm has the properties of the two previous algorithms (developed version of Iterative Soft Shrinkage algorithm) and provide a new functionality via shrinking the non-relevant coefficients to zeros in a block-wise manner. It works in the way that the norm of each individual class gets computed at each iteration and when the computed norm is equal or less than the given threshold the coefficients of that class will be jointly set to zero. Otherwise, another optimality condition operation must be assigned.

4.5. Condensation of the a-prior given dictionaries.

Dictionary is the core component of sparse recovery algorithms. Indeed, constructing an informative dictionary is a key step for sparsity-based model. Once the atoms in a dictionary presented in an optimal manner we are able to significantly reduce the computational time load and represent the given pixel in a sparsest manner. The dictionary should be inferred from data, and for a classification task, a label for each member should be given in prior.

4.5.1. Geometric base dictionary construction.

In linear inverse problem, it is expected that a discriminate dictionary can be learned from training samples so that a test sample can be truly represented for classification (Feng Z., Yang M., Zhang L., Liu Y., Zhang D., 2013). A dictionary is constructed by concatenating several sub-dictionaries. Therefore, constructing a sub-dictionary involves a main issue, which is being a redundant set of column vectors (samples/atoms) that present the main directional space. In another word, being a full rank matrix. The other issue is that it also might be considered as the inequality in the number of sample for each class (sub-dictionary). The number of training sample for each individual class may not be equal, which is a common problem in classification tasks. This problem of the data is called imbalanced data problem (Zou, X., Feng, Y., Li, H., and Jiang, S., 2017). Hence, imbalanced data refers to classification problems where we have unequal samples/instances for different classes. For example, assume we have three classes (sub-dictionary) that are stored in a dictionary D given by;

$$D = [D_1 \ D_2 \ D_3] \in \mathbb{R}^{B \times d},$$

Where B and d denote the number of sample dimension and number of training sample respectively. Let us consider $d = (d_1 + d_2 + d_3)$ where $d = 5500$, in which $d_1 = 100$, $d_2 = 2000$, and $d_3 = 3400$. Another words $D_1 \in \mathbb{R}^{B \times 100}$, $D_2 \in \mathbb{R}^{B \times 2000}$, and $D_3 \in \mathbb{R}^{B \times 3400}$. As shown the number of sample in each class is obviously different, particularly in the first class (d_1).

This is called imbalanced data problem that might have several consequences such as miss classification, and computational time load. Hence, to tackle the above issues, we proposed a geometric base dictionary where Principal Component Analysis (PCA) strategy elegantly implemented to be apply on each individual sub-dictionary. PCA can solve two main problems in data representation, on one hand, redundant training sample can decrease the computational time load and on the other hand, it leads to the enhancement of discrimination between different classes in dictionary (Feng Z., Yang M., Zhang L., Liu Y., Zhang D., 2013). The PCs space are given by an orthogonal linear transformation. This property allows us to choose the most informative samples that retain the most variation in the dataset while they are not in the span of each other, another words, the represented samples are a set of redundant set in which they are linearly independent (Jolliffe, 2002). There are two numerical methods to calculate the principal components. The first method is eigenvalue decomposition (EVD), that is applied only on diagonal matrix i.e. $A^T A$, while the second one uses the Singular Value Decomposition (SVD) method which can be applied directly on the data matrix. The implementation of PCs space via EVD described in the next section.

4.5.1.1. Principle Component Analysis via Eigenvalue Decomposition (EVD).

Principal Component Analysis (PCA) is an important dimension reduction tool which finds the orthogonal directions reflecting the maximal variation in the data. This allows us to present the data in a lower dimensional, by projecting data onto these directions (Shen, D., Shen, H., Marron, J.S., 2016). PCs space via eigenvalue decomposition, are obtained from a diagonal matrix, variance-covariance matrix (a positive semi-definite matrix) of the data.

Covariance Matrix

The covariance, or variance, matrix is a diagonal matrix of the data sample. The output of a covariance matrix is a square matrix. The covariance matrix is a positive semi definite $cov_{ij} = (cov_{ij})^T$ in which variance are represented in diagonal entries.

The input matrix for computing covariance matrix for each individual sub-dictionary can be given as follows:

Consider the D_1 to be sub-dictionary 1;

$$D_1 = \begin{pmatrix} x^{(1,1)} & \dots & x^{(1,B)} \\ \vdots & \ddots & \vdots \\ x^{(d,1)} & \dots & x^{(d,B)} \end{pmatrix} \in R^{d \times B}$$

Where B and d denote the number of bands (feature dimension) and number of training sample (observation) respectively. In the next stage the training sample is centered by subtracting the mean value of each training sample in B dimension from each data point in B dimension. Another words, the mean should be computed in row for the above matrix.

Compute mean in row:

$$\mu_i = \frac{1}{B} \sum_1^B x_i, \text{ where } i = 1, 2, 3, \dots, d \quad (4.16)$$

$$\begin{pmatrix} \mu_1 = \frac{1}{B} \sum_1^B x_1 \\ \mu_2 = \frac{1}{B} \sum_1^B x_2 \\ \mu_3 = \frac{1}{B} \sum_1^B x_3 \\ \vdots \\ \mu_i = \frac{1}{B} \sum_1^B x_i \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_i \end{pmatrix} \in R^d \quad (4.16)$$

Then the covariance of the data samples d_i and d_j where $i \neq j$ in B dimension is given by,

$$Cov(d_i, d_j) = \frac{1}{d-1} \sum_1^d (x_i - \mu)(x_j - \mu) \quad 4.17,$$

Or

$$\Sigma = Cov(D) = E((D - \mu)(D - \mu)^T).$$

Eventually the covariance matrix Σ can be represented as following;

$$\Sigma = \begin{pmatrix} var(d_1, d_1) & \cdots & Cov(d_1, d_d) \\ Cov(d_d, d_1) & \cdot & Cov(d_d, d_1) \\ \vdots & \ddots & \vdots \\ Cov(d_d, d_1) & \cdots & var(d_d, d_d) \end{pmatrix} \in R^{d \times d}$$

A positive value indicates the positive correlation between data samples and a negative value represent the negative relation and eventually a zero value of covariance represent the independency between the samples in a particular sub-dictionary.

Compute the Eigenvalue Decomposition

A (non-zero) vector V of dimension d is an eigenvector of a square $d \times d$ matrix Σ (covariance matrix) if it satisfies the linear equation given by,

$$V\Sigma = \lambda V \quad (4.18),$$

where V and λ denote the eigenvector and eigenvalue of covariance matrix respectively. The eigenvalues are the scalar values and eigenvector are the principle component vectors with non-zero entry. The eigenvectors represent the direction of PCA space and the eigenvalue are the scalar multiplication of eigenvector that represent the robustness of the eigenvector (Hyvärinen, 1970; Strang, G., & Aarikka, K., 1986). Accordingly, the problem is to solve the linear system of equation in (4.18) for eigenvalues given by,

$$p(\lambda) = \det(\Sigma - \lambda I) = 0 \quad (4.19)$$

Eventually the eigenvalue decomposition of the covariance matrix Σ is given by

$$\Sigma = Q\Lambda Q^{-1} \quad (4.20),$$

where Q is the square $d \times d$ matrix whose i -th column is the eigenvector q_i of Σ , and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues $\lambda_{ii} = \lambda_i$. After factorizing the covariance matrix of the given sub-dictionary, we are able to select the K -th eigenvalues that retain the maximum variance in respect to the magnitude of their corresponding eigenvalue. Figure 4.1 represent the working follows of computing PCs space via eigenvalues decomposition.

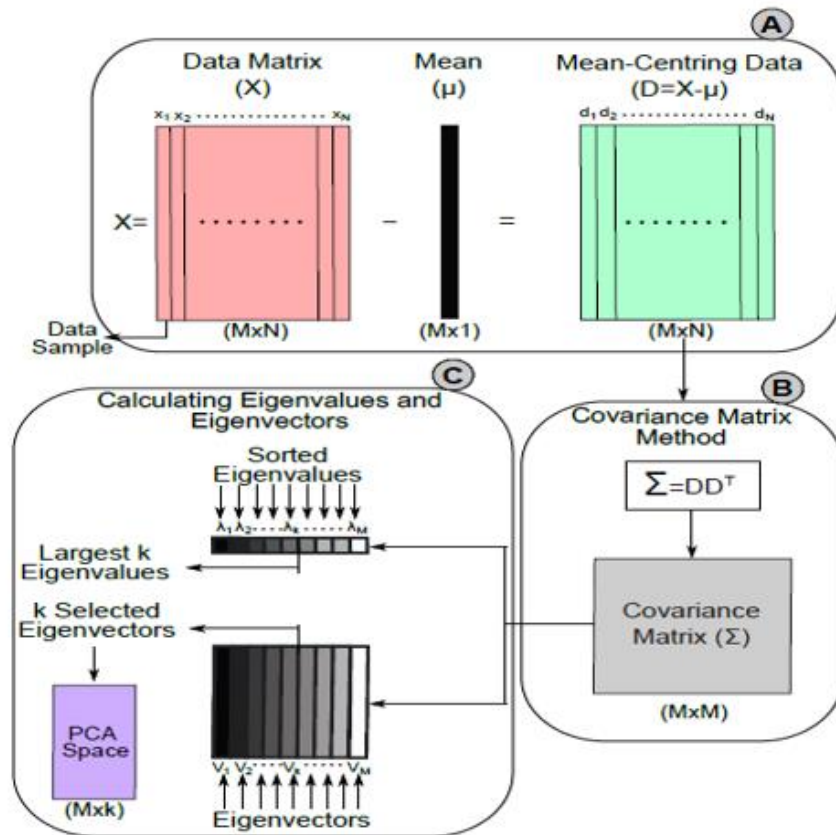


Figure2. Depict the steps for computing PCs space via eigenvalue decomposition.

The projection of the data to the lower dimension goes as follows:

In order to construct the lower dimension of PCA space, a linear combination of the first k number of eigenvector (PCs) that have the higher eigenvalues are selected in order to retain the maximum variance in the dataset. Thus, the lower dimension is find as $W = (v_1, v_2, v_3, \dots, v_i)$. Eventually the

diminution of the data is reduced by projecting the original data in sub-dictionary to the lower dimension of PCs space (figure 4.2) which is given by following,

$$Y = W^T D_1 = \sum_{i=1}^d W^T (x_i - \mu) \quad (4.21)$$

where $Y \in R^k$ is the projected data in lower dimension (W). Therefore, the dimension of sub-dictionary is reduced from

$$D_1 \in R^{d \times B} \quad \text{to} \quad Y \in R^{K \times B}$$

The projection depicted in figure 4.2.

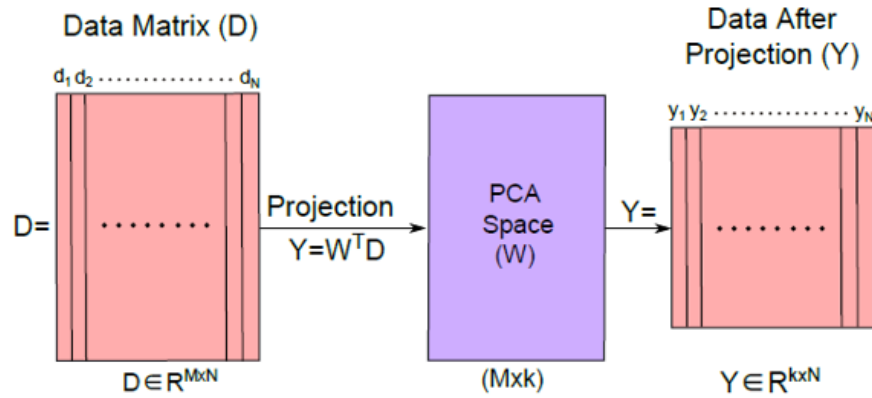


Figure 4.2. Illustrate the projection of the original data to the lower dimension.

Robustness of Pcs space.

Eventually the validity test of choosing the K number of principle component (eigenvalue) can be examined by their corresponding eigenvector. Indeed, the main parameter in PCA that needs to be adjusted is the number of PCs (k) to be selected. Hence, the robustness of PCA space can be control by k number of selected eigenvector and measured by the division of sum of the k selected eigenvalues by sum of all eigenvalues given by,

$$RR = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^B \lambda_i} \geq 95 \quad (4.21),$$

where RR is the Robustness Ratio for k (Abdi, H., & Williams, L. J., 2010).

Chapter 5

5.1. Experimental design

Hyperspectral images are the result of indirect measurements. This type of measurements always are contaminated with noise. In addition, they have several limitations that needs to be resolved. Lastly, this un-structural data sets do not follow the homogeneity property and prone to be big data. Thus, performing a classification task on such dataset needs advanced and efficient algorithms. Furthermore, using such data set for operational decisions needs scalable algorithms for the streaming application. Processing big data especially in stream application needs fast and simpler algorithms that also provide reliable result. Therefore, sparse representation is proposed as an effective algorithm. In sparse representation it turns out that many coefficients are not needed (Qazi Sami ul Haq, et all, 2010) by restricting them by a regularization parameter to keep them small and set to zero that also lead to avoid overfitting. Thus, the image size can be reduced. Linear representation methods have been extensively studied (B. K. Natarajan, 1995) and draws many attentions in real life application problems (M. Huang, W. Yang, J. Jiang, Y. Wu, Y. Zhang,, 2014). Sparse representation is one of the linear representation methods, which has been proven being highly efficient and powerful solution to a wide range of application particularly in signal processing, image processing, machine learning and computer vision such as image denoising, image classification and image segmentation (Zhang Z., Xu Y., Yang J., Li X., Zhang D., 2016).

5.2. Background and relevant work.

Sparse representation based classification where the main goal is to classify the given pixel based on a set of predefine categories. The sparse representation base classification (SRC) first assumes that there exist a set of linear combination mechanism between features/training samples that the approximation of their coefficients lead to an efficient representation (least non-zero entries) of the given test sample from the same subject. Eventually by calculating the residual of each class, employing the sparse representation coefficient and test samples will be assigned to that class with the minimum residual (Zhang Z., Xu Y., Yang J., Li X., Zhang D., 2016). Sparse representation classification can be done via supervised classification approach, where there exist a prior-knowledge for each class. Indeed, based on the availability of ground truth supervised learning preformed to classify new given pixel (test pixel) based on the training pixels, which used for the construction of dictionary. In recent years, sparse representation has gained a great attention in hyperspectral image classification (Chen, Y., Nasrabadi, N.M., Tran, T.D., 2011; Zhang, H.; Zhai, H.; Zhang, L.; Li, P., 2016; Zhang H., Li J., Huang Y., Zhang L., 2014; Chen Y., Nasrabadi N.M., Tran T.D., 2013; Wang J., Jiao L., Liu H., Yang S., Liu F., 2015; Bian X., Chen C., Xu Y., Du Q., 2016; Chen C., Chen N., Peng J., 2016). There are many algorithms for hyperspectral image classification.

Such algorithms like Support Vector Machine demonstrate a significant performance and results are comparable to the tradition algorithm used for remote sensing imagery classification (Fauvel, M., Benediktsson, J. A., Chanussot, J., Sveinsson, J. R., 2008). Despite the capability of SVM for classification, in hyperspectral images there exist several difficulties explained above that makes

the classification problem more challenging. In order to overcome those difficulties, the contribution of spatial information along with spectral signature becomes very popular (Bian, X., Zhang, T., Yan, L., Zhang, X., Fang, H., Liu, H., 2013). Such that (Hu, L., Qi, C., Wang, Q., 2018) developed a classification approach by integrating spatial information with spectral values using an SVM base classifier, and their result demonstrate a significant improvement in hyperspectral image classification. Recently, sparse representation algorithms have been used for hyperspectral image classification (Huang A, Zhang H , Pižurica A., 2017). In fact, sparse representation for remote sensing images classification has found its boundary by accepting the concept that via cooperating spatial information with spectral information the performance of classification will be highly improved (Song B., Li J., Mura M., Li P., Plaza A., José M., Dias B., Benediktsson J., Chanussot J., 2014). Song et. al. has proposed to exploit sparse representations of morphological attribute profiles for remotely sensed image classification. They have integrated both spatial and spectral information and have eventually used sparse recovery to reduce the high dimensionality of the given data for both multi and hyperspectral images (Song B., Li J., Mura M., Li P., Plaza A., José M., Dias B., Benediktsson J., Chanussot J., 2014). Similarly, Chen et.al has introduced a joint sparse representation classification (JSRC) algorithm for HSI classification. The proposed model is based on the prior knowledge that the pixels in a patch describe the same spectral characteristics, which can be represented by a common set of atoms in dictionary (Chen, Y., Nasrabadi, N.M., Tran, T.D., 2011). Huang et. al. have proposed another sparse representation for hyperspectral image classification. They have integrated sparse and Gaussian noise along with cooperating spatial and spectral information to develop the classification of HSI using sparse representation. Furthermore, they have called the successful proposed algorithm joint sparse recovery classification (Huang A, Zhang H , Pižurica A., 2017). These approaches involve the problem of choosing the window size for integrating of spatial information. However, utilizing sparse representation algorithms needs a comprehensive understanding of them, which is not in the scope of this thesis but for more details please refer to this paper (Zhang Z., Xu Y., Yang J., Li X., Zhang D., 2016). In all the previous work for HSI classification using sparse representation the development is only done on construction of dictionary by exploiting spatial and spectral information and eventually the optimization function is solved using greedy algorithm such as orthogonal matching pursuit (OMP) (Huang A, Zhang H , Pižurica A., 2017). In addition, the integration between spatial and spectral information can be affected by the Hughes phenomena (Hughes, 1968) (the limited amount of training sample). Hence, sparse representation classification considers this problem and shows a better performance (Wang, H., H., Turgay., 2018). In this thesis, a geometric base dictionary has been proposed along with an advanced version of proximity algorithm to recover the corresponding coefficient sparsely while providing high accuracy and efficient computational cost load. In this thesis, we only exploit spectral information and hence the spatial information of the given dataset is not considered in the construction of the dictionary.

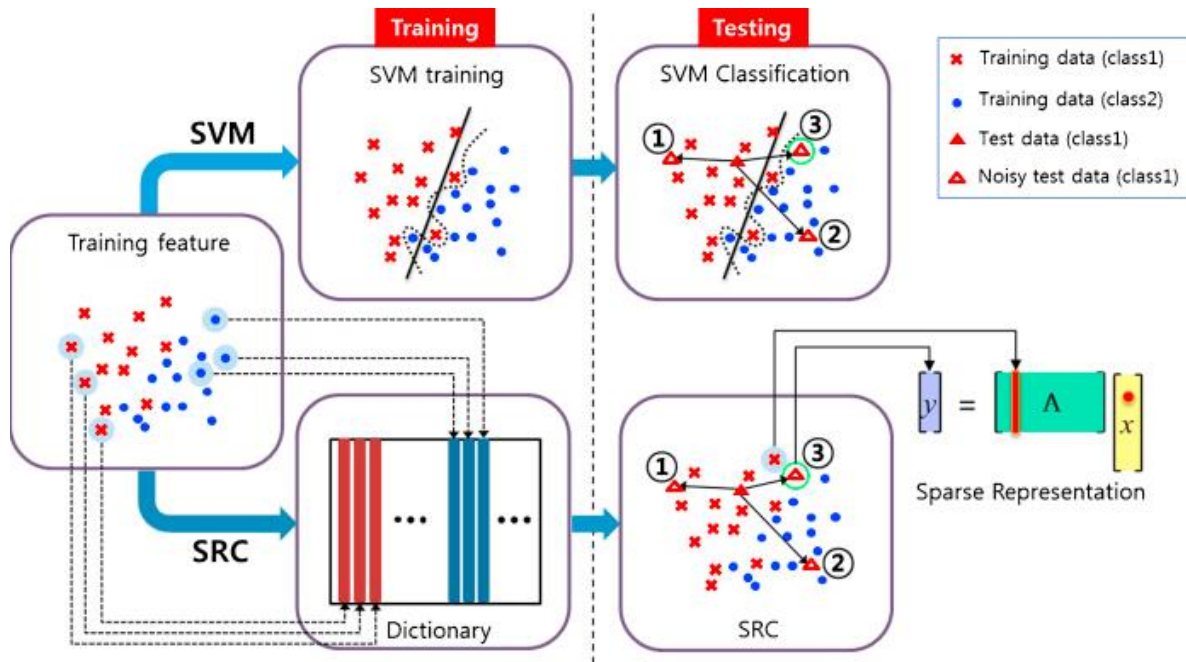


Figure 5.1. Demonstrates the classification task via SVM and Sparse representation Classification (SRC) (Shin, Y., Lee, S., Ahn, M., Cho, H., Jun, S.S., Lee, H., 2015).

5.3. Experiments

In this section, the details of applying the proposed sparse representation schemes for hyperspectral image classification are discussed, and the results in each stage of the development for hyperspectral image classification are provided. Later in the Section 5.3 and its subsections, we evaluate the performance of the proposed schemes, and eventually we end up with comparing the performance of the proposed package in each stage of its development. Figure 5.2 illustrates the general working follow of proposed efficient sparse signal recovery algorithm for hyperspectral image classification. The details of construction of the geometric based dictionary discussed in section 5.2.2.1.

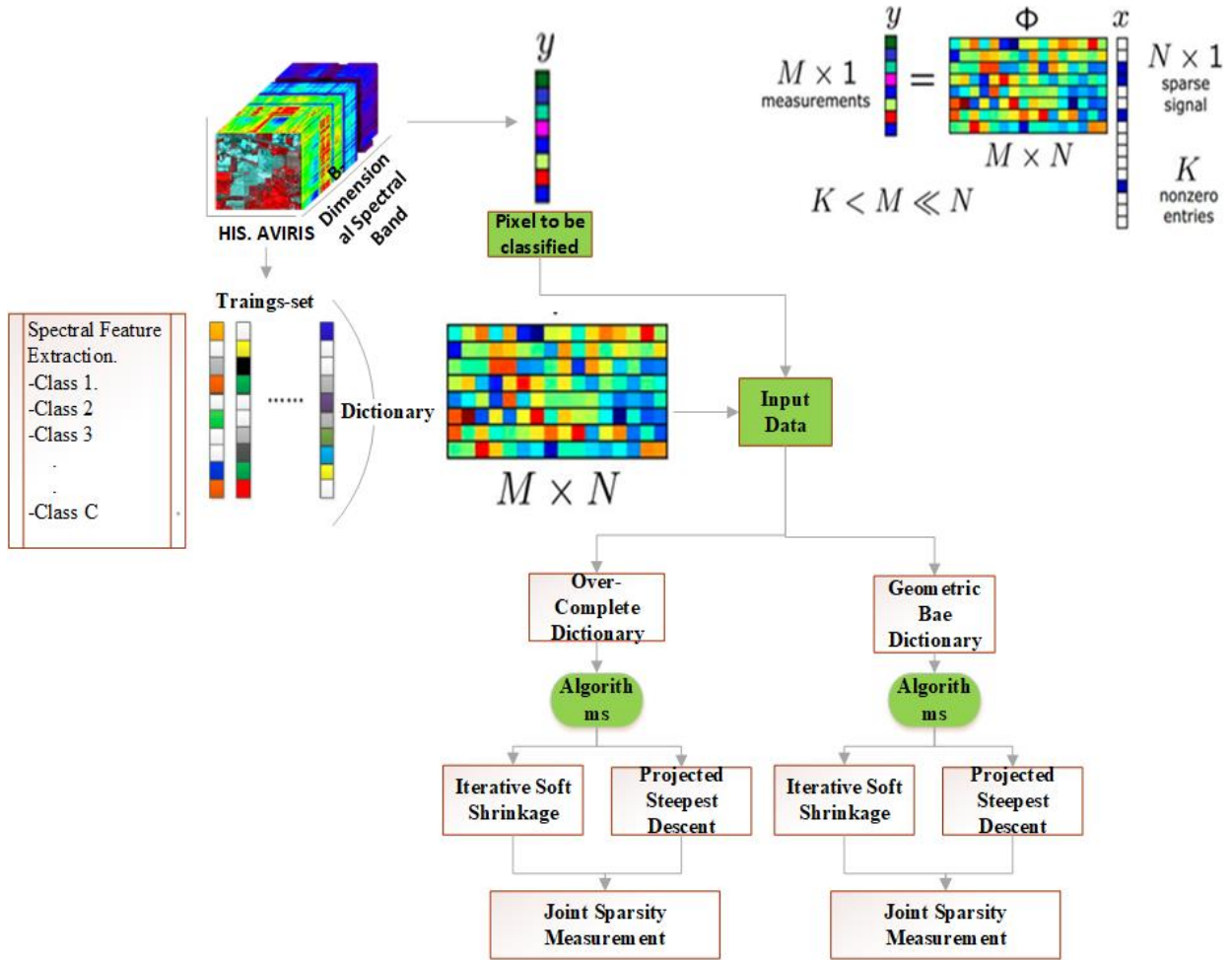


Figure 5.2. The General working diagram for proposed efficient sparse signal recovery algorithm for hyperspectral image classification.

5.3.1. Data Set Description

In this thesis in order to evaluate the capability of the proposed efficient sparse signal recovery algorithm for classification task, we use Indian pines dataset presented in 200 bands and the scene size of 145×145 pixels, which acquired by AVIRIS sensor¹.

¹ This dataset is freely available at: [http://www.ehu.es/ccwintco/index.php?title=Hyperspectral Remote Sensing Scenes](http://www.ehu.es/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes)

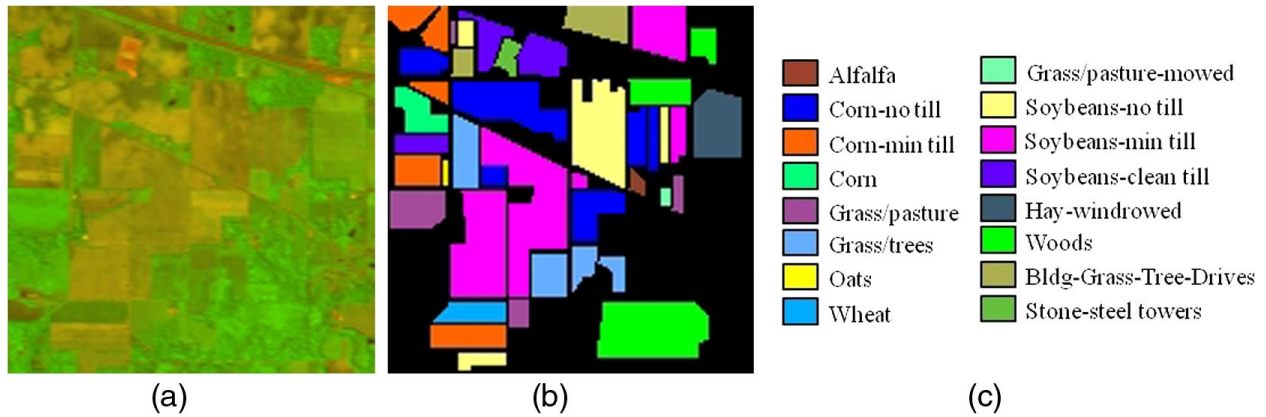


Figure 5.3. (a) Represent the Indian Pines scene, (b) its corresponding ground truth in 16 classes, and (c) the legend of the corresponding classes.

Indian Pines has the spatial resolution of 20 m. Two-thirds of this scene is covered by agriculture crops, and one-third by forest and other perpetual vegetation. The corresponding class for each pixel represented in 16 classes in an image called ground truth (figure 5.3). The associated classes for Indian Pines dataset presented in table 5.1.

#	Class	Samples
1	Alfalfa	46
2	Corn-notill	1428
3	Corn-mintill	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windrowed	478
9	Oats	20
10	Soybean-notill	972
11	Soybean-mintill	2455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1265
15	Buildings-Grass-Trees-Drives	386
16	Stone-Steel-Towers	93
17	Unknown samples	10776

Table 5.1. Sample size for Indian Pines dataset.

The classes are not mutually exclusive. Here, we choose only four classes for evaluation of our proposed package. The classes used in experimental designed are including class 4, 5, 14, and class 16. In order to present the data to the algorithms, the dataset randomly separated in two parts called training and testing/validation set. 70 percent is used to design the dictionary and 30 percent utilized for validation. Table 5.4 represent the samples size for each class and their pixel value in 200 dimensions.

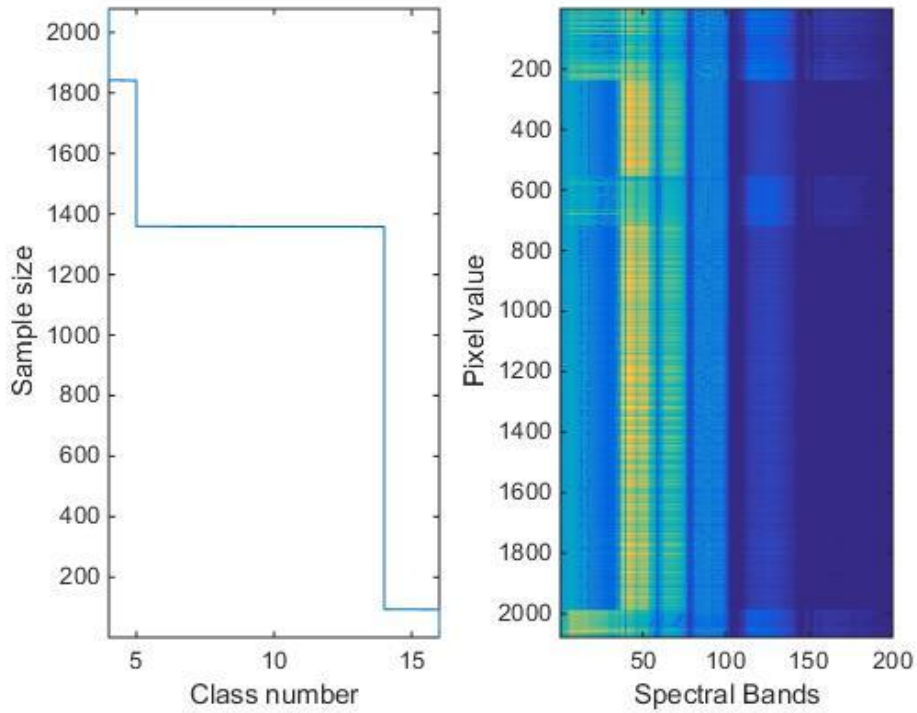


Figure 5.4. Left side represent the sample size for each class, and in the right side the pixel values are represented for the chosen classes in 200 spectral bands.

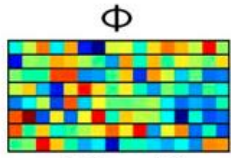
5.3.2. Experimental Design

Consider a given test pixel $x \in R^B$ to be classified. Hence, if x is a member of k -th class stored in the dictionary D , then it should be close to one of the atoms in D_k . In other words, a pixel $x \in R^B$ can be modeled as a linear combination of a set of vectors $d = \{d_1, d_2, d_3, \dots, d_n\}$ called atoms in dictionary. Hence, we can present the given class in (5.1). Figure 5.3 depicts a general form of sparse recovery.

$$x = D\alpha + \epsilon$$

$M \times 1$
 measurements
 y

=

Φ

 $M \times N$

$N \times 1$
 sparse
 signal
 x

$K < M \ll N$

K
 nonzero
 entries

Figure 5.4. Demonstrates the sparse recovery concept, where only a linear combination of few nonzero elements and their corresponding atoms are needed to represent the given test sample. Here $x \in R^B$ is the given test pixel in which B denotes the spectral dimension, $D \in R^{B \times d}$ denotes the dictionary that has the B dimensional spectral bands and d is the number of training sample

stored as column vectors, also called atoms, $\epsilon \in R^B$ is the Gaussian noise, and $\alpha \in R^d$ is the sparse coefficients, that is given by the following property

$$\|\alpha_k\|_{R^{d_k}}^2 \gg \|\alpha_j\|_{R^{d_j}}^2 \text{ for all } j \neq k. \quad (5.2)$$

Then, the optimization problem is given by

$$\min_{\alpha} \|x^\delta - D\alpha\|^2 + \lambda \|\alpha\|_1 \quad (4.8)$$

Since this problem is not smooth (discussed in chapter 4), we develop a classification principle by implementing a fundamental approach that has an easy and efficient scheme to tackle the problem. By employing proximity techniques, we could relax the problem (4.8), and in the next stage, we convert this unconstrained problem to a constrained problem in order to accelerate the implemented proximity algorithm and possibly improve the accuracy. Eventually, we develop the optimization formula further as a joint sparsity algorithm that helps to identify a unique dictionary that is clearly identify the given test sample.

For designing the efficient sparse signal recovery algorithm for hyperspectral data, we reshape the array of $145 \times 145 \times 200$ Indian Pines dataset to a matrix of 21025×200 , after that, the four mentioned classes extracted from the scene. Each class is represented as a sub-dictionary. In order to present an informative dictionary for the proposed schema, we design a geometric dictionary by applying singular value (SVD) on each sub-dictionary. Thus, the principle direction space of each sub-dictionary is represented by choosing a sufficient amount of redundant sample space. In the next step, we concatenate all the four sub-dictionary (classes) in a matrix called dictionary. Before applying the proposed algorithm on given dictionary and test set, in order to have a unit l_2 -norm the atoms in dictionary normalized. Atom normalization is a step that in some point become crucial especially when one implementing iterative and algorithm that needs to converge. The Normalization is done using the following equation,

$$z = \frac{d_c - \mu_c}{\sigma_c} \quad (5.6)$$

Here, d_c denotes the C-th atom, μ_i denotes the mean value over all feature dimension (bands) for C-th atom, and σ_c represent the standard deviation of C-th atom over all feature dimension (bands). Standard deviation expresses the data dispersion around the center (mean). The input matrix for scaling the atoms for each individual sub-dictionary can be given as follows,

$$D = \begin{pmatrix} x^{(1,1)} & \dots & x^{(1,d)} \\ \vdots & \ddots & \vdots \\ x^{(B,1)} & \dots & x^{(B,d)} \end{pmatrix} \in R^{B \times d}$$

where B and d denote number of bands (feature dimension) and number of atoms (observation) respectively, and x represent the pixel.

Compute mean:

$$\mu_D = \left[\bar{d}_1 = \frac{1}{B} \sum_{i=1}^B x_i, \bar{d}_2 = \frac{1}{B} \sum_{i=1}^B x_i, \dots, \bar{d}_C = \frac{1}{B} \sum_{i=1}^B x_i \right] \in R^{1 \times d} \quad (5.7)$$

Compute standard deviation:

$$\sigma_D = \left[\sigma_1 = \sqrt{\frac{1}{B} \sum_{i=1}^B (x_i - d_1)^2}, \sigma_2 = \sqrt{\frac{1}{B} \sum_{i=1}^B (x_i - d_2)^2}, \dots, \sigma_C = \sqrt{\frac{1}{B} \sum_{i=1}^B (x_i - d_C)^2} \right] \in R^{1 \times d} \quad (5.8)$$

Compute Standardization Equation:

$$\frac{\begin{pmatrix} x^{(1,1)} & \dots & x^{(1,d)} \\ \vdots & \ddots & \vdots \\ x^{B,1} & \dots & x^{(B,d)} \end{pmatrix} - \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_C \end{pmatrix}}{\begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_C \end{pmatrix}} = \frac{\begin{pmatrix} x^{(1,1)} - d_1 & \dots & x^{(1,d)} - d_C \\ \vdots & \ddots & \vdots \\ x^{B,1} - d_1 & \dots & x^{(B,d)} - d_C \end{pmatrix}}{\begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_C \end{pmatrix}} = \begin{pmatrix} x^{(1,1)} & \dots & x^{(1,d)} \\ \vdots & \ddots & \vdots \\ x^{B,1} & \dots & x^{(B,d)} \end{pmatrix} \in R^{B \times d}$$

Eventually the classification is done for each step of the development of the proposed efficient sparse signal recovery algorithm. The results of classifications for each stage of the development are given in the next sections.

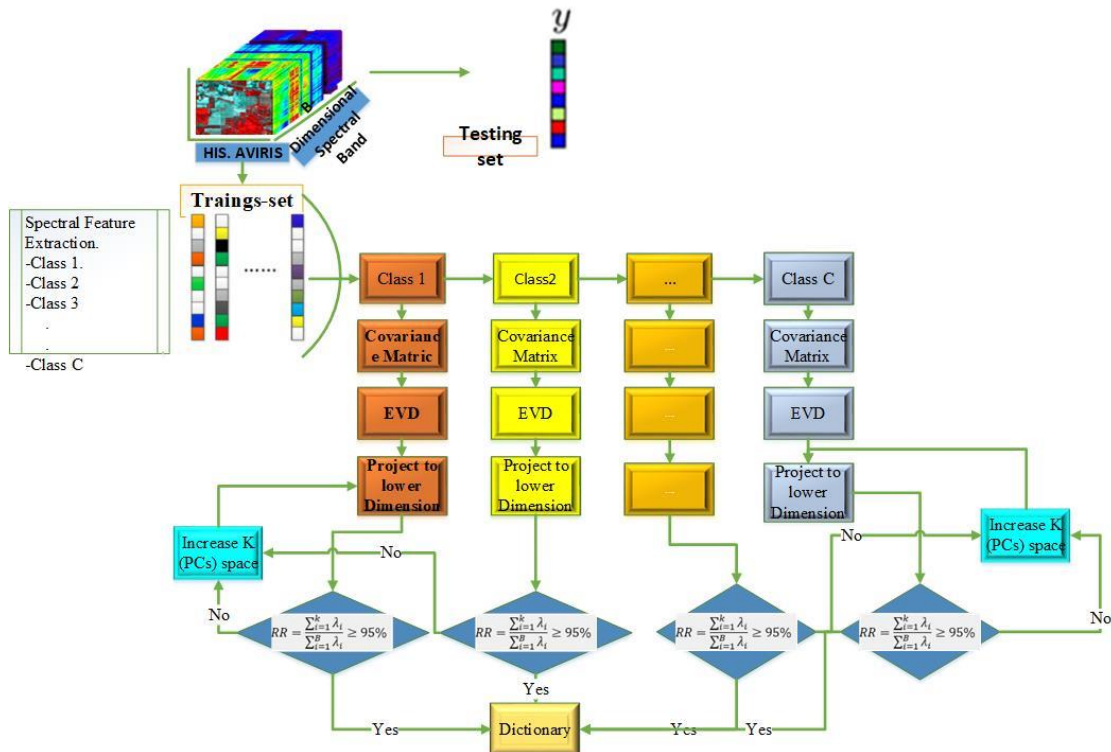
5.3.2.1. Dictionary Construction

Dictionary is the core component of sparse recovery algorithms. Indeed, constructing an informative dictionary is a key step for sparsity-based model. Once the atoms in a dictionary are presented in an optimal manner we are able to present the given test pixel while significantly reduce the computational time load. Hence, in order to ensure sufficient number of useful bases for sparse representation we apply PCA via Eigen Value Decomposition (EVD) on sub-dictionaries and the main components of all sub-dictionary concatenated together. Another words, we construct a full rank matrix for each individual sub-dictionary that presents only the main information of its class members. Therefore, we are able to transfer the data form a complex higher dimensional set to a lower and informative dimension in which the samples of each individual class is orthogonal set (linearly independent set). Having use of EVD allows us to construct a well informative dictionary that leads to the high level of sparsity. The dictionary construction procedure using EVD goes as follows;

Consider a dictionary $D \in R^{B \times d}$ that contains three sub-dictionary given by

$$D = [D_1, D_2, D_3] \quad (4.9)$$

where each sub-dictionary may have different number of sample that represent a specific class. Now let us consider one of these classes e.g. D_1 that has 2500 samples in B dimension, which is a huge amount of data to represent. In addition, hyperspectral images are prone to have mix-pixel, which makes it difficult to present the end member. Therefore, this idea comes from where that PCA utilized for dimensionality reduction and help to a better representation of the state of the art endmember that is also a common task in hyperspectral pix-unmixing, which is also called the intra-class variability problem (Andreou, C., Karathanassi, V, 2011; Deville, Yannick., Revel, C., Achard, V., Briottet, X, 2018). This motivation encourages us unlike the other research works in construction of dictionary to perform a PCA based eigenvalue on individual sub-dictionary and hence reduce the number samples and present the basses representation for each class. Hence, we would like to transfer $D_1 \in R^{B \times d}$ to $D_1 \in R^{B \times k}$, where $k < d$, and this process is generalized for all sub dictionaries. In order to reduce the number of training sample to a reasonable amount, we compute the eigenvalues and eigenvectors of the covariance matrices computed form each sub-dictionary. In the next stage, the eigenvalues on the diagonal sorted in a descending manner along with their corresponding eigenvalue and the projection done on the few first components according to the magnitude of their eigenvalues in which we could retain more than 95 present of the variation for each class. Figure 5.5 demonstrate the diagram of Geometric dictionary construction.



5.5. Geometric Dictionary construction approach.

5.4. Experimental Result

In this section, the proposed algorithm is applied on the Indian Pines dataset in each stage of its development. In the first stage, Iterative Soft –Shrinkage Thresholding (ISST) is applied and the result presented both in terms of accuracy and computational time. In the second stage, the steepest is contributes in order to accelerate the ISST and deal with large coefficients. Lastly, the Joint Sparsity measurement is applied on the given dataset in order to identify the corresponding sub-dictionary for the given test pixel. Figure 5.6 demonstrates the classification task on hyperspectral image via proposed efficient sparse signal recovery algorithm.

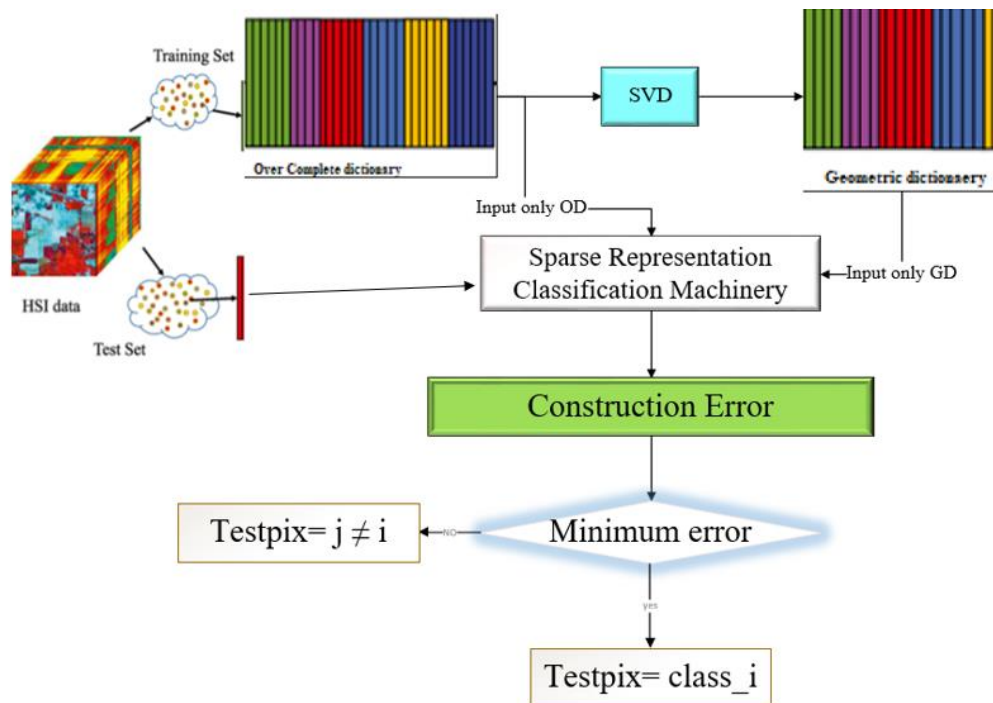


Figure 5.6. Illustrate the experimental design on hyperspectral image via proposed efficient sparse signal recovery. As it shown, the given test sample classify based on the minimum residual of the reconstruction error called residual.

5.4.1. l_1 Sparse recovery via Soft-Shrinkage Iteration.

As we discussed in chapter 4, we first start with an easy scheme to implement that leads to the sparsity measurement of the recover coefficients. Algorithm 1 represents the steps of implementation of Iterative Soft-Shrinkage Thresholding for hyperspectral image classification.

Algorithm 1. l_1 Sparse Recovery by Iterative Soft-Shrinkage

To solve:

$$\min_{\alpha} \|x^\delta - D\alpha\|^2 + \lambda \|\alpha\|_1 \quad (4.8)$$

Input: (a) Geometric Dictionary ($D \in \mathbb{R}^{B \times d}$), with normalized sample to have unit l_2 -norm (eq. 5.6), (b) Test sample $x \in \mathbb{R}^{B \times 1}$, (c) Threshold $\lambda > 0$, (d) Number of iteration,

(e) $C = \max(|eigs(D' * D)|)$;

Initialization: coefficient $\alpha \in \mathbb{R}^{d \times 1} = \mathbf{0}$

Step1: Compute the coefficient (gradient of first term in 4.8)

$$\alpha = (D^T(x^\delta - D\alpha)) ./ C$$

Step 2: In each iteration update α via Soft shrinkage

$$S_1(\alpha, \lambda) = \begin{cases} \mathbf{0}, & \text{if } |\alpha| \leq \lambda \\ \text{sgn}(\alpha)(|\alpha| - \lambda), & \text{if } |\alpha| > \lambda \end{cases}$$

Step 3: Update α until convergence

$$\alpha^{n+1} = S_{\lambda/2}(\alpha^n + (D^T(x^\delta - D\alpha)) ./ C) \quad (4.10)$$

Step 4: Compute the residual of each sub dictionary to assign test pixel to its class.

$$\text{class}(x) = \arg \min r_j(x) = \arg \min \|x - D_j \alpha_j\|_2, \quad j = 1, 2, 3, \dots, C \quad (4.6).$$

Step 5: Output label(x).

The result for the first step is presented in table 5.2.

Table 5.2. Performance of the ISST algorithm			Parameter	
Performance	Geometric Dictionary	Over complete dictionary (low rank matrix)		
Accuracy in percentage	93%	73%	Number of Iteration	150
Computation Time	1700 sec/28 min	12720 sec/212min		
Number of training sample	1455	1455	Threshed	0.1
Number of Atoms	20	1455		
Number of test sample	623	623		

As can be seen from the table, the performance of the geometric dictionary is significantly higher than the performance of the over complete dictionary. Indeed, both accuracy and computational time present a much better output using geometric dictionary than over complete dictionary.

5.4.2. l_1 Constrained Recovery via Projected Steepest Descent Iteration.

In this stage, we inject a steepest descent algorithm in the soft shrinkage Iteration in order to converge faster and avoiding bias in terms of large coefficients. Algorithm 2 describes the steps of implementation of l_1 Constrained Recovery via Projected Steepest Descent Iteration for hyperspectral image classification. The result for the second step presented in table 5.3.

Table 5.3. Performance of the ISSTSD algorithm			Parameter	
Performance	Geometric Dictionary	Over complete dictionary (low rank matrix)		
Accuracy in percentage	93%	75%	Number of Iteration	120
Computation Time	1360 sec/22 min	12300 sec/205min		
Number of training sample	1455	1455	Threshold	0.1
Number of Atoms	20	1455		
Number of test sample	623	623		

According to Table 5.3, the result of the geometric dictionary is again significantly higher than the result of the over complete dictionary. The over complete dictionary is also called low rank matrix

and presents a set of atoms that are linearly dependent. Therefore, finding the relevant atoms needs lots of computation and the chance of having the sparsest coefficients is very low.

Algorithm 2. ℓ_1 Constrained Recovery via Projected Steepest Descent Iteration

To solve:

$$\min_{\alpha \in B_k(\ell_1)} \|x^\delta - D\alpha\|^2 \quad (4.11)$$

Input: (a) Geometric Dictionary ($D \in \mathbb{R}^{B \times d}$), with normalized sample to have unit ℓ_2 -norm (eq. 5.6), (b) Test sample $x \in \mathbb{R}^{B \times 1}$, (c) Threshold $\lambda > 0$, (d) Number of iteration. (e) $C = \max(|eigs(D' * D)|)$;

Initialization: coefficient $\alpha \in \mathbb{R}^{d \times 1} = \mathbf{0}$, $step\ length(\beta) = 1$, $C1 = C * steplength$

Step1: Compute the coefficient (gradient of the 4.11)

$$\alpha = (D^T(x^\delta - D\alpha)) ./ C$$

Step 2: In each iteration update α via Soft shrinkage

$$\mathbb{S}(\alpha, \lambda) = \begin{cases} \mathbf{0}, & \text{if } |\alpha| \leq \lambda \\ \text{sgn}(\alpha)(|\alpha| - \lambda), & \text{if } |\alpha| > \lambda \end{cases}$$

Step 3: In each iteration Check whether;

$$Check = C|\alpha_{n+1} + \mathbf{1} - \alpha_n|^2 - \|D(\alpha_{n+1} - \alpha_n)\|^2 > 0$$

$$P_{B_k(\ell_1)} = \begin{cases} \beta = \beta * 0.8, \text{ and } C1 = C * \beta & \text{if } Check > 0 \\ \beta, \text{ and } C & \text{if } Check \leq 0 \end{cases}$$

Step 4:

Repeat step 1, 2, and 3 until convergence

$$\alpha^{n+1} = P_{B_k(\ell_1)}(\alpha^n + \beta^n D^T(x^\delta - D\alpha^n)) \quad (4.12)$$

Step 5: Compute the residual of each sub dictionary to assign test pixel to its class.

$$class(x) = \arg \min r_j(x) = \arg \min \|x - D_j \alpha_j\|_2, \quad j = 1, 2, 3, \dots, C \quad (4.6).$$

Step 6: Output label(x).

5.4.3. Joint Sparsity Measure Recovery via Projected Steepest Descent iteration.

Eventually, we end up with the proposed optimization strategy joint sparsity measurement that is able to select the relevant dictionary in a unique way. Algorithm 3 shows the steps of the implementation of Joint Sparsity Measure Recovery algorithm for hyperspectral image classification. The result for the second step presented in Table 5.4.

Table 5.4. Performance of the JSM algorithm			Parameter	
Performance	Geometric Dictionary	Over complete dictionary (low rank matrix)		
Accuracy in percentage	98%	80%	Number of Iteration	90
Computation Time	1058 sec/17 min	12000sec/200min		
Number of training sample	1455	1455	Threshold	0.1
Number of Atoms	20	1455		
Number of test sample	623	623		

As shown in Table 5.4, the resulting output for both dictionary is still significantly different, where the Geometric dictionary has a great impact on the representation of the relevant atoms (training sample) by presenting a set of redundant atoms that contains the main information of the given training set. Moreover, it should be mentioned that in this thesis only spectral information has been employed to construct the dictionary. It is possible to fuse the spatial information with spectral information and produce a spatial-spectral dictionary that may provide more discriminant details on individual classes.

Algorithm 3. Joint Sparsity Measure Recovery via Projected Steepest Descent iteration

To solve:

$$\min_{\alpha \in B_k(\ell_1)} \|x^\delta - D\alpha\|^2 + \lambda \sum_{K=1}^C \|\alpha_k\|_2 \quad (4.13)$$

Input: (a) Geometric Dictionary ($D \in \mathbb{R}^{B \times d}$), with normalized sample to have unit ℓ_2 -norm (eq. 5.6), (b) Test sample $x \in \mathbb{R}^{B \times 1}$, (c) Threshold $\lambda > 0$, (d) Number of iteration. (e) $C = \max(|\text{eigs}(D' * D)|)$, (f) N , an array of number of sample in each sub-dictionary.

Initialization: coefficient $\alpha \in \mathbb{R}^{d \times 1} = \mathbf{0}$, $\text{step length}(\beta) = 1$, $C1 = C * \text{steplength}$

Step1: Compute the coefficient (gradient of the first term in 4.8)

$$\alpha = (D^T(x^\delta - D\alpha)) ./ C$$

Step 2: In each iteration update α via Soft-block shrinkage

$$S(\alpha, \lambda, N) = \begin{cases} \mathbf{0}, & \text{if } \|\alpha_c\|_2 \leq \lambda \\ \alpha_c - \lambda * \alpha_c / \|\alpha_c\|_2, & \text{if } \|\alpha_c\|_2 > \lambda \end{cases}$$

Step 3: In each iteration Check whether;

$$\begin{aligned} \text{Check} &= C \|\alpha_{n+1} + \mathbf{1} - \alpha_n\|^2 - \|D(\alpha_{n+1} - \alpha_n)\|^2 > 0 \\ P_{B_k(\ell_1)} &= \begin{cases} \beta = \beta * 0.8, \text{ and } C1 = C * \beta & \text{if } \text{Check} > 0 \\ \beta, \text{ and } C & \text{if } \text{Check} \leq 0 \end{cases} \end{aligned}$$

Step 4:

Repeat step 1, 2, and 3 until convergence

$$\alpha^{n+1} = P_{\bar{B}_k(\ell_1)} \left(\alpha^n + \beta^n D^T(x^\delta - D\alpha^n) \right) \quad 4.15$$

Step 5: Compute the residual of each sub dictionary to assign test pixel to its class.

$$\text{calss}(x) = \arg \min r_j(x) = \arg \min \|x - D_j \alpha_j\|_2, \quad j = 1, 2, 3, \dots, C \quad (4.6).$$

Step 6: Output label(x).

5.5. Result and Discussion

In this section, the result of all three stages of the development presented and compared in terms of accuracy and computational time. The main aim of performing such experimental design is to check whether our proposed schema works fine in the two mentioned aspects (accuracy and computational time). Indeed, we expect having a satisfactory accuracy (almost same accuracy in each stage) and improvement in the computational time after injecting the Steepest Descent. After running the algorithm for each stage of its development, we bring all the results of each stage of the development in table 5.5.

Table 5.4. Performance comparison for each stage of the development

Stage 1			Stage 2			Stage 3		
ISST	Geometric Dictionary	Over complete dictionary	ISSTSD	Geometric Dictionary	Over complete dictionary	JSM	Geometric Dictionary	Over complete dictionary
Accuracy	93%	73%	Accuracy	93%	75%	Accuracy	98%	80%
Computation Time	1700 sec	12720 sec	Computation Time	1360 sec	12300 sec	Computation Time	1058sec	12000sec sec
Number of training sample	1455	1455	Number of training sample	1455	1455	Number of training sample	1455	1455
Number of Atoms	20	1455	Number of Atoms	20	1455	Number of Atoms	20	1455
Number of test sample	623	623	Number of test sample	623	623	Number of test sample	623	623
Number of iteration	150		Number of iteration	120		Number of iteration	90	
Threshold	0.1		Threshold	0.1		Threshold	0.1	
Step length β	0.8		Step length β	0.8		Step length β	0.8	

As it can be seen, we have achieved a significant accuracy for classifying four classes, including corn, grass-pasture, woods, and stone-steel-towers. Furthermore, the computation time has been significantly improved after injecting the steepest descent. Moreover, in the last step, the block sparsity measurement makes it much easier to identify the relevant classes for the given dictionary. We expect of having almost the same accuracy in each stage that is shown in table 5.5 that the resulting output for geometric dictionary is much higher than the over complete dictionary. Moreover, we see that the computational time effectively reduced after porting the Steepest Descent into the Iterative Soft-Shrinkage algorithm. In addition, we could reach the minimum residual in a very fast convergence manner, which firstly shows the power of Iterative Soft-Shrinkage and secondly the efficient development of implemented soft shrinkage algorithm via Steepest Descent. The figure 5.6 illustrate the cost function for 250 iterations in one test sample.

Chapter 6

6.1. Summary

The main objective of this thesis is to develop a classification principle using the sparsity-based model that can deal with big data with the application in classification of hyperspectral imagery data. The motivation of this thesis comes from where that the sparsity based model is an efficient tool for extracting the latent structure in hyperspectral images. Recently sparse representation has gained a great attention in remote sensing community. In most of the literatures in remote sensing domain sparse coding is only utilized and the development of the numerical solution for sparsity based model is less considering and the main focus is on the presenting/constructing a dictionary that can discriminately represent different classes. Some examples of this dictionary representation in remote sensing domain to be mentioned are special dictionary and spatial-spectral dictionary. In the second approach the spatial information in the image utilized along with the spectral information in order to give a better presentation for the test image. Many efforts have been made by the researchers for developing efficient mathematical procedures to solve the optimization problem in sparsity based model. Hence, the focus is on the formulation of the optimization functional. For sparse coding, many algorithms have been developed to solve the non-smooth optimization problem (least squares with l_1 -norm). Of those all, greedy algorithm such as OMP can be mentioned. Other methods such as IRLS, BP have been also proposed to find the solution which minimize the objective functional, but it turns out that these algorithms need a lot of iterations and computations to converge. Therefore, a new numerical solution called Iterative Shrinkage Thresholding (IST) that has been motivated by Donoho-Johnston shrinkage method built to address the problem of dealing with high dimensionality for big data. This algorithm is a proximity algorithm that separates the non-differentiable part of the objective function from the convex part. Hence, the minimization of the convex part becomes easy with a global rate and the penalty term transformed to a shrinkage operation that is controlled by a threshold. Despite of the suite operation of shrinkage there is still obstacle for achieving a reliable convergence. The obstacle is the universal thresholding for the optimality condition where it cannot deal with large coefficients and needs many iterations to sparsity the recover information and obtain a reliable result. It is for this reason, in this thesis after implementing the Iterative Soft Shrinkage Thresholding (ISST) algorithm, we inject Steepest Descent into the algorithm that can deal with bias in the estimation of the coefficients via a step length in which this also lead to an accelerated version of the ISST. Eventually, we present a joint sparsity measurement comprising of the two previous steps and a new feature as optimization function that computes the norm of each sub-dictionary in each iteration and via an optimality condition set the irrelevant dictionary to zero that leads to a block sparsity measurement. This approach is even able to uniquely identify the relevant dictionary for the given test pixel. The procedure for the construction of the dictionary is done by firstly extracting the corresponding pixel for each class (a prior-knowledge of the membership for each class). The extracted spectral feature only inferring the spectral information in the image and spatial information is not considered. The quest of having a well-structured dictionary encourages us to construct a geometric dictionary via singular value decomposition (SVD). Indeed, the SVD

has been implemented to be applied on each sub-dictionary in order to remove the redundant atoms and only present the main information in each class. Furthermore, we also consider all existing training samples for constructing the dictionary and this dictionary called is over complete dictionary which has the property of low rank matrix with a high degree of linear dependency between the atoms. The proposed algorithm has been applied separately on Geometric and Over-complete dictionary to check whether the proposed Geometric dictionary works well on our developed algorithm. The evaluation of the proposed sparse signal recovery has been done in each stage of its development and the result for both over complete and geometric dictionary has been presented in Chapter 5 Table 5.4. The result shows that with Geometric dictionary only a few sample (e.g. the first five Principal components space) needed to be represented in each sub-dictionary for the construction of the main dictionary. Hence, the amount of atoms has been reduced to a lower number where we could also sparsify the recovered coefficients at most. While in over complete dictionary case the computation time compared to the Geometric dictionary is significantly higher and the accuracy is very low and even the amount of sparse coefficient is much less than the case, using Geometric dictionary. Apart from the dictionary representation, our implementation of the ISST algorithm gives a significant accuracy for the classification of the four given classes including Corn, Grass-Pasture, woods, and Stone-Steel-Towers and the development of the ISST in each stage demonstrate a significant enhancement in the computational time while preserving the accuracy at best. The proposed efficient sparse signal recovery contains bunch of parameters that one needs to adjust them to get the best output. The parameters are including, threshold rate, number of iteration, step length of gradient, and the number of atoms representation (dictionary construction which is a long story itself), but when a good presentation of the feature dataset is provided then, by adjusting the parameters of the model, the result can be surprisingly reliable for even more complex dataset.

6.2. Conclusion

Our experiment on the four relatively close spectral classes of AVIRIS sensor in Indian Pines data set gives us a promising result for our proposed algorithm for the big data classification, particularly for hyperspectral image classification. The experimental design for each stage of the development present an instruction for the proper using of this package for the future researches and even using for solving the real world problems. Indeed, after applying the ISST and then ISSTSD, and eventually joint sparsity measurement we have observed computational time and the accuracy of each stage in a fixed adjustment of the model parameters. The accuracy demonstrates almost the same result for each stage of the development while the number of required iteration for convergence is significantly decreased. Indeed, after reducing the number of iteration following up injecting the Steepest Descent, in the second stage of development the amount of computation time significantly reduced and the accuracy preserved almost the same as before with low number of iteration. Furthermore, the Geometric dictionary gives us also the ability to have an effective computational time and accuracy also more importantly sparsify the recovered coefficients at most. In other words, the proposed Geometric dictionary approach contributes to the general performance of the proposed algorithm.

This thesis addresses one of the open questions in big data mining and analysis by providing an efficient sparse signal recovery that can be used for classification task for higher dimensional dataset. In addition, we provide an insight to our proposed package for a proper usage. Lastly, by employing a concept from linear algebra, we create a geometric dictionary that can inspire the researchers for the extending this work.

6.3. Future Direction

There are several potential research directions based on this thesis that can contribute to build up new researches. The directions are as follows;

In Chapter 4, the dictionary construction can be extended by contributing spatial information along with spectral information. Indeed, the data fusion techniques can be used for construction of the dictionary, which has a significant impact on the performance of the proposed sparsity based algorithm in this thesis. One of the famous tools is the simple linear iterative clustering algorithm inspired by K-means algorithm in which the spatial and spectral information is captured for clustering. Hence, one may apply this algorithm before constructing the dictionary that also can solve the problem of atoms in the dictionary. In addition, the other one may be concerned about removing the redundant spectral bands before constructing the dictionary. Indeed, hyperspectral images contains overlapping spectral region that conveys almost the same information. Hence, the redundant dimensions can be sufficiently remove by maximizing the variance in the dataset using data mining techniques such as PCA.

This proposed algorithm in this thesis can be also examined by using the end members for both classification and regression task.

The proposed optimization algorithm can perform effectively once the issues of dictionary representation along well adjustment of the parameters solved effectively. Hence, one may study the analytical solutions for the optimal choice of the model parameters.

Lastly, due to the certain connections between sparse approximation and deep learning, the proposed variation regularization sparsity based model can also be extended into deep networks; one may expect faster inference, large learning capability, and better scalability.

Appendix:

Intuition and output of the proposed algorithm.

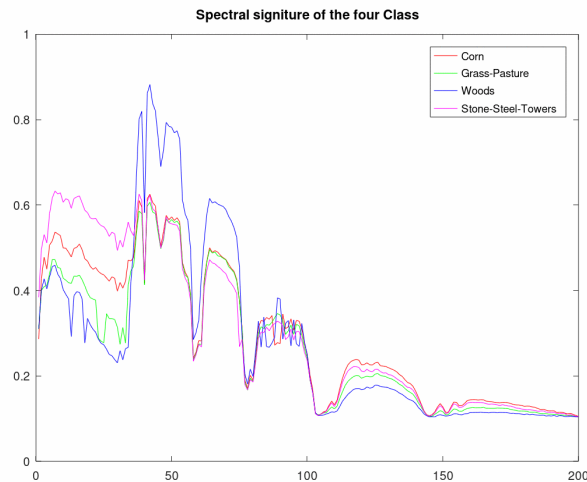


Figure 1. Presents the spectral signature of each class. As shown, the discrimination between these four classes are roughly represent in the first few bands.

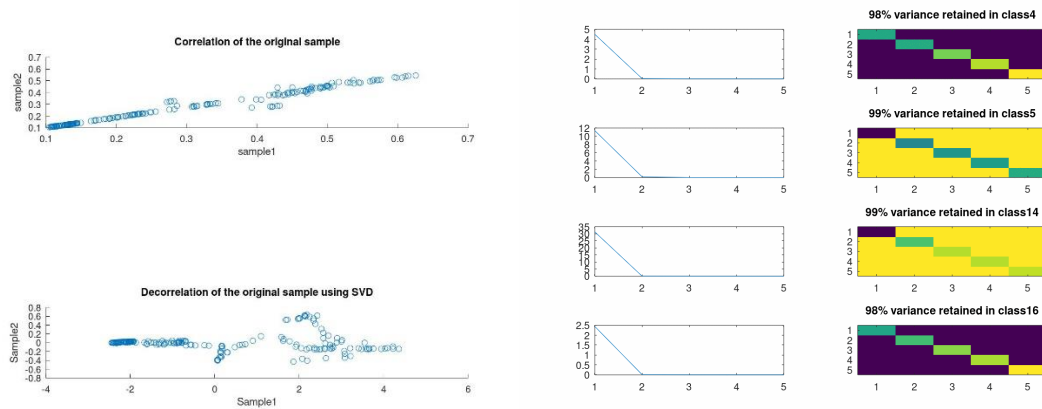


Figure 2. The left figure on the top presents the correlation between 2 sample from class 4, and in below the de-correlated of the redundant sample via SVD illustrated. The figure in the right shows the number of PCs space for each sub-dictionary. After reduction in almost all sub-dictionary more than 98 percentage of the variance retained. This shows the advantage of Geometric dictionary for better representation of the Atoms In each class that directly contributes to the result of proposed package.

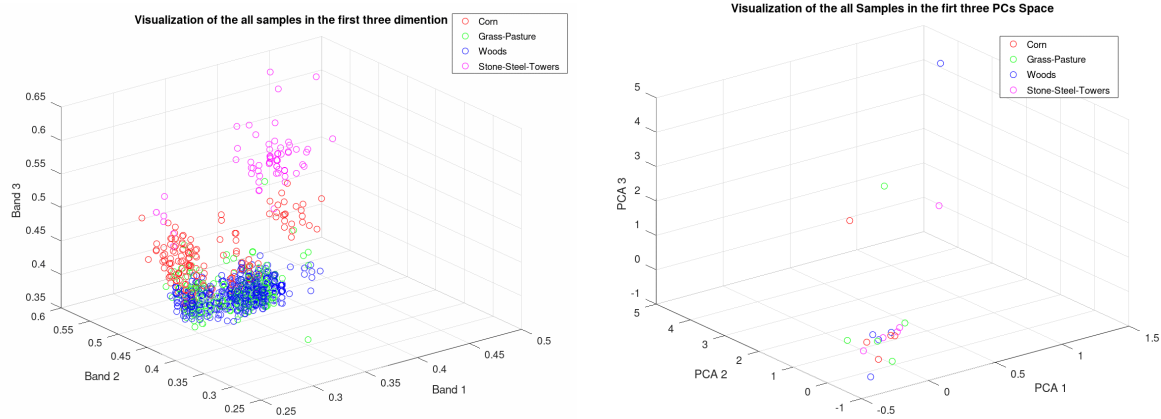


Figure 3. This figure visualizes the training sample in 3D. In the left the original data from four classes plot in the first three spectral bands. The right figure illustrates the first three PCs space of training sample. It can be seen that the number of training sample is significantly reduced. Although, some outliers can be seen in the data that are going to have impact on the output of the model. Nevertheless, we do not take any step to fix this problem. Furthermore, this concept also can be a remark for one who wants to use PCA base analysis that consider outlier as a critical issue for PCA base analysis. Hence, in the case where endmembers are existence SVD can preforms much better.

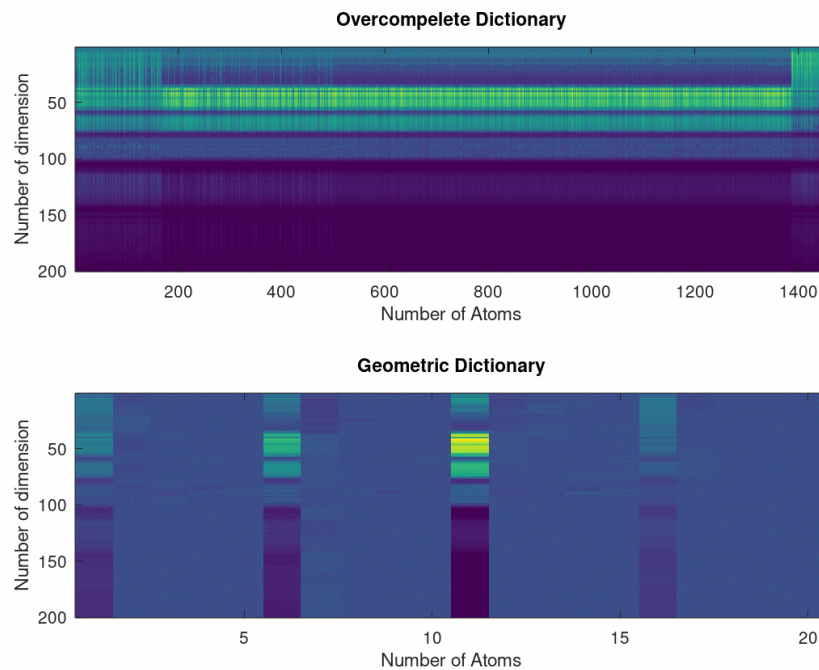


Figure 4. Depicts Over-complete and Geometric dictionary. As seen in the geometric dictionary, first Atoms are the reach information among all the training set. Note that in this thesis for Geometric dictionary we peek the first five principal components just to be in the safe side.

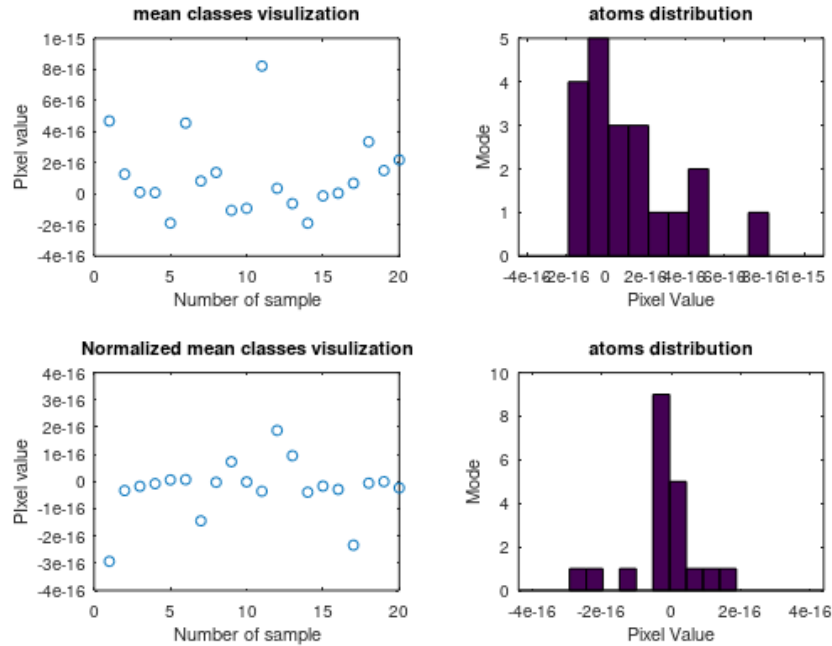
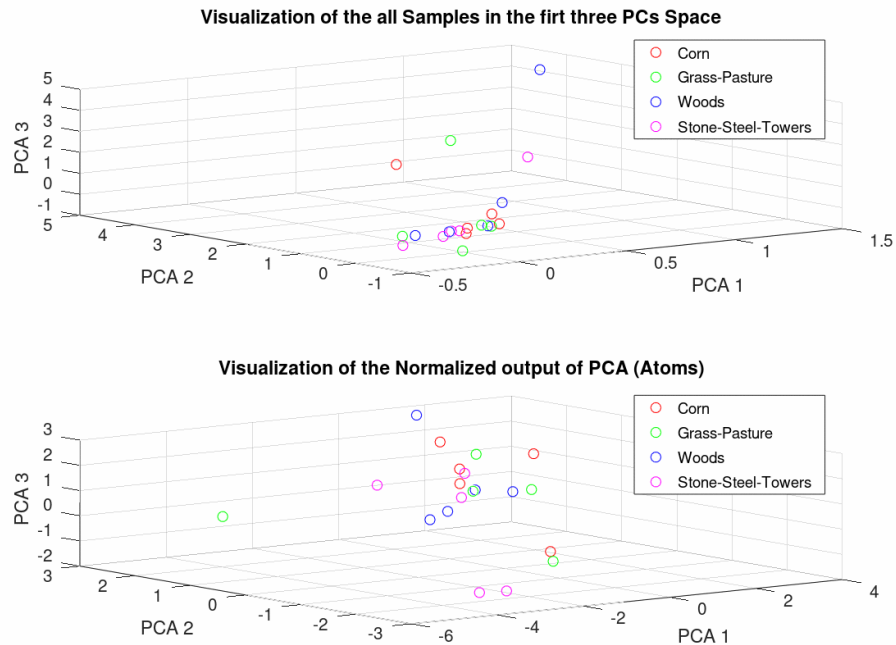


Figure 4. Shows the atoms in the top before normalization and the bottom is after normalization. This normalization turns the atoms to have unit l_2 -norm. Simply get rid of floating points that affect the approximation. Therefore the



The figure 5. Shows a remark that is normalizing the atoms right after SVD. As demonstrated the Normalized atoms after SVD gives a better-discriminated information on the classes.

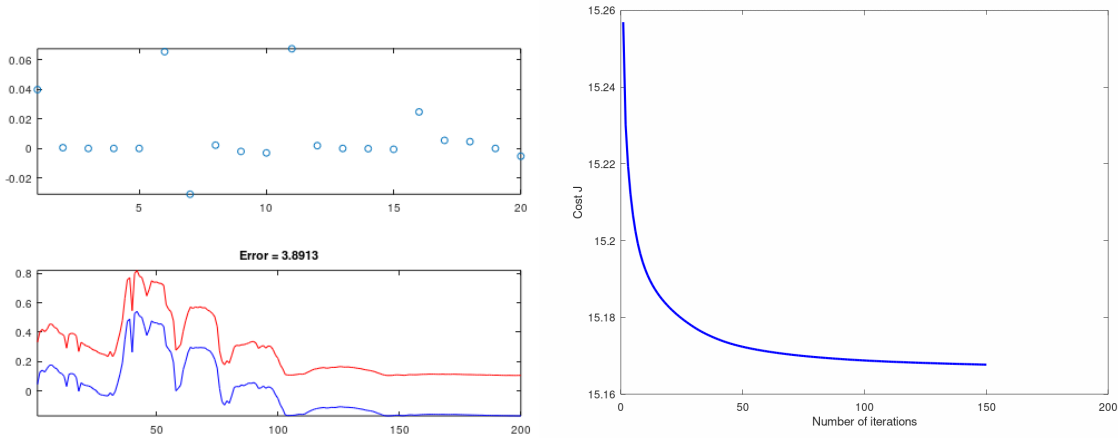


Figure 1. Demonstrate the output of the Iterative soft-Shrinkage algorithm. The left side of the figure depict the sparse solution for the given test pixel via our scratch implementation of iterative soft-shrinkage algorithm. The convergence rate is even less than the number of iteration. The classification of given test pixel failed to identify the real class. Nevertheless, there is also the case, one reduce the number of iteration and may obtain the right answer from this machinery.

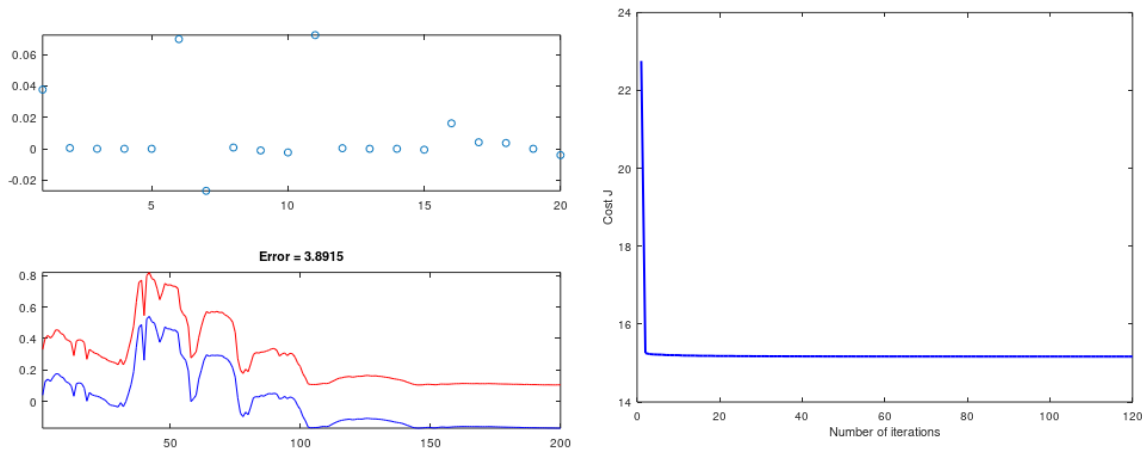


Figure 1. Demonstrate the output of the optimize version of Iterative soft-Shrinkage algorithm (ISSA) with steepest descent. It can be seen that within the developed version of iterative soft-shrinkage the result remains the same but the convergence (minimization of the objective function) is significantly increase. In addition, the number of sparse coefficients remains the same like ISSA while the number of iteration set from 150 to 130 in this stage.

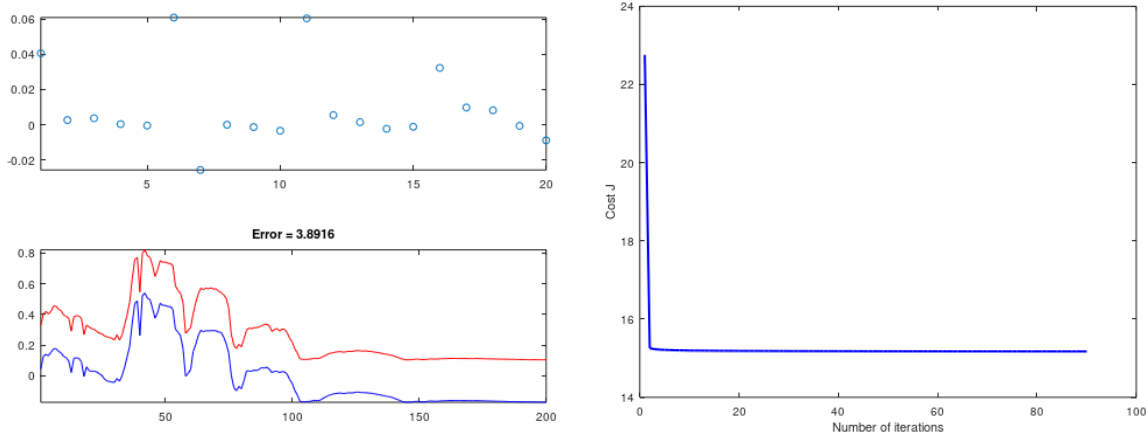


Figure1. Present the output of the proposed efficient signal recovery called Joint Sparsity Measurement (JSM). In the left side, the sparsity of the model remains same while an interesting competition starts between the two effective coefficients on the top of the graph. This leads to the wining of the classification task. Furthermore, the number of iteration is even increase in this stage to check the promise of our algorithm (fast convergence and promote a reliable accuracy). Indeed, the number of iteration decrease from 150 to 90.

References

- A. Green, M. Berman, P. Switzer and M. Craig. (1998). A transformation for ordering multispectral data in terms of image quality with implications for noise removal,. *Geoscience and Remote Sensing*, 65–74. Retrieved from <https://pdfs.semanticscholar.org/6ae0/0ebd3a91c0667c79c39035b5163025bfcad.pdf>
- Abdi, H., & Williams, L. J. (2010). Principal component analysis', *Wiley Interdisciplinary Computational Statistics*, 433-459. Retrieved from <https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.101#accessDenialLayout>
- Ahmed , A-M., Duran, O., Zweiri, Y., Smith, M. (2017). Hybrid Spectral Unmixing: Using Artificial Neural Networks for Linear/ Non-Linear Switching. *Remote Sensing*, 1-22. Retrieved from <https://doi.org/10.3390/rs9080775>
- Amaldi, E., and Kann, V. (1998). On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 237–260.
- Anand, R., Veni, S., Aravinth, J. (2017). Big Data Challenges in Airborne Hyperspectral Image for Urban Landuse Classification. *International Conference on Advances in Computing, Communications and Informatics* (pp. 2-8). Udupi, India: IEEE. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8126107/authors#authors>
- Anaraki, F.B., Hughes S.M. (2013). Compressive K-SVD. *International Conference on Acoustics, Speech and Signal Processing* (pp. 1-6). Vancouver, BC, Canada: IEEE. Retrieved from <https://ieeexplore.ieee.org/document/6638709/authors#authors>
- Andreou, C., Karathanassi, V. (2011). Using principal component analysis for endmember extraction. *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*. (pp. 1-6). Lisbon, Portugal: IEEE. Retrieved from <https://ieeexplore.ieee.org/document/6080955>
- B. K. Natarajan. (1995). Sparse approximate solutions to linear systems of equation. *SIAM journal on computing*, 227-234. Retrieved from <https://pdfs.semanticscholar.org/f629/5fd69d76d606f66cc15f58767a8161d60335.pdf>
- Baraniuk, R.G., Candes, E., Elad, M., and Ma, Y. (2010). Applications of sparse representation and compressive sensing. . *Proceedings of the IEEE*, 906–909. Retrieved from <https://ieeexplore.ieee.org/document/5466604>
- Bellman, R. (1956). Dynamic programming and lagrange multipliers. *Proceedings of the National Academy of Sciences of the USA.*, 767-9. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC528332/?page=1>
- Bian X., Chen C., Xu Y., Du Q. (2016). Robust Hyperspectral Image Classification by Multi-Layer Spatial–Spectral Sparse Representations. *Remote Sensing*, 1-24. Retrieved from <https://doi.org/10.3390/rs8120985>
- Bian, X., Zhang, T., Yan, L., Zhang, X., Fang, H., Liu, H. (2013). Spatial–spectral method for classification of hyperspectral images. *Opt. Lett.*, 815-817. Retrieved from <https://www.osapublishing.org/ol/viewmedia.cfm?uri=ol-38-6-815&seq=0>
- Bioucas-Dias, J-M., Plaza, A., Dobigeon, N, Parente, M., Du, Q., Gader, P., Chanussot, J. (2012). Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based

- Approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 354 - 379. Retrieved from 10.1109/JSTARS.2012.2194696
- Bruckstein, A.M., Donoho D. L., and Elad M. (2009). sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 34-81. Retrieved from <https://dl.acm.org/citation.cfm?id=1654037>
- Chan, S.S., Donoho, D.L., and Saunders, M.A. (2001). *Atomic Decomposition by Basis Pursuit*. Stanford: SIAM Review. Retrieved from <https://dl.acm.org/citation.cfm?id=588850>
- Chang, C.-I. (2013). *Hyperspectral data processing : algorithm design and analysis*. the United States of America: JohnWiley & Sons, Inc.
- Chen C., Chen N., Peng J. (2016). Nearest Regularized Joint Sparse Representation for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens*, 424-428. Retrieved from <https://ieeexplore.ieee.org/document/7395320>
- Chen Y., Nasrabadi N.M., Tran T.D. (2013). Hyperspectral image classification via kernel sparse representation. *IEEE Trans. Geosci. Remote Sens*, 217–231. Retrieved from <https://ieeexplore.ieee.org/document/6236130>
- Chen, S.S., Donoho, D.L., Saunders, M.A. (2001). Atomic Decomposition by Basis Pursuit. *SIAM review*, 33-61.
- Chen, Y., Nasrabadi, N.M., Tran, T.D. (2011). Hyperspectral image classification using dictionary-based sparse. *IEEE Trans. Geosci. Remote Sens*, 3973-3985. Retrieved from 10.1109/TGRS.2011.2129595
- Chen, Y., Nasrabadi, N.M., Tran, T.D. (2011). Hyperspectral image classification using dictionary-based sparse representation. *IEEE Trans. Geosci. Remote Sens.*, 3973–3985. Retrieved from 10.1109/TGRS.2011.2129595
- Deville, Yannick., Revel, C., Achard, V., Briottet, X. (2018). Application and Extension of PCA Concepts to Blind Unmixing of Hyperspectral Data with Intra-class Variability. In Y. R. Deville, *Advances in Principal Component Analysis* (pp. 225-252). Toulouse, France.: Springer.
- Dias, J.M.B., Plaza, A., Valls, G.C., Scheunders, P., Nasrabadi, N., Chanussot, J. (2013). Hyperspectral Remote Sensing Data Analysis and Future Challenges. *IEEE Geoscience and Remote Sensing Magazine*, 6 - 36. Retrieved from 10.1109/MGRS.2013.2244672
- Donoho, D. (1995). De-noising by soft-thresholding. *IEEE*, 613–627. Retrieved from <https://ieeexplore.ieee.org/document/382009>
- Donoho, D. (2006). For most large underdetermined systems of linear equations the minimal ℓ_1 -norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 797–829. Retrieved from <https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.20132>
- Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 1289–1306. Retrieved from <https://dl.acm.org/citation.cfm?id=2272089>
- Donoho, D-L., I-M-J Biometrika. (1994). Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, 425-455. Retrieved from <https://www.jstor.org/stable/2337118>

- Du, Q., Raksuntorn, N., Younan, N.H., King, R.L. (2008). End-member Extraction for Hyperspectral image Analytsis. *Applied Optics*, F77-F84. Retrieved from <https://www.osapublishing.org/ao/abstract.cfm?uri=ao-47-28-f77>
- Eerens, H., Haesen, D., Rembold, F., Urbano F., Tote C., Bydekerke, L. (2014). Image time series processing for agriculture monitoring. *Environmental Modelling & Software*, 154-162. Retrieved from <https://doi.org/10.1016/j.envsoft.2013.10.021>
- El_Rahman, S. (2016). Hyperspectral Image Classification Using Unsupervised Algorithms. *International Journal of Advanced Computer Science and Applications*, 198-205. Retrieved from <https://thesai.org/Publications/ViewPaper?Volume=7&Issue=4&Code=IJACSA&SerialNo=25>
- Elad, M. (2013). *Sparse and Redundant Representations*. Haifa, Israel: Springer.
- Engan, K., Aase, S.O., Husoy, J.H. (1999). Method of optimal directions for frame design. *International Conference on Acoustics* (pp. 2443-2446). Phoenix, AZ, USA: IEEE. Retrieved from <https://ieeexplore.ieee.org/document/760624/authors#authors>
- Fauvel, M., Benediktsson, J. A., Chanussot, J., Sveinsson, J. R. (2008). Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 3804–3814.
- Feng Z., Yang M., Zhang L., Liu Y., Zhang D. (2013). Joint Discriminative Dimensionality Reduction and Dictionary Learning for Face Recognition. *Pattern Recognition*, 2134-2143. Retrieved from <https://doi.org/10.1016/j.patcog.2013.01.016>
- Figueiredo, M.A.T., Nowak, R.D., Wright, S.J. (2007). Gradient Projection for Sparse Reconstruction: Application to Compressed Compressed Sensing and Other Inverse Problems. *IEEE*, 586 - 597. Retrieved from <https://ieeexplore.ieee.org/document/4407762>
- Fletcher, R. (2013). *Practical Methods of Optimization, Second Edition*. Scotland: John Wiley & Sons, Ltd. Retrieved from <https://onlinelibrary.wiley.com/doi/book/10.1002/9781118723203>
- Fornasier, M., Peter S. (2015). An Overview on Algorithms for Sparse Recovery. In M. P. Fornasier, *An Overview on Algorithms for Sparse Recovery* (pp. 1-74). munich: semantic scholar. Retrieved from <https://www.semanticscholar.org/paper/An-Overview-on-Algorithms-for-Sparse-Recovery-Fornasier-Peter/c028b61164eac11720708234dfe7dedc5b738a73?navId=paper-header>
- Geladi, L.M.P., Grahn, H.F., Burger J.E. (2007). *Multivariate Images, Hyperspectral Imaging: Background and Equipment*. Southern Gate: John Wiley & Sons, Ltd. . Retrieved from <https://doi.org/10.1002/9780470010884.ch1>
- Gill, P.R., Wang A., Molnar, A. (2010). The In-Crowd Algorithm for Fast Basis Pursuit Denoising. *IEEE*, 4595 - 4605. Retrieved from <https://ieeexplore.ieee.org/document/5940245>
- Gislason, P.O., Benediktsson J.A. (2006). Random Forests for Land Cover Li et al. *Pattern Recognition Letters*, 294-300.
- H. Cheng., Z. Liu., L. Yang., and X. Chen. (2013). Sparse representation and learning in visual recognition : Theory and applications. . *Signal Processing*, 1408–1425.

- Hao S., Wang, W., Bruzzone, L. (2017). Class-wise dictionary learning for hyperspectral image classification. *Elsevier*, 121-129. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0925231216309067>
- Hu, L., Qi, C., Wang, Q. (2018). Spectral-Spatial Hyperspectral Image Classification Based on Mathematical Morphology Post-Processing. *International Conference on Identification, Information and Knowledge in the Internet of* (pp. 93–97). Qufu,China: Procedia Computer Science.
- Huang A, Zhang H , Pižurica A. (2017). A Robust Sparse Representation Model for Hyperspectral Image Classification. *Sensors*, 1-18. Retrieved from <https://pdfs.semanticscholar.org/b955/358f273cd4b64dff1d99ccd702755bd373a.pdf>
- Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 55–63.
- Hyvärinen, L. (1970). Principal component analysis. *Mathematical Modeling for Industrial*, 82-104.
- Iordache, M-D., Bioucas-Dias, J., Plaza, A. (2011). Sparse Unmixing of Hyperspectral Data. *IEEE*, 2014-2039. Retrieved from <https://ieeexplore.ieee.org/document/5692827>
- J. A. Tropp, A. C. Gilbert, and M. J. Strauss. (2006). Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. . *Signal Processing*, 572–588.
- J. Bioucas-Dias and J. Nascimento. (2008). Hyperspectral subspace identification. *Geoscience and Remote Sensing, IEEE.*, 2435–2445. Retrieved from <https://ieeexplore.ieee.org/document/4556647>
- Jolliffe, I. (2002). *Principal Component Analysis, Second Edition* . New York : Springer .
- Kowalski, M. (2015). Thresholding RULES and iterative shrinkage/thresholding algorithm: A convergence study. *International Conference on Image Processing (ICIP)* (pp. 1-6). Paris, France: IEEE. Retrieved from <https://ieeexplore.ieee.org/abstract/document/7025843/authors#authors>
- Kowalski, M. (2015). Thresholding RULES and iterative shrinkage/thresholding algorithm: A convergence study. *International Conference on Image Processing* (pp. 1-6). Paris, France: IEEE. Retrieved from <https://ieeexplore.ieee.org/abstract/document/7025843/authors>
- Li, M., Zang, S., Zhang, B., Li, S., Wu, C. (2014). A Review of Remote Sensing Image Classification Techniques: the Role of Spatio-contextual Information. *An official journal of the Italian Society of Remote Sensing*, 389-411. Retrieved from <https://www.tandfonline.com/doi/abs/10.5721/EuJRS20144723>
- Li, W., Prasad, S., Fowler, J.E., Bruce, L.M.,. (2011). Locality-Preserving Dimensionality Reduction and Classification for Hyperspectral Image Analysis. *Geoscience and Remote Sensing*, 1185 - 1198. Retrieved from 10.1109/TGRS.2011.2165957
- Li, Y., Wu, Zebin Wu., Wei J., Plaza, A., Li, J., Wei Z. (2015). Fast principal component analysis for hyperspectral imaging based on cloud computing. *International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 513-516). Milan, Italy: IEEE. Retrieved from 10.1109/IGARSS.2015.7325813

- Liu W., Wen, Y., Li, H., Zhu, B. (2014). Dictionary construction for sparse representation classification: A novel cluster-based approach. *IEEE Symposium on Computers and Communications (ISCC)* (pp. 1-6). Funchal, Portugal: IEEE. Retrieved from <https://ieeexplore.ieee.org/document/6912545?denied=>
- M. Elad. (2010). *Sparse and redundant representations: from theory to applications in signal and image processing*. Haifa, Israel: Springer.
- M. Huang, W. Yang, J. Jiang, Y. Wu, Y. Zhang,. (2014). Brain extraction based on locally linear representation-based classification. *NeuroImage*, 322-339. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/24525169>
- Ma, W-K., Bioucas-Dias J.M., Chan, T-H., Gillis N., Gader, P., Plaza, A-J., Ambikapathi, A., Chi C-H. (2014). A Signal Processing Perspective on Hyperspectral Unmixing: Insights from Remote Sensing. *IEEE Signal Processing Magazine*, 67-81. Retrieved from <https://ieeexplore.ieee.org/document/6678258>
- Moroni, M., Lupo, E., Marra, E., & Cenedese, A. (2013). Hyperspectral Image Analysis in Environmental Monitoring: Setup of a New Tunable Filter Platform. *Procedia Environmental Sciences*, 885–894. Retrieved from https://www.researchgate.net/publication/270916981_Hyperspectral_Image_Analysis_in_Environmental_Monitoring_Setup_of_a_New_Tunable_Filter_Platform
- Mountrakis, G., Im, J. (2011). Support Vector Machines in Remote Sensing: A Review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 247-259.
- Mukherjee, S., Basu, R., Seelamantula, CS. (2016). ℓ_1 -K-SVD: A robust dictionary learning algorithm with simultaneous update. *Signal Processing*, 4252. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0165168415004351>
- Nascimento, José M. P., Bioucas-Dias, José M. (2012). Hyperspectral Unmixing Based on Mixtures of Dirichlet Components. *IEEE Transactions on Geoscience and Remote Sensing*, 863-878. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.386.5287&rep=rep1&type=pdf>
- Olshausen, B.A., Field, D.J. (1997). Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1 ? *Visin research*, 3311-3325. Retrieved from [https://doi.org/10.1016/S0042-6989\(97\)00169-7](https://doi.org/10.1016/S0042-6989(97)00169-7)
- Pal, M.; Foody, G. (2010). Feature Selection for Classification of Hyperspectral Data by SVM. *IEEE Trans. Geosci. Remote Sens*, 2297–2307. Retrieved from <https://ieeexplore.ieee.org/document/5419028>
- Parikh, N., and Boyd S. . (2013). Proximal algorithms. *Foundations and Trends in Optimization*, 123–231.
- Patel V. M., and Chellappa, R. (2014). Sparse representations, compressive sensing and dictionaries for pattern recognition. *First Asian Conference on Pattern Recognition* (pp. 325–329). Beijing, China: IEEE. Retrieved from <https://ieeexplore.ieee.org/document/6166711>

- Pati, Y.C., Rezaifar, R., Krishnaprasad PS. (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. *Annual Asilomar Conference on Signals Systems and Computers* (pp. 40-44). PACIFIC GROVE, CALIFORNIA: IEEE.
- Plaza, J., Hendrix E.M.T., García I., Martín, G., Plaza, A. (2012). On Endmember Identification in Hyperspectral Images Without Pure Pixels: A Comparison of Algorithms. *Mathematical Imaging and Vision*, 163–175. Retrieved from <https://doi.org/10.1007/s10851-011-0276-0>
- Plaza, J., Plaza, A., Perez, R., Martinez, P. (2009). On the use of small training sets for neural network-based characterization of mixed pixels in remotely sensed hyperspectral images. *Pattern Recognit*, 3032–3045. Retrieved from <https://doi.org/10.1016/j.patcog.2009.04.008>
- Puletti N., Perria R., Storchi P. (2014). Unsupervised classification of very high remotely sensed images for grapevine rows detection. *European Journal of Remote Sensin*, 45-54. Retrieved from <https://www.tandfonline.com/doi/abs/10.5721/EuJRS20144704>
- Qazi Sami ul Haq, et all. (2010). Hyperspectral Data Classification via Sparse Representation in Homotopy. *The 2nd International Conference on Information Science and Engineering* (pp. 3748-3752). Hangzhou, China: IEEE. Retrieved from <https://ieeexplore.ieee.org/document/5689027>
- R., T. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society*, 273-282. Retrieved from <https://pdfs.semanticscholar.org/6b5e/99c128b9cd7b7fbc817a2843a47ce8a1c35d.pdf>
- Razaviyayn, M., Tseng, H-W., Luo Z-Q. (2014). Dictionary learning for sparse representation: Complexity and algorithms. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5284-5288). Florence, Italy: IEEE. Retrieved from 10.1109/ICASSP.2014.6854604
- Rish, I., Grabarnik, G. (2014). *Sparse Modeling: Theory, Algorithms, and applications*. New York : Taylor & Francis Group.
- Rodarmel, C., Shan J. (2002). Principal Component Analysis for Hyperspectral Image Classification. *Surveying and Land Information Science*, 115-123. Retrieved from https://www.researchgate.net/publication/265198128_Principal_Component_Analysis_for_Hyperspectral_Image_Classification
- Rubinstein, R., Bruckstein, A.M., Elad, M. (2010). Dictionaries for Sparse Representation Modeling. *IEEE*, 1045 - 1057. Retrieved from <https://ieeexplore.ieee.org/document/5452966>
- Rubinstein, R., Bruckstein, AM., Elad, M. (2010). Dictionaries for Sparse Representation Modeling. *Proceedings of the IEEE*, 1045 - 1057. Retrieved from <https://ieeexplore.ieee.org/document/5452966>
- Rubinstein, R., Peleg, T., Elad, M. (2013). Analysis K-SVD: A Dictionary-Learning Algorithm for the Analysis Sparse Model. *IEEE Transactions on Signal Processing archive*, 661-677. Retrieved from <https://dl.acm.org/citation.cfm?id=2710711>
- Schmidt, M. (2005). Least Squares Optimization with L1-Norm Regularization. *citeseerx*, 1-12. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.186.3602>

- Schmidt, M., Fung, G., Rosales, R. (2009). Optimization Methods for L1-Regularization. *University of British Columbia, Technical Report TR*, 1-20. Retrieved from <https://www.cs.ubc.ca/cgi-bin/tr/2009/TR-2009-19.pdf>
- Schweizer, S.M, Moura, J.M.F. (2001). Efficient Detection in Hyperspectral Imagery. *IEEE Transactions*, 584–597. Retrieved from http://users.ece.cmu.edu/~moura/papers/schweitzer_effhyperspectral.pdf
- Shalaby A., Tateishi, R. (2007). Remote Sensing and GIS for Mapping and Monitoring Land Cover and Land-use Changes in the Northwestern Coastal Zone of Egypt. *Applied Geography*, 28-41.
- Shaw, G.A. and Burke, H.K. (2003). Spectral Imaging for Remote Sensing. *LINCOLN LABORATORY*, 3-28. Retrieved from <https://pdfs.semanticscholar.org/5ce6/339aca93ca69c00f4558c5a1bd08708d02e8.pdf>
- Shen, D., Shen, H., Marron, J.S. (2016). A general framework for consistency of principal component analysis. *The Journal of Machine Learning Research*, 5218-5251. Retrieved from <https://dl.acm.org/citation.cfm?id=3053432>
- Shin, Y., Lee, S., Ahn, M., Cho, H., Jun, S.S., Lee, H. (2015). Noise robustness analysis of sparse representation based classification method for non-stationary EEG signal classification. *Biomedical Signal Processing and Control*, 8-18.
- Singh, A. and A. Harison. (1985). Standardized principal component analysis. *Int. J. Rem. Sens*, 883–896. Retrieved from https://www.researchgate.net/publication/248975686_Standardized_principal_components
- Skretting, K., and Engan, K. (2010). Recursive Least Squares Dictionary Learning Algorithm. *IEEE Transactions on Signal Processing*, 2121 - 2130. Retrieved from <https://ieeexplore.ieee.org/document/5382523>
- Song B., Li J., Mura M., Li P., Plaza A., José M., Dias B., Benediktsson J., Chanussot J.,. (2014). Remotely Sensed Image Classification Using Sparse Representations of Morphological Attribute Profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 5122 - 5136. Retrieved from 10.1109/TGRS.2013.2286953
- Strang, G., & Aarikka, K. . (1986). Introduction to applied mathematics. *Wellesley-Cambridge Press*, 82-104.
- Tang, W., Shi, Z., Wu, Y., Zhang C. (2014). Sparse Unmixing of Hyperspectral Data Using Spectral A Priori Information. *IEEE Transactions on Geoscience and Remote Sensing*, 770 - 783. Retrieved from <https://ieeexplore.ieee.org/document/6840362>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 267–288. Retrieved from https://www.jstor.org/stable/2346178?seq=1#page_scan_tab_contents
- Tong, F., Tong, H., Jiang, J., Zhang, Y. (2017). Multiscale union regions adaptive sparse representation for hyperspectral image classification. *Remote Sens.*, 1-19. Retrieved from <https://doi.org/10.3390/rs9090872>

- Tropp, J. A. (2006). Algorithms for simultaneous sparse approximation part ii: Convex relaxation. *Signal Processing*, 589–602.
- Ülkü, i., Kizgut E. (2018). Large-scale hyperspectral image compression via sparse representations based on online learning. *International Journal of Applied Mathematics and Computer Science*, 197–207. Retrieved from <https://doi.org/10.2478/amcs-2018-0015>
- Valls, G-C., Tuia, D., Chova L-G., Jiménez, S., Malo J. (2012). *Remote Sensing Image Processing*. Spain: morgan and claypool publishers. Retrieved from <https://doi.org/10.2200/S00392ED1V01Y201107IVM012>
- Vasanth Raj, P.T., and Hans W.J. (2015). Sparse Representation Based Single Image Dictionary Construction For Image Super Resolution. *Australian Journal of Basic and Applied Sciences*, 386-390. Retrieved from <http://www.ajbasweb.com/old/ajbas/2015/Special%20ICSCS/386-390.pdf>
- Wang J., Jiao L., Liu H., Yang S., Liu F. (2015). Hyperspectral Image Classification by Spatial–spectral Derivative–Aided Kernel Joint Sparse Representation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2485–2500. Retrieved from <https://ieeexplore.ieee.org/document/7052294?denied=>
- Wang, H., H., Turgay. (2018). Sparse representation-based hyperspectral image classification. *Signal, Image and Video Processing*, 1009–1017.
- Wang, W., Qian, Y. (2016). Kernel based sparse NMF algorithm for hyperspectral unmixing. *International Geoscience and Remote Sensing Symposium (IGARSS)* (p. %0 Journal Article). Beijing, China: IEEE. Retrieved from <https://ieeexplore.ieee.org/document/7730818/authors#authors>
- Weber, C., Briottet, Xavier, B., Aguejdad R., Aval, Josselin, A. (2018). Hyperspectral Imagery for Environmental Urban Planning. *IGARSS* (pp. 1628-1631). Valencia, Spain: IEEE. Retrieved from 10.1109/IGARSS.2018.8519085
- X. Lu, H. Wu, Y. Yuan, P. Yan, and X. Li. (2013). Manifold regularized sparse nmf for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 2815–2826.
- Xie ID Fuding., F,Li., Lei, C., Ke, L. (2018). Representative Band Selection for Hyperspectral Image Classification. *nternational Journal of Geo-Information*, 537-86. Retrieved from <file:///C:/Users/admin/Downloads/ijgi-07-00338.pdf>
- Xu M., Watanachaturaporn P., Varshney P., Arora M. (2005). Decision Tree Regression for Soft Classification of Remote Sensing Data. *Remote Sensing of Environmen*, 322-336. Retrieved from <https://www.sciencedirect.com/science/article/abs/pii/S0034425705001604>
- Y. Xu, D. Zhang, J. Yang, and J. Yang. (2011). A two-phase test sample sparse representation method for use with face recognition. *IEEE Transactions on Circuits and Systems for Video Technology.*, 1255–1262.
- Y. Yuan, X., Li, Y., Pang, X., Lu, and D., Tao. (2009). Binary sparse nonnegative matrix factorization. *IEEE Transactions on Circuits and Systems for Video Technology*, 772–779.
- Yan, H. (2013). Sparsity Preserving Score for Joint Feature Selection. *Intelligence Science and Big Data Engineering*, 635-641.

- Yan, Hand., Yang, J. (2015). Sparse discriminative feature selection. *Pattern Recognition*, 1827-1835.
- Z. Zhang, Z. Li, B. Xie, L. Wang, and Y. Chen. (2014). Integrating globality and locality for robust representation based classification. *Mathematical Problems in Engineering*, 1-10. Retrieved from <http://dx.doi.org/10.1155/2014/415856>
- Zhang H., Li J., Huang Y., Zhang L. (2014). A nonlocal weighted joint sparse representation classification method for hyperspectral imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2056–2065. Retrieved from <https://ieeexplore.ieee.org/document/6522858>
- Zhang Z., Xu Y., Yang J., Li X., Zhang D. (2015). A survey of sparse representation: algorithms and applications. *IEEE Biometrics Compendium*, 490 - 530. Retrieved from 10.1109/ACCESS.2015.2430359
- Zhang Z., Xu Y., Yang J., Li X., Zhang D. (2016). A survey of sparse representation: algorithms and applications. *IEEE Biometrics Compendium*, 490 - 530. Retrieved from 10.1109/ACCESS.2015.2430359
- Zhang, H.; Zhai, H.; Zhang, L.; Li, P. (2016). Spectral-Spatial Sparse Subspace Clustering for Hyperspectral Remote Sensing Image. *IEEE Trans. Geosci. Remote Sens.*, 3672–3684. Retrieved from 10.1109/TGRS.2016.2524557
- Zhang, X., Sun, Y., Shang, K., Zhang, L., & Wang, S. (2016). Crop Classification Based on Feature Band Set Construction and Object-Oriented Approach Using Hyperspectral Images. *Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4117–4128. Retrieved from <https://ieeexplore.ieee.org/document/7500077>
- Zhu, Z., Qi, G., Chai, Y., Li, P. (2017). A Geometric Dictionary Learning Based Approach for Fluorescence Spectroscopy Image Fusion. *Applied Science, MDPI.*, 1-18. Retrieved from <https://www.semanticscholar.org/paper/A-Geometric-Dictionary-Learning-Based-Approach-for-Zhu-Qi/0baa73ddd9b17c1b84e595b97422081b2d383ce4>
- Zou, X., Feng, Y., Li, H., and Jiang, S. (2017). Sparse representation-based over-sampling technique for classification of imbalanced dataset. *Earth and Environmental Science* (pp. 1-10). Zhuhai, China: IOP Publishing Ltd. Retrieved from <https://iopscience.iop.org/article/10.1088/1755-1315/81/1/012201>
- Zuo, W., Meng D., Zhang, L., Feng, X., Zhang, D. (2014). A Generalized Iterated Shrinkage Algorithm for Non-convex Sparse Coding. *International Conference on Computer Vision* (pp. 1-8). Sydney, NSW, Australia: IEEE. Retrieved from <https://ieeexplore.ieee.org/document/6751136>