

# Abstract

Not more than 350 words

In many practical applications in machine learning, computer vision, data mining and information retrieval one is confronted with datasets whose intrinsic dimension is much smaller than the dimension of the ambient space. This has given rise to the challenge of effectively learning multiple low-dimensional subspaces from such data. Multi-subspace learning methods based on sparse representation, such as sparse representation based classification (SRC) and sparse subspace clustering (SSC) have become very popular due to their conceptual simplicity and empirical success. However, there have been very limited theoretical explanations for the correctness of such approaches in the literature. Moreover, the applicability of existing algorithms to real world datasets is limited due to their high computational and memory complexity, sensitivity to data corruptions as well as sensitivity to imbalanced data distributions.

This thesis attempts to advance our theoretical understanding of sparse representation based multi-subspace learning methods, as well as develop new algorithms for handling large-scale, corrupted and imbalanced data. The first

## ABSTRACT

contribution of this thesis is a theoretical analysis of the correctness of such methods. In our geometric and randomized analysis, we answer important theoretical questions such as the effect of subspace arrangement, data distribution, subspace dimension, data sampling density, and so on.

The second contribution of this thesis is the development of practical subspace clustering algorithms that are able to deal with large-scale, corrupted and imbalanced datasets. To deal with large-scale data, we study different approaches based on active support and divide-and-conquer ideas, and show that these approaches offer a good tradeoff between high accuracy and low running time. To deal with corrupted data, we construct a Markov chain whose stationary distribution can be used to separate between inliers and outliers. Finally, we propose an efficient exemplar selection and subspace clustering method that outperforms traditional methods on imbalanced data.

[Include readers on the same pages as the abstract](#)

**Primary Reader and Advisor:** René Vidal

**Secondary Reader:** Daniel P. Robinson

OPTIONAL

# Acknowledgments

If included, place after abstract and before table of contents

I am very grateful to my advisor, Professor René Vidal, for introducing me to the fantastic world of machine learning, and for his guidance along the road of my PhD study. René's profound insights and broad vision in research have been a role model to me and have been the incentive to me to pursue my research. I am also thankful to him for always being patient and for giving me the freedom to explore different research ideas and conduct research through trial and error, which not only helped me to shape my independent research interests but also made the research process enjoyable.

I am also grateful to Professor Daniel P. Robinson for always being supportive in my research. I have learned a lot from Daniel on the importance of precise and rigorous writing in scientific papers and reports. I would also like to express my gratitude to Prof. Trac Tran, Prof. Vishal Patel, Prof. Enrique Mallada and Prof. Yi Ma for serving in my thesis proposal and dissertation committees, and to Prof. Gregory Hager, Prof. Sridevi Sarma for serving in my Graduate Board Oral committee.

OPTIONAL

# **Dedication**

If included, place after the abstract and before the table of contents.

This thesis is dedicated to my parents, Fengyun Li and Zhendong You, for their eternal love, trust and support.

Acknowledgements, dedication, prefacing quotes, etc may all be placed in any order between the abstract and table of contents

# Contents

<b>Abstract</b>	Front matter is included in Table of Contents Each element begins on a new page	<b>ii</b>
<b>Acknowledgments</b>		<b>iv</b>
<b>List of Tables</b>		<b>xv</b>
<b>List of Figures</b>		<b>xvi</b>
	Chapters and subtitles are listed with accurate page numbers	
<b>1 Introduction</b>		<b>1</b>
1.1 Multi-subspace data . . . . .		2
1.2 Multi-subspace learning . . . . .		3
1.2.1 Subspace classification . . . . .		4
1.2.2 Subspace clustering . . . . .		5
1.3 Sparse methods for multi-subspace learning . . . . .		7
1.4 Thesis contributions . . . . .		9
1.4.1 Geometric and probabilistic analysis of multi-subspace learn- ing methods . . . . .		10

## CONTENTS

<b>7 Conclusions</b>	<b>299</b>
<b>Bibliography</b>	<b>301</b>
<b>Vita</b> <a href="#">Not required, but if included, use continuous page numbering from main text</a>	<b>332</b>

Only required if tables are included in the document

## List of Tables

If included, place after the table of contents and before the list of figures

1.1	Evaluation of subspace clustering methods on real data: a summary . . . . .	21
4.1	Dataset information for testing subspace clustering on real image databases. . . . .	215
4.2	Performance of subspace clustering methods on large-scale data - clustering accuracy . . . . .	216
4.3	Performance of subspace clustering methods on large-scale data - running time . . . . .	217
4.4	Performance of SSC-DC on the MNIST dataset . . . . .	221
5.1	Outlier detection results on the Extended Yale B database . . . . .	243
5.2	Outlier detection results on the Caltech-256 database . . . . .	246
5.3	Outlier detection results on the Coil-100 database. . . . .	247
5.4	Running time of outlier detection on Extended Yale B data . . . . .	248
6.1	Subspace clustering on the GTSRB street sign database. . . . .	288
6.2	Subspace clustering on a small subset of the GTSRB street sign database. . . . .	291
6.3	Classification from subsets on the Extended Yale B face database	293

Includes table number and accurate page number

Only required if figures are included in the document

# List of Figures

If included, place after list of tables

1.1	An illustration of multi-subspace structure in a face dataset . . .	3
1.2	An illustration of the subspace classification problem . . . . .	5
1.3	An illustration of the subspace clustering problem . . . . .	6
1.4	An illustration of subspace-preserving representation for multi-subspace data . . . . .	9
1.5	The effect of subspace separation on multi-subspace learning . .	11
1.6	The effect of point distribution in each of the subspaces on multi-subspace learning . . . . .	12
1.7	Challenge and contribution for subspace clustering on large-scale data . . . . .	15
1.8	Challenge and contribution for subspace clustering on corrupted data . . . . .	17
1.9	Challenge and contribution for subspace clustering on imbalanced data . . . . .	19
2.1	Trade-off between subspace-preserving and connectedness properties in SSC-BP and LSR . . . . .	43
2.2	Representation matrix of SSC-BP in the presence of outliers . . .	51
2.3	Representation matrix of SSC-BP for imbalanced dataset . . . . .	53
3.1	Illustration of the geometric characterization of the dictionary . .	61
3.2	Illustration of the geometry associated with subspace-preserving representations . . . . .	82
3.3	Summary of the results of universal recovery conditions . . . . .	83
3.4	A comparison of the condition in (3.66) and the condition in (3.70).	105
3.5	Illustration for proving bounds for area of spherical cap. . . . .	110
3.6	Illustration of the structure of the solution for elastic-net . . . . .	157
3.7	Illustration of oracle point and oracle region for elastic-net problem	160
3.8	The structure of the solution for EnSC . . . . .	166

Includes figure number and accurate page number

Marks beginning of Main Text

# Chapter 1

## Introduction

The significant increase in the ability to collect and store diverse information in the past decades has led to an exceptional growth in the availability of data. In the field of computer vision, for instance, portable and affordable digital cameras and smartphones interconnected with high-speed mobile networks have produced image and video datasets of unprecedented scale, which are being collected by giant Internet companies such as Google and Amazon through services they provide to billions of customers. The proliferation in dataset size and complexity is accompanied by the challenge of successfully analyzing the data to discover patterns of interest. Aside from being large-scale, modern datasets very often possess significant amounts of corruptions in various forms such as noise, corrupted entries, outliers and missing entries. All these features pose stark challenges to the development of techniques for modern data