RICE UNIVERSITY

Endogenous Sparse Recovery

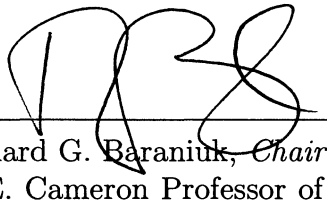
by

Eva L. Dyer

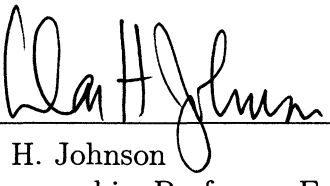
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Masters of Science

APPROVED, THESIS COMMITTEE:



Dr. Richard G. Baraniuk, *Chair*
Victor E. Cameron Professor of Electrical and Computer Engineering



Dr. Don H. Johnson
J.S. Abercrombie Professor Emeritus of Electrical and Computer Engineering

Dr. Wotao Yin
Assistant Professor of Computational and Applied Mathematics

HOUSTON, TEXAS
DECEMBER 2011

Abstract

Endogenous Sparse Recovery

by

Eva L. Dyer

Sparsity has proven to be an essential ingredient in the development of efficient solutions to a number of problems in signal processing and machine learning. In all of these settings, sparse recovery methods are employed to recover signals that admit sparse representations in a pre-specified basis. Recently, sparse recovery methods have been employed in an entirely new way; instead of finding a sparse representation of a signal in a fixed basis, a sparse representation is formed "from within" the data. In this thesis, we study the utility of this *endogenous sparse recovery* procedure for learning unions of subspaces from collections of high-dimensional data. We provide new insights into the behavior of endogenous sparse recovery, develop sufficient conditions that describe when greedy methods will reveal local estimates of the subspaces in the ensemble, and introduce new methods to learn unions of overlapping subspaces from local subspace estimates.

This thesis is dedicated to my mother.
I couldn't have done any of this without your support and love.

Contents

Abstract iii

1 Introduction 2

1.1 Unions of Subspaces 2

1.2 Endogenous Sparse Recovery 4

1.3 Motivating Example 6

1.4 Thesis Organization 8

1.5 Notation and Preliminaries 10

2 Background 12

2.1 Sparse Approximation 12

2.1.1 Sparse Representation in Overcomplete Dictionaries 12

2.1.2 Sparse Recovery Methods 14

2.1.3 Exact Recovery Conditions 15

2.2 Applications of Endogenous Sparse Recovery 17

3 Greedy Feature Selection from Unions of Subspaces 20

3.1 Preliminaries 20

3.1.1 Signal Model 20
3.1.2 Projective Space 21
3.2 Greedy Feature Selection 23
3.3 Exact Feature Selection 25

4 EFS from Unions of Disjoint Subspaces **27**
4.1 Principal Angles 27
4.2 Greedy Selection Lemma 29
4.3 Exact Feature Selection Theorem 31

5 EFS for Uniformly Bounded Unions **33**
5.1 Uniformly Bounded Unions 33
5.2 EFS Lemma for Bounded Unions 34
5.3 Theorem for EFS from Bounded Unions 37

6 Empirical Study of EFS from Unions of Subspaces **39**
6.1 Generating Structured Unions of Subspaces 39
6.2 Influence of Cross-spectra on EFS 41
6.3 Phase Transitions for EFS 43
6.3.1 Oversampling of Subspaces 44
6.3.2 Energy in Subspace Intersections 44
6.3.3 Comparison of ℓ_0 and NN-graphs 46

7 Methods for Learning Unions of Subspaces **48**
7.1 Methods for Subspace Learning 48
7.2 Clustering or Consensus? 49
7.3 Consensus on the ℓ_0-graph 51
7.4 Experimental Results 51

8 Learning Unions of Subspaces from Image and Text Data **55**

8.1 Face Illumination Subspaces 55

8.2 Motion Segmentation Data 56

8.3 Multispectral Image Segmentation 57

8.4 Document Clustering 61

8.4.1 Clustering Documents with ℓ_0-graphs 61

8.4.2 Visualizing Information Flow Across Documents 62

9 Discussion **65**

9.1 Summary of Results 65

9.2 Implications of this Work 66

9.2.1 Cross-spectral Minimization 67

9.2.2 'Data Driven' Sparse Approximation 67

9.2.3 Learning Block Sparse Signal Models 68

9.3 Going Beyond Coherence 68

9.4 Open Questions 69

References **70**

List of Figures

1.1 *Comparison of the nearest neighbor graph and ℓ_0-graph for unions of illumination subspaces* 9

3.1 *Covering radius in the projective space.* 23

3.2 *Covering radius in the ambient space.* 24

3.3 *Demonstration of exact feature selection (EFS) from unions of subspaces* 26

6.1 *Generating unions of subspaces from shift-invariant dictionaries* . . . 40

6.2 *Probability of EFS for unions with structured cross-spectra* 42

6.3 *Probability of EFS for different sampling conditions* 45

6.4 *Phase transitions for sparse recovery and nearest neighbor graphs.* . . 47

7.1 *Probability of subspace recovery* 53

7.2 *Performance of subspace consensus.* 54

8.1 *Cross-spectra for illumination subspaces* 57

8.2 *Comparison of ℓ_0-graphs with NN-graphs.* 58

8.3 *Classification performance for segmenting two motions.* 59

8.4 *Classification performance for segmenting three motions.* 59

8.5 *Multispectral image segmentation.* 60

8.6 *Document affinity matrices.* 61

8.7 *Document support matrices.* 62

8.8 *Visualizing information flow across documents.* 63

CHAPTER 1

Introduction

1.1 Unions of Subspaces

With the emergence of novel sensing systems capable of acquiring data at scales ranging from the nano to the tera, modern sensor and imaging data is becoming increasingly high-dimensional and heterogenous. To cope with this explosion of complex high-dimensional data, we must exploit the fact that 'natural signals'[1] have intrinsic structure of much lower dimension than that of the ambient space.

Linear subspace models are one of the most widely used signal models for characterizing the intrinsic low-dimensional structure contained within collections of high-dimensional data, with applications throughout signal processing, machine learning, and the computational sciences. This is in part due to the simplicity of linear models but is also due to the fact that principal components analysis (PCA) provides an elegant closed-form solution to the problem of finding an optimal low-rank approximation to a collection of data (an ensemble of points in $\mathbb{R}^n$). More formally, if we stack a collection of d points in $\mathbb{R}^n$ into the columns of a matrix $Y \in \mathbb{R}^{n \times d}$, PCA

[1]Natural signals arise when studying natural phenomenon. Examples of natural signals include: images captured from a structured light field, the trajectory of a protein when moving from its native to unfolded state, or the acoustic waveform arising from the pluck of a guitar string.

seeks the best rank-k estimate of Y by solving

$$(PCA) \qquad \min_{X} \quad \|Y - X\|_F \quad \text{subject to} \quad \text{rank}(X) \leq k. \tag{1.1}$$

Despite the power of linear subspace models, mounting evidence suggests that a wide range of data may not be succinctly represented in terms of a single linear subspace but instead admit an *union of subspaces*. For instance, ensembles ranging from collections of images taken of objects under different illumination conditions [1], motion trajectories of point-correspondences [2, 3], to structured sparse and block-sparse signals [4, 5, 6], can all be well-approximated by a union of low-dimensional subspaces or a union of affine hyperplanes (flats). Unions of subspace models have also found utility in the classification of signals collected from complex systems at different points in time, e.g., local field potentials collected from the motor cortex over different days [7].

Unions of subspaces provide a natural extension of linear subspace models, but providing a *provable* extension of PCA that is capable of determining an optimal union of subspaces that well-approximate a collection of data, is extremely challenging. This is due to the fact that segmentation—the identification of points that live in the same subspace—and subspace estimation must be performed simultaneously. However, if we can accurately sift through the points in the ensemble and determine which subsets of points lie near the same subspace, then subspace estimation becomes trivial.

A common approach for identifying sets of points that live in the same subspace is to determine the 'multi-way affinity' between points in the set from locally linear approximations to the data [8] . To be precise, these methods compute the affinity between two test points by fitting a linear approximation to the points within an euclidean neighborhood of each test point and computing the the similarity between these subspace estimates. After determining the affinity between points in the set,

spectral clustering is performed on the resulting affinity matrix. Methods that use nearest neighbor sets to form locally linear approximations to data include: local subspace affinity (LSA) [9], spectral clustering based on locally linear approximations [8], spectral curvature clustering [10], and local best-fit flats [11, 12].

When the subspaces present in the ensemble are independent and/or are linearly separable, linear approximations obtained from neighboring points typically provide reliable and stable estimates of the affinity between points in the ensemble. However, neighborhood-based approaches quickly begin to fail as the overlap between the two structures increases and as the subspace dimension increases. This is due to the fact that as the overlap between two subspaces increases, the set of points that live in neighborhoods of one another are less likely to be contained within the same subspace. This suggests that if we can find another *feature selection* strategy that improves our probability of selecting a feature set that contains points from the same subspace, then we can use this alternate set of points to form a *local subspace estimate* [1] instead of forming linear approximations from sets of near neighbors.

1.2 Endogenous Sparse Recovery

Recently, Elhamifar et al. have set forth an entirely new proposal for feature selection which remedies a number of the issues that arise when using neighborhood-based subspace estimates [13]. The idea behind this approach is to select a subset of points from the ensemble that provide a 'sparse representation' of another point in the same ensemble. By enforcing sparsity in this representation, one can show that a representation formed from points in the same subspace is more efficient than a representation formed from points outside of the subspace that the point is contained in [14]. For

[1] We refer to subspace estimates as being local if they are formed from a subset of points in the ensemble. In contrast, we refer to standard low-rank approximation over the entire set of points as a global subspace estimate.

this reason, the resulting feature sets selected via sparse recovery methods are likely to contain points that all belong to the same subspace.

We will refer to this application of sparse recovery methods as *endogenous sparse recovery* due to the fact that representations are not formed from an external source—as in standard applications of sparse recovery—but are formed "from within" the data. Formally, for a collection of d signals, $\mathcal{Y} = \{y_1, \ldots, y_d\}$ each of dimension n, we seek a sparse representation of the i^{th} point with respect to the remaining $d-1$ points in the ensemble. The sparsest endogenous representation of the i^{th} point is defined as the representation with smallest "ℓ_0-norm"[1]

$$c_i^* = \arg\min_c \|c\|_0 \quad \text{subject to} \quad y_i = \sum_{j \neq i} c(j) y_j, \tag{1.2}$$

where the $\|c\|_0$ counts the number of non-zeroes in its argument. Let $\Lambda^{(i)} = \text{supp}(c_i)$ denote the subset of points from $\mathcal{Y}$ selected to represent the point y_i and $c_i(j)$ denote the contribution of the j^{th} point to the sparse representation of y_i. We will also refer to $\Lambda^{(i)}$ as the feature set selected for the i^{th} point. In general, finding the optimal subset of columns from the ensemble that possess the smallest cardinality has combinatorial complexity; rather, sparse recovery methods such as basis pursuit (BP) [15] or a greedy pursuit (OMP) [16] may be employed to find approximate solutions to this problem.

When the data live on a union of disjoint subspaces, i.e., subspaces intersect only at the origin, and are sufficiently separated, Elhamifar et al. demonstrated that the feature sets selected via BP will only contain points from the same subspace [14]. These results provide new insight into the role that 'sparsity' may play in feature selection from unions of subspaces; however, the practical performance of this tech-

[1]The "ℓ_0-norm" is placed in quotes because it is not actually a norm. The ℓ_0-penalty $\|x\|_0$ simply counts the number of non-zeros in its argument.

nique has quickly outpaced the theoretical results that exist in the literature. In particular, there has been no study of the utility of greedy feature selection strategies for endogenous sparse recovery as well as the application of endogenous sparse recovery to non-disjoint or *overlapping subspaces.* Examples of unions of non-disjoint or *overlapping subspaces* include natural image ensembles [17], illumination subspaces [18], and overlapping block-sparse signals [19, 20].

The aim of this thesis is to provide new insights into the behavior of greedy feature selection strategies for learning local subspace estimates from collections of high-dimensional data. The contributions of this thesis can be summarized in terms of our efforts of three main fronts:

1. Theoretical analysis of greedy feature selection from unions of subspaces.
2. Empirical study of EFS from unions of overlapping subspaces.
3. Study and comparison of methods for learning unions of subspaces from local subspace estimates.

1.3 Motivating Example

Before proceeding, we begin by revealing an interesting property of greedy feature selection from unions of subspaces—the feature sets selected by matching pursuits exhibit *diversity.* When we say that the feature sets are diverse, we mean that each point in the set is sufficiently different from the rest of the points in the set. This is due to the fact that orthogonal greedy methods such as OMP find points in the dataset that are highly correlated with the signal of interest; however, each time we select a point from the dataset, the signal is projected into the space orthogonal to the subspace spanned by the points selected at previous iterations. This guarantees that all of the points in our feature set will point in different directions and are not

redundant. As a consequence, we avoid accumulating redundant points in our feature set that will skew our local subspace estimates. This is particularly useful in the case where our nearest neighbor graph exhibits hubs (nodes with very high degree); in this setting, we find that greedy feature selection can be used to recover affinities that are hub-free.

We now provide an example of this hub-breaking phenomenon. In Figure 1.1, we show an example of a union of subspaces formed from two different faces under various illumination conditions. To visualize the affinity between points across different subspaces, the data is sorted such that all of the images from a single face are in a contiguous block. By sorting the data in this way, we expect to see clustering in the block-diagonal component and minimal edges contained in the off-diagonal component. On the left, we show the adjacency matrix A for the near neighbor (NN) graph, which is laden with hubs in the off-diagonal which link points belonging to different subspaces. At the bottom of the NN affinity matrix, we show an example of two points from different subspaces that are linked via one of these hubs on the NN graph.

On the right, we show the ℓ_0-*graph* formed from the same ensemble. The ℓ_0-graph of the ensemble $G = (V, E)$ contains $|V| = d$ vertices, where each vertex corresponds to a particular point in the dataset. If we assume that each point in the ensemble can be expressed with no more than k points in the set,[1] then the number of edges $|E| \leq kd$. In general, the edge weight between vertex i and j can take on any number of different values, as long as the edge weight is non-zero when point i and j use one another in their sparse representations. To ensure that the ℓ_0-graph of $\mathcal{S}(\mathcal{Y}) = \{\Lambda^{(i)}$

[1]We will refer to this property as self-expression. Our implicit assumption will be that for points living in k-dimensional subspaces, we can represent each point in the set in terms of at most k other points in the ensemble.

is symmetric, i.e., $e_{ij} = e_{ji}$, we will define the edge weights e_{ij} as follows:

$$e_{ij} = \begin{cases} 0 & \text{if } j \notin \Lambda^{(i)}, i \notin \Lambda^{(j)}, \\ |c_i(j)| + |c_j(i)| & \text{else.} \end{cases}$$

This example provides striking visual evidence of the power of endogenous sparse recovery from unions of subspaces, where we observe perfect clustering in the correct components. In contrast, the NN-graph contains a great deal of energy in its off block-diagonal, suggesting that points from different subspaces (images of different people) will be clustered together in terms of their NN relationship. This results suggests that endogenous sparse recovery is capable of "breaking hubs in high-dimensions". This is in stark contrast to NN-graphs which are known to be susceptible to hubs.[1]

1.4 Thesis Organization

We provide a roadmap of the main contributions of this thesis below.

Chapter 2. We describe relevant work in sparse recovery and discuss methods for forming sparse approximations from overcomplete dictionaries. Following this, we provide a summary of applications of endogenous sparse recovery to subspace learning problems.

Chapter 3. We provide a method for greedy feature selection and introduce the notion of exact feature selection (EFS).

Chapter 4–5. We develop sufficient conditions for EFS with greedy methods that reveals an intimate relationship between the covering of subspaces in the ensemble and the geometry of the union of subspaces. In Chapter 5, we extend these

[1] See [21] for a description of the 'hubness' phenomenon that plagues NN-graphs and classifiers in high-dimensions.

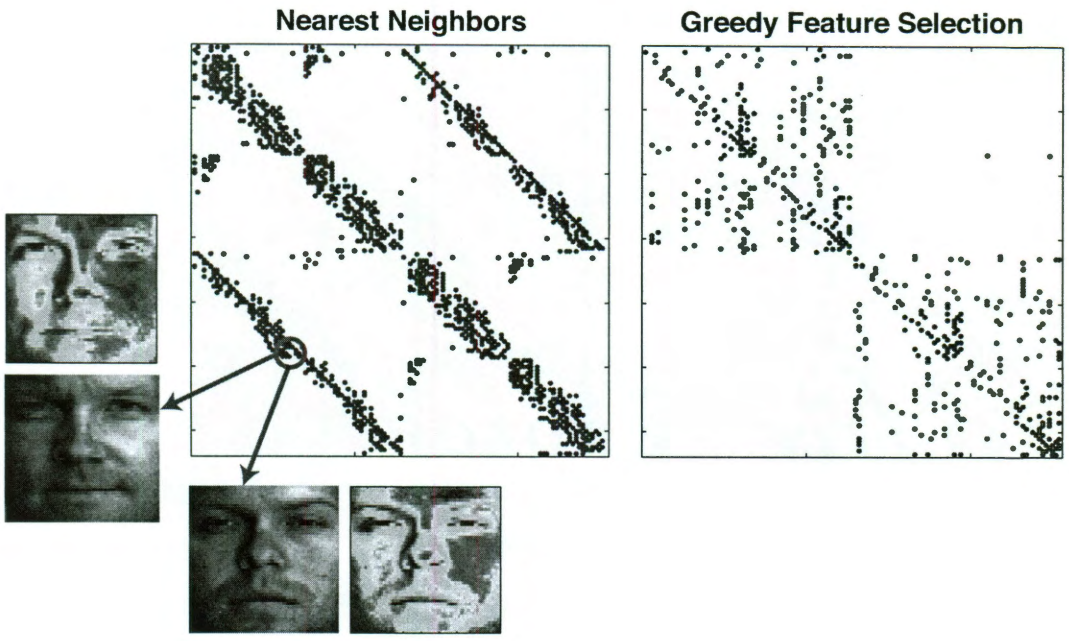


Figure 1.1: *Comparison of nearest neighbor graph and ℓ_0-graph for unions of illumination subspaces.* The highlighted point corresponds to a hub where points from different subspaces are identified as near neighbors. On the right, we show that greedy feature selection avoids selecting points from these hubs when we form endogenous representations of the data.

results to the setting where our data admits an 'uniformly bounded union'; this enables us to reveal further dependencies between EFS and the the distribution of principal angles between subspaces.

Chapter 6. We study and characterize the empirical phase transitions for EFS from structured unions of overlapping subspaces.

Chapter 7. We introduce a new method for subspace recovery and compare this method with other methods for learning unions of subspaces from local feature sets.

Chapter 8. We study the application of greedy selection and our new methods for subspace consensus to segmentation and clustering problems arising in a num-

ber of applications which include: segmentation of multispectral images, face subspaces, and motion trajectories, as well as document cluster analysis from educational material.

Chapter 9. We discuss the implications of our analysis and subsequent studies on sparse approximation, dictionary learning, feature selection methods. We conclude with a number of interesting open problems.

1.5 Notation and Preliminaries

In this paper, we will work solely in real finite-dimensional vector spaces or in $\mathbb{R}^n$. We write vectors x in lowercase script, matrices A in uppercase script, and scalar entries of vectors as $x(j)$. The standard p-norm is defined as

$$\|x\|_p = \left(\sum_{j=1}^{n} |x(j)|^p \right)^{1/p},$$

where $p \geq 1$. The ℓ_0 quasi-norm of a vector x is defined as the number of non-zero elements in x. The support of a vector x, often written as $\text{supp}(x)$, is defined as the set containing the indices of its non-zero coefficients; hence, $\|x\|_0 = |\text{supp}(x)|$. The key matrix norms that we will employ in our subsequent analysis include: $\|A\|_{1,1}$, the maximum ℓ_1-norm across all the columns in A and the spectral norm $\|A\|_{2,2}$, which is also equivalent to the maximum singular value of A.

We denote the Moore-Penrose pseudoinverse of a matrix A as $A^\dagger$. If $A = U\Sigma V^T$ then $A^\dagger = V\Sigma^+ U^T$, where we obtain Σ^+ by taking the reciprocal of the entries in Σ, leaving the zeros in their place, and taking the transpose of this matrix. An orthonormal basis (ONB) is known to satisfy the following two properties, $\Phi_i^T \Phi_i = I_k$, and $\text{range}(\Phi_i) = \mathcal{W}_i$, where I_k is the $k \times k$ identity matrix. The ortho-projector onto the subspace spanned by the sub-matrix X_Λ is defined as $P_\Lambda = X_\Lambda X_\Lambda^\dagger$.

We denote the ℓ_2 sphere of radius r as

$$\mathbb{S}^{n-1}(r) = \{x \in \mathbb{R}^n \ : \ \|x\|_2 = r\}. \tag{1.3}$$

We will write the unit sphere as $\mathbb{S}^{n-1}$, without specifying the radius.

CHAPTER 2

Background

In this chapter, we review relevant work in sparse signal recovery and describe applications of endogenous sparse recovery to subspace learning and clustering problems.

2.1 Sparse Approximation

2.1.1 Sparse Representation in Overcomplete Dictionaries

Sparsity has proven to be an essential ingredient in the development of efficient and, in some cases, unique solutions to a number of fundamental problems in signal processing and machine learning, from compression and denoising of signals [15, 22] to compressive sensing [23, 24], morphological components analysis [25, 26] and sparse-representation based classification [27]. In all of these settings, sparse recovery methods, e.g., ℓ_1-minimization [15, 28] or greedy pursuits [16], are employed to recover signals that admit sparse representations in a fixed and pre-specified basis or overcomplete dictionary.

To make this precise, we refer to a finite collection of unit-norm atoms $\mathcal{D} = \{\varphi_i\}_{i=1}^d$ as a *dictionary*. If the dictionary is complete, i.e., spans $\mathbb{R}^n$, then an exact recon-

struction of any input signal $x \in \mathbb{R}^n$ can be formed by finding a linear combination of the atoms in the dictionary as follows

$$x = \sum_{i=1}^{d} a(i)\varphi_i = \Phi a, \tag{2.1}$$

where $\Phi \in \mathbb{R}^{n \times d}$ contains the atoms in $\mathcal{D}$ in its columns and $a(i)$ indexes the i^{th} entry of the coefficient vector $a \in \mathbb{R}^d$. When Φ forms an ONB, the coefficients in (2.1) are uniquely determined by the projection of each basis vector onto the signal of interest, i.e., $a(i) = \langle \varphi_i, x \rangle$. In this case, the representation of x with respect to Φ is unique. However, when Φ is overcomplete, i.e., $d > n$, an infinite number of representations can be formed from the atoms in Φ. Hence, the simplest or most parsimonious explanation can be sought by finding a *sparse representation* of x with respect to the atoms in Φ. To find the sparsest representation of x, our aim is to find a solution to the following problem

$$\text{(EXACT)} \qquad \min_{\mathrm{a}} \quad \|\mathrm{a}\|_0 \quad \text{subject to} \quad \mathrm{x} = \Phi\mathrm{a}. \tag{2.2}$$

Instead of looking for the sparsest representation directly, we may instead fix the sparsity level k and then search for the best k-term approximation of x

$$\text{(SPARSE)} \qquad \min_{a} \quad \|x - \Phi a\|_2 \quad \text{subject to} \quad \|a\|_0 \leq k. \tag{2.3}$$

Although the objectives of both (EXACT) and (SPARSE) are similar, we will find that the structure of each problem lends itself well to different classes of methods designed for sparse signal recovery.

2.1.2 Sparse Recovery Methods

Methods for sparse recovery fall broadly into one of two classes, convex optimization-based approaches and greedy pursuit methods. The first class of methods transform the non-convex objective function in (EXACT) into a convex objective by replacing the ℓ_0 penalty with the ℓ_1 norm. This relaxation results in a formulation which is known to as Basis Pursuit (BP)

$$\text{(BP)} \qquad \min_{\mathrm{a}} \quad \|\mathrm{a}\|_1 \quad \text{subject to} \quad \mathrm{x} = \Phi\mathrm{a}. \tag{2.4}$$

We may also relax this equality constraint by trading off the sparsity of the solution with the ℓ_2 approximation error; this results in convex formulations known as basis pursuit denoising (BPDN) [15] and the LASSO [28].

The second class of sparse recovery methods employ greedy pursuit strategies to find an approximate solution to (SPARSE). Examples of greedy pursuit strategies include matching pursuit (MP) [29], orthogonal matching pursuit (OMP) [16], or subspace pursuits such as CoSaMP [30]. Greedy methods work by selecting atoms iteratively, subtracting the contribution of each selected atom from the current signal residual. This selection process is then repeated until a stopping criterion is satisfied: either a target sparsity $\|\widehat{a}\|_0 = k$ is reached, or the residual magnitude becomes smaller than a pre-specified value.

Greedy pursuits will serve as the algorithmic framework for our subsequent study. For this reason, we detail OMP in Algorithm 1 to familiarize the reader with the algorithm.

Algorithm 1 : Orthogonal Matching Pursuit

Input: Input signal $y \in \mathbb{R}^n$, a dictionary $\Phi \in \mathbb{R}^{n \times d-1}$, and a stopping criterion (either number of atoms k or the norm of signal residual ϵ).
Output: Index set Λ containing the indices of all *atoms* chosen in the pursuit.
Initialize: Set the residual to the input signal $r_0 = y$.
1. Compute the analysis coefficients for the current residual as $\Phi^T r_n$.
2. Find the largest analysis coefficient in absolute magnitude. Call the *atom* corresponding to this maximum analysis coefficient φ_j and place its index in the support set $\Lambda = \Lambda \cup j$.
3. Update the residual, $r_{n+1} = (I - \Phi_\Lambda \Phi_\Lambda^\dagger)y$.
4. Repeat steps (1-3) until a stopping criterion is reached, i.e., either k *atoms* are selected or $\|r_n\| \leq \epsilon$.

2.1.3 Exact Recovery Conditions

In this section, we will describe geometric constraints on the dictionary required to guarantee *exact support recovery* for a signal that lies in the span of a particular subset of atoms from $\mathcal{D}$. If we assume that x has been synthesized from a linear combination of atoms in the sub-dictionary $\Phi_\Lambda \in \mathbb{R}^{n \times k}$, then we will be interested in when we can uniquely recover an approximation of x that consists solely of the elements in Λ. In this case, we say that exact support recovery occurs. After recovering the support of our signal, the best ℓ_2-approximation of the signal is then found by projecting the onto the subspace spanned by the recovered set of atoms, where $\widehat{x} = \Phi_\Lambda \Phi_\Lambda^\dagger x$.

To guarantee that exact support recovery occurs for all signals supported over a particular sub-dictionary, Tropp introduced a general *exact recovery condition* (ERC) for both BP [31] and OMP [32]. The ERC is defined as follows.

Definition 1 (Exact Recovery Condition) *For any signal supported over the sub-dictionary Φ_Λ, exact support recovery is guaranteed for both OMP and BP if the following condition holds*

$$ERC(\Lambda) \equiv \max_{i \notin \Lambda} \|\Phi_\Lambda{}^\dagger \varphi_i\|_1 < 1.$$

A geometric interpretation of this condition is that the ERC provides a measure of how far a projected atom $P_\Lambda \varphi$ outside of the set Λ lies from the antipodal convex hull of the atoms in Λ. In words, the ERC(Λ) provides a measure of how unique a representation drawn from a superposition of atoms in Λ is with respect to the rest of the elements in the dictionary. If the atoms outside of Λ are similar to the atoms in Λ, then exact support recovery is not guaranteed.

Although the ERC provides some intuition about when a signal can be uniquely recovered from a certain sub-dictionary, to guarantee that all k-sparse signals can be uniquely recovered, we must ensure that all sub-dictionaries of size k satisfy the condition that $ERC(\Lambda) < 1$. Thus, in practice, the exact recovery condition is impossible to check because it requires evaluating $ERC(\Lambda)$ for all $\binom{d}{k}$ sub-dictionaries of size k and finding the maximum over this set. Instead, these conditions are often translated into constraints on the geometric structure of the dictionary.

Two such quantities that we will be interested in are the maximum coherence and the cumulative coherence of the dictionary. We supply a formal definition of both quantities below.

Definition 2 (Maximum coherence) *The maximum coherence of a dictionary of unit-norm atoms $\mathcal{D} = \{\varphi_i\}_{i=1}^d$ is defined as*

$$\mu \equiv \max_{i \neq j} |\langle \varphi_i, \varphi_j \rangle|.$$

Definition 3 (Cumulative coherence) *The cumulative coherence of a dictionary of unit-norm atoms $\mathcal{D} = \{\varphi_i\}_{i=1}^d$, is defined as*

$$\mu_1(k) \equiv \max_{|\Lambda| \leq k} \max_{i \notin \Lambda} \sum_{j \in \Lambda} |\langle \varphi_i, \varphi_j \rangle|.$$

Whereas the maximum coherence describes the maximum amount of coherence that

exists between two atoms in the dictionary, the cumulative coherence measures the accumulation of coherence between a fixed atom and k other atoms in the dictionary. Moreover, the cumulative coherence gives an upper bound on the absolute off-diagonal row (or column) sum of the Gram matrix obtained for any sub-dictionary, where the Gram matrix list the inner products between the atoms in the sub-dictionary or $G = \Phi_\Lambda^T \Phi_\Lambda$ for any set $|\Lambda| \leq k$. For a detailed review of the geometry of sparse approximation and ERC-based conditions for noisy signals see [33].

2.2 Applications of Endogenous Sparse Recovery

We now discuss applications of endogenous sparse recovery in both supervised and unsupervised subspace learning problems.

P1. *Subspace clustering:* The goal of subspace clustering is to partition points in an ensemble in accordance with the subspace membership of each point. In [13], endogenous sparse recovery (BP) is used to form an adjacency matrix A for the ensemble, where the $(i,j)^{th}$ entry of A is given by

$$A_{ij} = |c_i(j)| + |c_j(i)|.$$

Following this, spectral clustering is performed on the graph Laplacian of the adjacency matrix A. Applications of subspace clustering include: segmenting motion trajectories, data-driven object recognition, and segmentation of diffusion tensor imaging data.

P2. *Subspace consensus:* The goal of subspace consensus is to find a linear approximation to a subset of points from the dataset (local subspace estimate) and look for agreement or 'consensus' amongst local subspace estimates obtained

from different subsets of the data. A standard approach for selecting subsets of points from the data is to select points that live in a local euclidean neighborhood around a point. In [34], endogenous sparse recovery is instead used to select subsets of points from which we may form local subspace estimates. At a high-level, this application of endogenous recovery can be summarized as follows:

1. Solve the endogenous sparse recovery problem in (1.2) for each point in the ensemble to obtain a collection of support sets $\mathcal{S}(\mathcal{Y}) = \{\Lambda^{(i)}\}_{i=1}^{d}$.
2. Compute local subspace estimates for each point by finding a low-rank approximation to the points indexed by each of the support sets in $\mathcal{S}(\mathcal{Y})$. This can be done either with PCA or a robust variant [35].
3. Determine the local estimates to be included in the model by letting points vote upon which of the estimates they agree upon, e.g., find the mode in the subspace estimates or count the number of points that lie within a fixed region around each of the local subspace estimates.

Applications of this include decoding trajectories from local field potentials in the motor cortex [7] and for dictionary learning in audio source separation problems [34].

P3. *Supervised subspace classification:* The goal of supervised subspace classification methods that employ endogenous sparse recovery is to determine which subspace structure a point belongs to based upon the energy of the sparse coefficients used to decorate points across each class in the dataset [18]. If we assume that Ω_s denotes the set of points in Y that belong to class s, then for a point y_i with an endogenous sparse representation given by $c_i \in \mathbb{R}^k$, we

determine the class of the point by solving the following maximization problem

$$s^* = \max_s \sum_{j \in \Omega_s} |c_i(j)|,$$

where s^* corresponds to the class for which c_i contains the most energy. Applications of this approach include robust face recognition [18] and local image analysis [36].

CHAPTER 3

Greedy Feature Selection from Unions of Subspaces

In this section, we will introduce a generative model for our data and describe how greedy algorithms can be employed for selecting features from data living on unions of subspaces. Following this, we introduce an intuitive constraint that we will enforce on our procedure for greedy feature selection.

3.1 Preliminaries

3.1.1 Signal Model

Although in general, our data may live in some arbitrary subset of $\mathbb{R}^n$, we will assume that the data is centered about the origin and that the set is bounded such that it lives within the n-dimensional unit hypercube $[0,1]^n$. Given a set of p subspaces of $\mathbb{R}^n$, $\mathcal{W} = \mathcal{W}_1, \ldots, \mathcal{W}_p$, each of dimension less than or equal to k, we generate 'subspace clusters' by sampling d_i points in $\mathbb{X}_i = \mathcal{W}_i \cap [0,1]^n$. Let $\widetilde{\mathcal{S}}_i$ denote the resulting set of points and let $\widetilde{\mathcal{Y}} = \cup_{i=1}^{p} \widetilde{\mathcal{S}}_i$ denote the union of these p sets.

We define the mapping $g : \mathbb{R}^n \to \mathbb{S}^{n-1}$ from a point $y \in \mathbb{R}^n$ to the unit sphere

$\mathbb{S}^{n-1}$ as follows

$$g(\mathcal{S}) = \left\{ \frac{y}{\|y\|_2} | y \in \mathcal{S}, y \neq 0 \right\}.$$

If we apply this mapping to each of points in our set $\mathcal{S}$, we may write the mapping of each of our subspace clusters onto the sphere as $\mathcal{S}_i = g(\widetilde{\mathcal{S}})$ and their union as $\mathcal{Y} = \cup_{i=1}^{p} g(\mathcal{S}_i)$.

3.1.2 Projective Space

In the sequel, we will be interested in studying the use of our dataset as a dictionary. Thus, the projective space provides a natural setting for our study, i.e., in the projective space we consider all points along a line to be equivalent. In particular, we will be interested in how well the points in $\mathcal{Y}$ cover the projection each of the subsets $\mathbb{X}_i$ onto $\mathbb{S}^{n-1}$. If we consider the mapping of each k_i-dimensional subspace $\mathbb{X}_i$ onto the unit sphere given by $g(\mathbb{X}_i)$, each surface $\mathbb{X}_i$ is mapped to a $(k_i - 1)$-dimensional ring that encircles the $n-1$-dimensional sphere. We show a mapping of a union of three planes (2D subspaces) in 3D to the sphere $\mathbb{S}^2$ in Figure 3.1.

To measure the degree to which the points in each subset $\mathcal{S}_k$ cover their span, we will define the covering radius of the set relative to the projective distance. The projective distance between two vectors u and v is defined relative to the acute angle between the vectors

$$\text{dist}(u, v) = \sqrt{1 - \frac{|\langle u, v \rangle|^2}{\|u\|_2 \|v\|_2}}. \tag{3.1}$$

In the projective space, the covering radius of the set $\mathcal{S}_k$ is defined as

$$\text{cover}_{g(\mathbb{X}_k)}(\mathcal{S}_k) \equiv \max_{u \in g(\mathbb{X}_k)} \min_{y \in \mathcal{S}_k} \text{dist}(u, y) \tag{3.2}$$

The covering radius can be interpreted as the size of the largest open ball that can be placed in $g(\mathbb{X}_k)$ without encompassing a point in the set $\mathcal{S}_k$. We provide a visualization

of the covering radius of subspaces in the projective space in Figure 3.1 and in the ambient space in Figure 3.2.

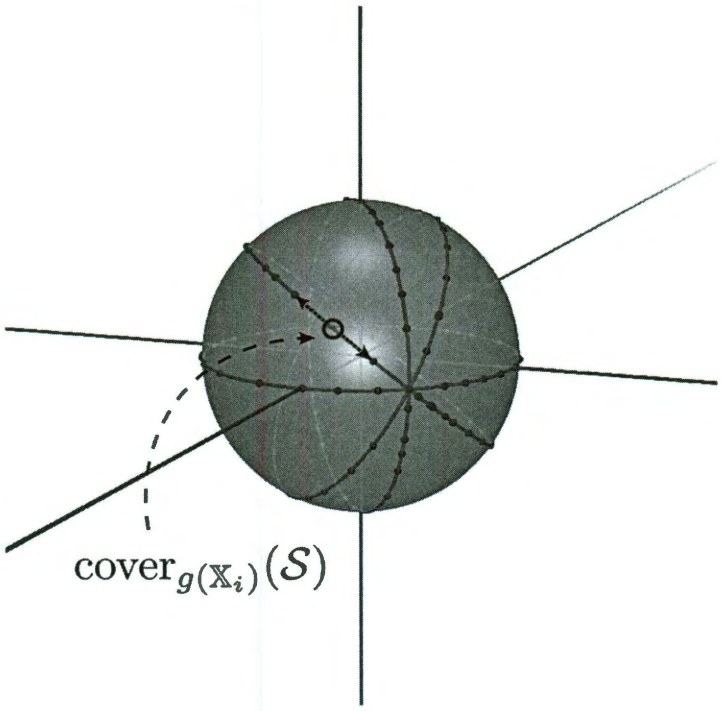


Figure 3.1: *Covering radius in the projective space.* A union of four 2D subspaces mapped to the unit sphere $\mathbb{S}^2$. The covering radius between two points in a subspace is shown.

3.2 Greedy Feature Selection

To form an endogenous sparse representation of a point from the set $\mathcal{S}_k$, we will employ a greedy algorithm known as orthogonal matching pursuit (OMP). At the first step, we find the point that is closest to our signal of interest in terms of its angular distance. If we consider this greedy selection for some point $y_i \in \mathcal{S}_k$, we will select the point from $\mathcal{Y} = \{y_j\}_{j=1}^d$ that maximizes this expression

$$j^* = \arg \max_{i \neq j} |\langle y_i, y_j \rangle|.$$

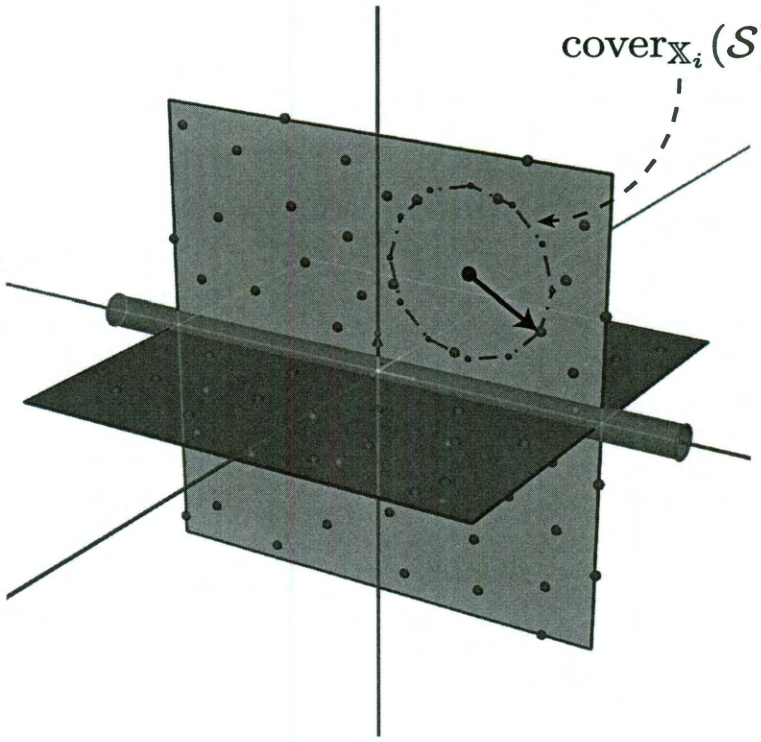


Figure 3.2: *Covering radius in the ambient space.* A union of two intersecting 2D subspaces in 3D. The covering radius between two points in a subspace is shown and the point that attains the covering radius is highlighted. We show an ϵ-tube around the subspace intersection.

We place this index in our feature set $\Lambda = j^*$ and update the residual by removing the projection of y_{j^*} onto y_i. To be precise, we set the residual to

$$r = y_i - \langle y_i, y_{j^*} \rangle y_i.$$

After computing the residual, we will look for the next point that is maximally correlated with the signal residual,

$$j^* = \arg \max_j \frac{|\langle r, y_j \rangle|}{\|r\|_2}.$$

This point is added to our feature set, $\Lambda = j^* \cup \Lambda$, and the residual is computed by projecting y_i into the space orthogonal to to the subspace spanned by the points in the current feature set. Formally, at the m^{th} step of the algorithm, the residual is computed as

$$r = y_i - P_\Lambda y_i,$$

where P_Λ is an ortho-projector onto the subspace spanned by the current feature set Λ. If we have knowledge that y_i lives on a k-dimensional subspace, this selection procedure is repeated k times or until the norm of the signal residual drops below a certain pre-specified threshold. Let $\Lambda^{(i)}$ denote the feature set selected for the i^{th} point in $\mathcal{Y}$.

3.3 Exact Feature Selection

In order to learn local subspace estimates from our ensemble, we will be interested in determining when the feature set $\Lambda^{(i)}$ returned by our greedy feature selection strategy contains points that all belong to the same subspace. We will refer to this event as *exact feature selection* (EFS). We now supply a formal definition.

Definition 4 (Exact feature selection) *Let $\Omega_k = \{y : y = P_k y,\ y \in \mathcal{Y}\}$ index the set of points that live in the span of $\mathbb{X}_k$, where P_k is an ortho-projector onto the span of $\mathcal{W}_k$. For a point $y_i \in \mathcal{W}_k$ with feature set $\Lambda^{(i)}$, we say that $\Lambda^{(i)}$ contains exact features if $\Lambda^{(i)} \subseteq \Omega_k$.*

Exact feature selection is essential for studying the performance of algorithms for unsupervised subspace learning problems, because when EFS occurs for a point in the set, this will yield a subspace estimate that coincides with one of the true subspaces contained within the data. For this reason, EFS provides a natural metric for studying

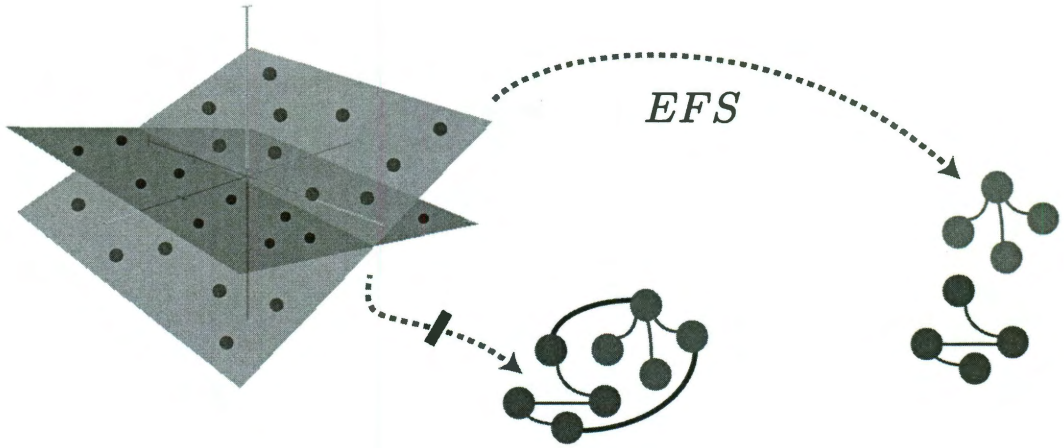


Figure 3.3: *Demonstration of exact feature selection (EFS) from unions of subspaces.* On the left, we show a union of intersecting planes and points tiling each of these planes. To form an ℓ_0-graph from these points, we place an edge between any two points that use one another in their sparse representations. To guarantee accurate subspace identification, the ℓ_0-graph must have a minimal number of edges linking points that live in different subspaces. On the top, we show a hypothetical ℓ_0-graph for which the points all admit exact features and below, we show a graph where EFS is violated (two vertices from different subspace clusters are connected).

the performance of both subspace consensus and subspace clustering methods that are based upon feature sets drawn from the data.

EFS is also important in supervised learning problems that rely on sparse representations of the data to determine the class membership of a new point as in [18]. In supervised learning problems, EFS is not required to guarantee accurate classification but is sufficient to ensure that accurate classification occurs; however, by studying the fundamental properties of the ensemble that govern EFS, we may also understand supervised classification methods that employ endogenous sparse recovery more deeply as well.

CHAPTER 4

EFS from Unions of Disjoint Subspaces

In this chapter, we will develop sufficient conditions that guarantee that EFS occurs for points in a particular subset $\mathcal{S}_k$. To do this, we must characterize the properties of the data related to the covering of each subspace as well as the geometry of pairs of subspaces in the ensemble. Before proceeding, we will quickly introduce the notion of principal angles between subspaces.

4.1 Principal Angles

To characterize the interaction between pairs of subspaces in the ensemble, the *principal angles* between subspaces will prove useful. The first principal angle $\theta_{i,j}$ between k-dimensional subspaces $\mathcal{W}_i$ and $\mathcal{W}_j$ is defined as the smallest angle between a pair of unit vectors (u_1, v_1) drawn from $\mathcal{W}_i \times \mathcal{W}_j$.

$$\theta_0 = \min_{(u_1,v_1)\in\mathcal{W}_i\times\mathcal{W}_j} \arccos \langle u_1, v_1 \rangle \quad \text{subject to} \quad \|u_1\|_2 = 1, \|v_1\|_2 = 1. \tag{4.1}$$

The vector pair (u_1^*, v_1^*) that attain this minimum are referred to as the first set of principal vectors. The second principal angle is defined much like the first, except

that the second set of principle vectors that define the second principal angle, are required to be orthogonal to the first set of principal vectors (u_1^*, v_1^*). The remaining principal angles are defined recursively in this way. The sequence of $k = \max(k_i, k_j)$ principal angles, $\theta_{i,j}^1 \leq \theta_{i,j}^2 \leq \cdots \leq \theta_{i,j}^k$, is non-decreasing and all of the principal angles lie between $[0, \pi/2]$.

A pair of subspaces is said to be disjoint if the minimum principal angle is greater than zero. This implies that the only point of intersection between disjoint subspaces is the origin. Non-disjoint or overlapping subspaces are defined as subspaces with minimum principal angle equal to zero.

Let $Y \in \mathbb{R}^{n \times d}$ denote our data matrix, where we have simply stacked the points in $\mathcal{Y}$ into the columns of $Y = [Y_1 \ldots Y_p]$ such that all of the points from a particular subset $\mathcal{S}_i$ are placed into a contiguous block. The sub-matrix $Y_i \in \mathbb{R}^{n \times d_i}$ which contains the points in $\mathcal{S}_i$, can be expanded in terms of an ONB $\Phi_i \in \mathbb{R}^{n \times k_i}$ that spans $\mathcal{W}_i$ and subspace coefficients $A_i \in \mathbb{R}^{k_i \times d_i}$, where $Y_i = \Phi_i A_i$. We will refer to the singular values of the matrix $G = \Phi_i^T \Phi_j$ as the *cross-spectra* of the subspace pair $(\mathcal{W}_i, \mathcal{W}_j)$. Formally, the cross-spectra is defined to be a k-dimensional vector in the unit hypercube, $\sigma \in [0, 1]^k$, where $k = \min(k_i, k_j)$. The cross-spectra is intimately related to the k principal angles between the subspace pair. In particular, the cross-spectra can be written in terms of the sorted principal angles, where $\sigma(m) = \cos(\theta_{ij}^m)$.

We define the mutual coherence with respect to the sets $\mathcal{S}_k$ and $\mathcal{S}_j$ as follows

$$\mu_c(\mathcal{S}_j, \mathcal{S}_k) \quad \equiv \quad \max_{u \in \mathcal{S}_j, v \in \mathcal{S}_k} |\langle u, v \rangle|. \tag{4.2}$$

By definition, the minimum principal angle bounds the mutual coherence. In particular, the mutual coherence $\mu_c(\mathcal{S}_j, \mathcal{S}_k) \leq \cos(\theta_{jk})$, where θ_{jk} denotes the minimum principal angle between subspaces $\mathcal{W}_j$ and $\mathcal{W}_k$.

4.2 Greedy Selection Lemma

Now that we have introduced relevant definitions needed to develop sufficient conditions for EFS, we will develop the main Lemma that will lay the foundation for our subsequent analysis of greedy feature selection.

Let us assume that Λ represents the set of features that we have selected to represent a point $y \in \mathcal{S}_k$ at the previous $m-1$ iterations. Recall that at the m^{th} iteration of the algorithm, we select the point from $\mathcal{Y}$ that maximizes the normalized inner product with the signal residual $r = (I - P_\Lambda)y$.

To guarantee that we select a point from the correct set $\mathcal{S}_k$, we require that the inner product between the residual and another point in $\mathcal{S}_k$ is larger than the inner product between the residual and a point outside of $\mathcal{S}_k$; we denote the set of all such points as $\mathcal{S}_k^c = \mathcal{Y} \backslash \mathcal{S}_k$. Formally, we require that the following greedy selection criteria holds

$$\max_{v \in \mathcal{S}_k} |\langle r, v \rangle| \quad > \quad \max_{v \in \mathcal{S}_k^c} |\langle r, v \rangle|. \tag{4.3}$$

The following lemma provides a sufficient condition that guarantees that this selection criterion will hold at a particular iteration of OMP.

Lemma 1 *Suppose that r lies in the span of $\mathcal{W}_k$. Let θ_0 denote the minimum principal angle between $\mathcal{W}_k$ and all other subspaces in the ensemble. A sufficient condition for the selection criterion in (4.3) to hold is that*

$$\text{cover}_{\mathbf{X}_k}(\mathcal{S}_k) \quad < \quad \sin(\theta_0). \tag{4.4}$$

Proof. To guarantee that we select a point from $\mathcal{S}_k$, we seek a lower bound on the maximum normalized inner product between a signal that lies in the span of $\mathcal{W}_k$ and a point in the set $\mathcal{S}_k$. To do this, we will consider the unit norm signal $u^* \in \mathcal{W}_k$ that

attains the minimum correlation with all of the elements in the set $\mathcal{S}_k$

$$u^* = \arg \min_{u \in \mathcal{W}_k, \|u\|_2=1} \max_{y \in \mathcal{S}_k} \sqrt{1 - |\langle u, y \rangle|^2}.$$

We can relate the maximum inner product between our signal residual r is related to the covering radius of $\mathcal{S}_k$ as follows

$$\begin{aligned} \max_{y \in \mathcal{S}_k} \frac{|\langle r, y \rangle|^2}{\|r\|_2^2} &\geq \max_{y \in \mathcal{S}_k} |\langle u^*, y \rangle|^2 \\ &= 1 - \text{cover}^2_{\mathcal{W}_k}(\mathcal{S}_k). \end{aligned}$$

Since the covering radii of $g(\mathbb{X}_i)$ and $\mathcal{W}_k$ are equivalent in the projective space, i.e., $\text{cover}_{g(\mathbb{X}_k)}(\mathcal{S}_k) = \text{cover}_{\mathcal{W}_k}(\mathcal{S}_k)$, we conclude that

$$\max_{y \in \mathcal{S}_k} \frac{|\langle r, y \rangle|^2}{\|r\|_2^2} \geq 1 - \text{cover}^2_{g(\mathbb{X}_k)}(\mathcal{S}_k).$$

Now, our aim is to find an upper bound on the maximum inner product between our residual and a point outside of $\mathcal{S}_k$. We will use the fact that the maximum inner product between points in different subspaces is bounded by the minimum principal angle between pairs of subspaces

$$\max_{y \in \mathcal{S}_k^c} \frac{|\langle r, y \rangle|}{\|r\|_2} \leq \cos(\theta_0),$$

where $\theta_0 = \min_{j \neq k} \{\theta_{jk}^1\}$ is the smallest principal angle shared between $\mathcal{W}_k$ and any other subspace in the union. This bound holds because the minimum principal angle defines the smallest angle (or maximum correlation) that any two points from different subspaces can share.

4.3 Exact Feature Selection Theorem

The greedy selection lemma that we developed in the previous section enables us to develop our main theorem for EFS from disjoint subspaces.

Theorem 1 *Let θ_0 denote the minimum principal angle between $\mathcal{W}_k$ and all other subspaces in the ensemble. A sufficient condition for EFS to occur for all $y \in \mathcal{S}_k$, is that the covering radius*

$$\text{cover}_{\mathbb{X}_k}(\mathcal{S}_k) < \sin(\theta_0). \tag{4.5}$$

Proof. We will prove this theorem by induction. At the first iteration, one can show that this condition easily leads to correct selection of a point in $\mathcal{S}_k$ in the first iteration. This is due to the fact that the mutual coherence between points is also upper bounded by the minimum principal angle. Now, suppose that after m iterations, we have already selected $m-1$ points from the optimal subset $\mathcal{S}_k$ to included in our feature set Λ. The residual at the m^{th} step of the algorithm equals the original signal $y \in \mathcal{S}_k$ minus a linear combination of $m-1$ points that also lie in $\mathcal{S}_k$. Since all of these points lie in the span of $\mathcal{W}_k$, then by our induction hypothesis, the residual also lies in $\mathcal{W}_k$. We note that we have not assumed that the residual is contained in the original subset of the space $\mathbb{X}_i$ where the points $\mathcal{S}_i$ are confined to. However, based upon our assumption that the points in $\mathcal{S}_k$ provide a covering of the image of $\mathbb{X}_i$ in the projective space, i.e., tile a $(k-1)-$dimensional ring on $\mathbb{S}^{n-1}$, we simply need apply the Greedy Selection Lemma to guarantee that we select a point from $\mathcal{S}_k$ at each iteration. Since our sufficient condition enforces a global sampling constraint on the points in $\mathcal{S}_k$, this condition is also sufficient to guarantee that EFS occurs for all points in the subset $\mathcal{S}_k$. □

Remarks. Although computing the covering radius of a set is in general a difficult problem, this theorem provides us with a geometric interpretation of what is happening at each step of our greedy selection algorithm. In particular, at each iteration, we seek a point that is close to our signal residual in angular distance. However, because we restrict our residual to be orthogonal to all of the points selected at previous iterations, this requires that we select a new point that is sufficiently different from the previous features. For this reason, we require that for any residual formed during our recovery procedure, is closer to a point in the correct subspace than any point in a different subspace. This naturally imposes a sampling constraint on our subspace clusters—namely, if we do not have a covering of our space, it is likely that we select a point from the incorrect subspace. The minimum principal angle provides a natural constraint on how close each point in our sets must be.

CHAPTER 5

EFS for Uniformly Bounded Unions

In this chapter, we extend our results for EFS to the case where we have a uniformly bounded union of subspaces.

5.1 Uniformly Bounded Unions

The sufficient condition that we developed in the previous section revealed an interesting relationship between the covering of each subspace in the set and the minimum principal angle between the subspaces in the ensemble. However, we have yet to reveal any dependence upon principal angles beyond the minimum angle. To make the connection between the geometry of our subspace union more apparent, we will make additional assumptions on the 'spread' of our principal vectors and distribution of points in the subspace. In particular, we will assume that our principal vectors and subspace coefficients are uniformly bounded.

To make this precise, we will consider the singular value decomposition (SVD) of $G = \Phi_i^T \Phi_j = U \Sigma V^T$, where $\Phi_i \in \mathbb{R}^{n \times k_i}$ is an ONB that spans $\mathcal{W}_i$ and the left and right singular vectors in $U \in \mathbb{R}^{k_i \times k_i}$ and $V \in \mathbb{R}^{k_j \times k_j}$ are referred to as the *principal vectors* between $\mathcal{W}_i$ and $\mathcal{W}_j$. Let $\mathcal{U} = \{u_m\}$ and $\mathcal{V} = \{v_m\}$ denote the set of left

and right principal vectors contained in the columns of U and V respectively. We can write the points in each set with respect to the same ONB, where $y \in \mathcal{S}_i$ may be expressed as $y = \Phi_i \alpha$ and the points $y \in \mathcal{S}_j$ can be expressed as $y = \Phi_j \beta$. Let $\mathcal{A} = \{\alpha_m\}_{m=1}^{d_i}$ and $\mathcal{B} = \{\beta_m\}_{m=1}^{d_j}$ denote the subspace coefficients for points in $\mathcal{S}_i$ and $\mathcal{S}_j$ respectively.

We will assume that the entries of the principal vectors and subspace coefficients are uniformly bounded such that they satisfy the following property

$$\max_{\alpha \in \mathcal{A}} |\langle u, \alpha \rangle|, \max_{\beta \in \mathcal{B}} |\langle v, \beta \rangle| \quad < \quad \gamma \qquad \forall\, u \in \mathcal{U},\ \forall\, v \in \mathcal{V}. \tag{5.1}$$

We will refer to ensembles that satisfy this property as 'uniformly bounded unions' of subspaces. in words, this constraint requires that the inner products between our subspace coefficients and the principal vectors of G are all bounded by the constant $\gamma \in (0, 1]$. If our principal vectors are 'spread' or that U and V are uniformly bounded as one assumes in [37] to provide guarantees for matrix completion, then our constraint above may easily be satisfied when paired with weak constraints on the magnitude of the subspace coefficients for points in $\mathcal{S}$.

5.2 EFS Lemma for Bounded Unions

Under the assumption that our union is uniformly bounded, we can prove the following Lemma.

Lemma 2 *Assume that we have a uniformly bounded union of subspaces $\mathcal{W}_i$ and $\mathcal{W}_j$ as defined in (5.1) with bounding constant $\gamma < \sqrt{1/r}$, where $r = \text{rank}(G)$. Let $\sigma \in \mathbb{R}^k$ denote the cross-spectra of the union. The maximum normalized inner product*

between $r \in \mathcal{W}_i$ and a point in $\mathcal{S}_j$ is bounded by

$$\max_{y \in \mathcal{S}_j} \frac{|\langle r, y \rangle|}{\|r\|_2} < \gamma \|\sigma\|_1. \tag{5.2}$$

Proof. We are interested in providing an upper bound on the maximum coherence between our residual $r \in \mathcal{W}_i$ and points in the set $\mathcal{S}_j$. Our aim is to exploit the fact that our principal vectors and subspace coefficients are bounded to find a tighter bound than the one obtained by bounding the mutual coherence with the minimum principal angle.

Since we have assumed that $r \in \mathcal{W}_i$, we can write the residual as $r = \Phi_i \widehat{\alpha}$. Similarly, we can write all points $y \in \mathcal{S}_j$ as $y = \Phi_j \beta$, where $\|\beta\| = \|y\|_2 = 1$ because Φ_j is a unitary matrix which preserves the ℓ_2-norm of y. We can expand our inner product as follows

$$\begin{aligned}
\max_{y \in \mathcal{S}_j} \frac{|\langle r, y \rangle|}{\|r\|_2} &= \max_{\beta \in \mathcal{B}} \frac{|\langle \Phi_i \widehat{\alpha}, \Phi_j \beta \rangle|}{\|\widehat{\alpha}\|_2} \\
&= \max_{\beta \in \mathcal{B}} \frac{|\langle \widehat{\alpha}, \Phi_i^T \Phi_j \beta \rangle|}{\|\widehat{\alpha}\|_2} \\
&= \max_{\beta \in \mathcal{B}} \frac{|\langle \widehat{\alpha}, U \Sigma V^T \beta \rangle|}{\|\widehat{\alpha}\|_2} \\
&= \max_{\beta \in \mathcal{B}} \frac{|\langle U^T \widehat{\alpha}, \Sigma V^T \beta \rangle|}{\|\widehat{\alpha}\|_2} \\
&\leq \max_{\beta \in \mathcal{B}} \frac{\|U^T \widehat{\alpha}\|_\infty}{\|\widehat{\alpha}\|_2} \|\Sigma V^T \beta\|_1.
\end{aligned}$$

Where the last step comes from an application of Holder's inequality, i.e., $|\langle w, z \rangle| < \|w\|_\infty \|z\|_1$.

First, we will tackle the term on the right, which we can write as $\|\Sigma V^T \beta\|_1 = \|\Sigma \widehat{\beta}\|_1$, where we assume that $\widehat{\beta}$ is a bounded unit-norm vector. This term can be simplified by writing it as an inner product between the cross-spectra $\sigma = \text{diag}(\Sigma)$ and $\widehat{\beta}$, where we have assumed that $\|\widehat{\beta}\|_\infty = \gamma \in (0, 1]$ and that $\|\widehat{\beta}\|_2 = 1$.

To develop an upper bound on this quantity, we seek the maximum of this constrained linear program with constraint set $\mathcal{C} = \{\beta \in \mathbb{R}^k : \|\beta\|_2 = 1,\ \|\beta\|_\infty \leq \gamma\}$,

$$\beta^* = \arg \max_{\beta \in \mathcal{C}} \quad \sigma^T \beta. \tag{5.3}$$

Suppose that $\gamma^2 r < 1$, where $r = \text{rank}(G) = \|\sigma\|_0$. In this case, we can maximize the expression above by setting the n^{th} entry of $\beta^*(n) = \gamma$ whenever $\sigma(n) \neq 0$. In this case,

$$\max_{\widehat{\beta} \in \mathcal{C}} \ \|\Sigma \widehat{\beta}\|_1 = \gamma \|\sigma\|_1.$$

When we relax our constraint on the maximum entry of β, then the ℓ_2-norm provides an upper bound on this quantity

$$\max_{y \in \mathcal{S}_j} \ \|\Sigma \Phi_j^T y\|_1 \leq \max_{\|\widehat{\beta}\|_2 = 1} \ \|\Sigma \widehat{\beta}\|_1 = \|\sigma\|_2.$$

This bound is due to the fact that when we relax our constraint on the maximum entry of β and only require that it is unit norm, $\beta^* = \sigma / \|\sigma\|_2$. Note that our bound that depends on $\|\sigma\|_1$ can be made arbitrarily small by requiring $\gamma \ll 1$. However, when we relax our constraint on the maximum value of γ, the resulting bound is uninformative because in general, $\|\sigma\|_2$ is greater than the cosine of the minimum principal angle.

When $\gamma^2 r < 1$, we can plug this bound into our original expression

$$\max_{y \in \mathcal{S}_j} \ \frac{|\langle r, y \rangle|}{\|r\|_2} \leq \gamma \|\sigma\|_1 \frac{\|U^T \Phi_j^T r\|_\infty}{\|r\|_2} \tag{5.4}$$

$$= \gamma \|\sigma\|_1 \frac{\max_{u \in \mathcal{U}} |\langle u, \alpha \rangle|}{\|\alpha\|_2}. \tag{5.5}$$

$$= \gamma \|\sigma\|_1 \|U\|_{2,2} = \gamma \|\sigma\|_1. \tag{5.6}$$

Since we assumed that $\gamma < \sqrt{1/r}$, this implies that $\gamma\|\sigma\|_1 < \sqrt{1/r}\|\sigma\|_1$. This constraint also requires that $\gamma\|\sigma\|_1 < 1$. This completes our proof. □

5.3 Theorem for EFS from Bounded Unions

This lemma, coupled with Theorem 4.5 for EFS, enables us to develop the following sufficient for EFS from uniformly bounded unions of subspaces. To do this, we simply need replace our earlier upper bound which depended upon the minimal principal angle with our new bound that depends upon the trace norm of G or equivalently, the ℓ_1-norm of the cross-spectra.

Theorem 2 *Assume that we have a uniformly bounded union of subspaces as defined in (5.1) with bounding constant $\gamma < \sqrt{1/r}$, where $r = \text{rank}(G)$. Let $\sigma \in \mathbb{R}^k$ denote the cross-spectra of the union. A sufficient condition for EFS to occur for all of the points in $\mathcal{S}_i$, is that the covering radius in the projective space*

$$\text{cover}_{g(\mathbf{x}_i)}(\mathcal{S}_i) < \sqrt{1 - \gamma\|\sigma\|_1}.$$

Remarks. To interpret this condition, we observe that when we have 'uniformly bounded unions', this allows us to bound the maximum inner product between points in different subspaces. When we have sufficient separation between points in different subspaces, this allows us to relax our constraint on the covering of the subspaces that we required which was based upon the minimum principal angle.

To contrast this condition with our earlier result, this condition nicely reveals the connection between EFS and higher order principal angles. This suggests that when the points in our sets are sufficiently spread along each subspace structure, the decay of the cross-spectra is likely to play an important role in determining whether points

from each set will admit EFS or not. This condition also suggests that unions with different cross-spectral decay properties are likely to behave differently in terms of their respective probability of EFS. To test this hypothesis, we will study the role that the cross-spectra plays in EFS in the following section.

CHAPTER 6

Empirical Study of EFS from Unions of Subspaces

In the previous section, we revealed an intimate connection between the covering of subspaces in the ensemble and the principal angles between subspaces in the ensemble. We will now conduct an empirical study to explore the dependence both on the cross-spectra between pairs of subspaces (geometry of the subspaces) as well as the density and distribution of points along each subspace (sampling of subspaces).

6.1 Generating Structured Unions of Subspaces

In order to study the probability for EFS for unions of subspaces with structured cross-spectra, we will generate data from unions of overlapping *block-sparse signals.* We define our construction as follows: take two subsets of k atoms from a dictionary $\mathcal{D} = \{d_m\}_{m=1}^{d}$, $|\Omega_1| = |\Omega_2| = k$. Let $\Psi \in \mathbb{R}^{n \times k}$ denote the subset of atoms indexed by Ω_1 and let $\Phi \in \mathbb{R}^{n \times k}$ denote the subset of atoms indexed by Ω_2.

We will select our sub-dictionaries Ψ and Φ such that $G = \Psi^T \Phi$ is diagonal, i.e., $\langle \psi_i, \varphi_j \rangle = 0$, if $i \neq j$. In this case, the cross-spectra is defined as $\sigma = \text{diag}(G)$, where $\sigma \in [0, 1]^k$. We assume that the 'overlap' or the rank of $G = \Psi^T \Phi$ is fixed to $q \in [0, k)$.

To generate a pair of k-dimensional subspaces with a q-dimensional intersection,

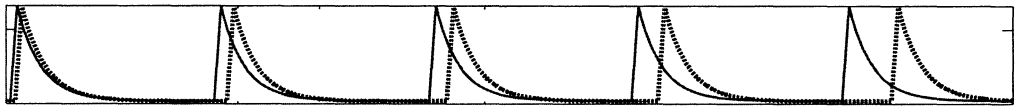


Figure 6.1: *Generating unions of subspaces from shift-invariant dictionaries.* An example of a collection of two sub-dictionaries of five atoms, where each of the atoms share some non-zero inner product with one other element. This produces cross-spectra with overlap equal to the subspace dimension, i.e., $q/k = 1$.

we can pair the elements from Ψ and Φ such that the i^{th} entry of the cross-spectra equals

$$\sigma(i) = \begin{cases} |\langle \psi_i, \varphi_i \rangle| & \text{if } 1 \leq i \leq q, \\ 0 & \text{if } i = q+1 \leq i \leq k. \end{cases}$$

We can leverage the banded structure of shift-invariant dictionaries to generate subspaces with arbitrary cross-spectra as follows. First, we fix a set of k incoherent atoms from our shift-invariant dictionary $\mathcal{D}$, which we place in the columns of Ψ. We set the i^{th} atom of our second sub-dictionary to be a shifted version of the i^{th} atom ψ_i. To be precise, if we set $\psi_i = d_m$, where d_m is the m^{th} atom in our shift-invariant dictionary, then we will set $\varphi_i = d_{m+\Delta}$ for a particular shift Δ. By varying the shift Δ, we can easily control the coherence between ψ_i and φ_i. In Figure 6.1, we show an example of one such construction for $k = q = 5$.

Since $\sigma \in (0, 1]^k$, the worst-case q-dimensional union that we can construct is when we pair q of the same atoms and $k - q$ orthogonal atoms. In this case, the cross-spectra attains its maximum over its entire support and equals zero otherwise. We will refer to this class of block-sparse signals as *orthoblock sparse signals.*

6.2 Influence of Cross-spectra on EFS

In this section, we study the impact that the cross-spectra plays on EFS. For our experiments, we generate pairs of subspaces from shift-invariant dictionaries as we describe in the previous subsection. We show the cross-spectra arising from three different unions of block-sparse signals along the top row of Figure 6.2. On the left, we show the cross-spectra for a orthoblock sparse signal model with $q/k = 3/4$. We show cross-spectra attained from pairing shifted Lorentzian and exponential atoms in the middle and right respectively.

We generate 'subspace clusters' by sampling m points from the span of the subspaces generated by each of our two sub-dictionaries Ψ and Φ. Denote the set of points generated from the first k-dimensional subspace as $Y_1 = \Psi A_1 \in \mathbb{R}^{n \times m}$ and the second k-dimensional subspace as $Y_2 = \Phi A_2 \in \mathbb{R}^{n \times m}$. For all of our experiments, we generate the subspace coefficients independently at random according to a standard normal distribution and then map all of the points in Y_1 and Y_2 to the unit sphere. We set $k = 20$ and $m = 100$.

In Figure 6.2, we show the average probability of EFS for these three subspace unions as we vary the overlap between subspaces. For a q-dimensional intersection, we select the first q elements from the full cross-spectra shown in Figure 6.1for $q/k = 1$ and set the remaining $k - q$ elements to be orthogonal.

Remarks. The results of this study are striking. In particular, we observe very different behavior for each of the three unions. For orthoblock sparse signals (worst-case unions), the probability of EFS for ℓ_0-graphs lies strictly above that obtained for the NN-graph, but the gap is relatively small. In the second union, where the cross-spectra exhibits nearly linear decay, both ℓ_0-graphs and NN-graphs maintain a high probability of EFS, with ℓ_0-graphs admitting nearly perfect feature sets P(EFS)$\approx$ 1,

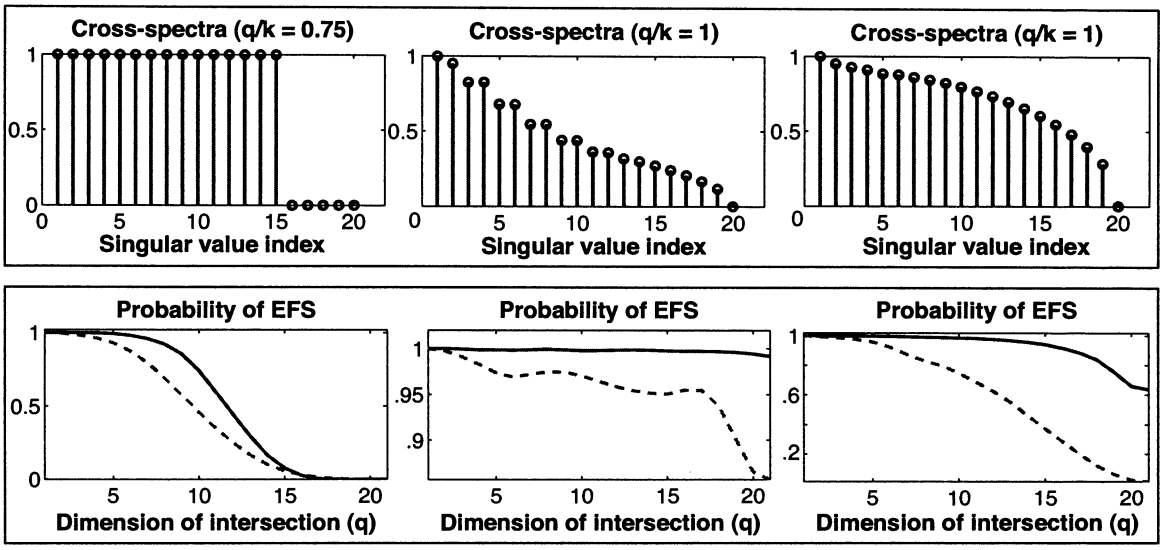


Figure 6.2: *Probability of EFS for unions with structured cross-spectra.* In the top row, we show the cross-spectra for different unions of block-sparse signals. On the bottom, we show the probability of EFS averaged over 50 trials as we vary $q \in [0, k]$ for ℓ_0-graphs (solid) and NN-graphs (dash).

even when the overlap ratio is maximal or where $q/k = 1$. The gap between ℓ_0 and NN-graphs is most pronounced for the third union with superlinear decay. In this example, we see the probability of EFS for the NN graph plunge to around $p = 0.1$, while the ℓ_0-graph maintains a very high probability of EFS even when the overlap ratio $q/k = 1$.

This study provides a number of interesting insights into the role that higher-order principal angles between subspaces play in feature selection. These results further support our claims that in order to truly understand and predict the behavior of endogenous sparse recovery across unions of subspaces, we require a description that relies on the entire cross-spectra.

6.3 Phase Transitions for EFS

In this section, we will study the probability of EFS as we vary the density and distribution of points along each subspace in the ensemble. To study the probability of EFS as we vary the overlap and sampling, we will study the *phase transitions* in the probability of EFS as we vary the distribution of points along each subspace and the relative overlap between the subspace unions. By visualizing the probability of EFS in this way, we can more easily study the behavior of greedy selection from pairs of overlapping subspaces. Of particular interest to our study will be a characterization of the:

1. *Phase boundary*: the contour that separates the phase space into regions where the $P(EFS) = 1$ and where the $P(EFS) < 1$. When we traverse this boundary, we transition between regions where all of points in the ensemble admit exact features (exact recovery for all) and regions where EFS occurs for some of the points in the set.

2. *Transition width*: the area of the phase space where $0 <$ P(EFS) < 1. We find that the phase transitions in EFS are not sharp (decay immediately to zero). Instead, the transition width tells us how quickly the probability of EFS decays as we increase the overlap between planes.

In Figure 6.3, we show the phase transitions for the probability of EFS for a union of orthoblock sparse signals for $k = 20$ and $k = 50$ on the left and right respectively. For these experiments, we generate data from each subspace by generating i.i.d. coefficients from a standard normal and mapping each point to the unit sphere. The results are averaged over 400 trials.

6.3.1 Oversampling of Subspaces

In the top row of Figure 6.3, we show how the probability of EFS varies as we increase the overlap ratio $q/k \in [0, 1]$ in conjunction with the oversampling ratio $k/m \in [0, 1]$, To see the rapid shift in the phase boundary when we approach critical sampling $m = k$, we display these results in terms of the logarithm of the oversampling ratio.

Remarks. In this study, we observe that the oversampling ratio has a big impact on the phase boundary for EFS. When the subspaces are densely sampled, i.e., $m \gg k$, the phase boundary is shifted dramatically from $q/k \in (0, 0.7]$. This result seems to confirm our covering arguments in Section 4.2, where we studied the interplay between the covering of the space and the overlap between subspaces. In particular, as we sample each subspace more densely, the covering of the space becomes sufficient to ensure that even when the overlap between planes is high, we will still select exact features. In contrast, when we approach critical sampling, where $m = k$, the phase boundary is shifted all the way back to $q/k = 0.1$.

As we increase the oversampling of each subspace, we also observe that the width of the transition region increases as the oversampling ratio increases. This suggests that there is a smooth transition between exact recovery (all points admit exact features) and the point where no points admit EFS as we vary the overlap; the transition width or smoothness of this transition seems to be tightly coupled with the oversampling.

6.3.2 Energy in Subspace Intersections

In the bottom row of Figure 6.3, we show the phase transitions for EFS as we vary the overlap ratio $q/k \in [0, 1]$ and the amount of energy that each point contains within the subspace intersection which we denote by $\tau \in [0, 1)$. To generate points with restricted energy in their intersection, we generate points with gaussian coefficients

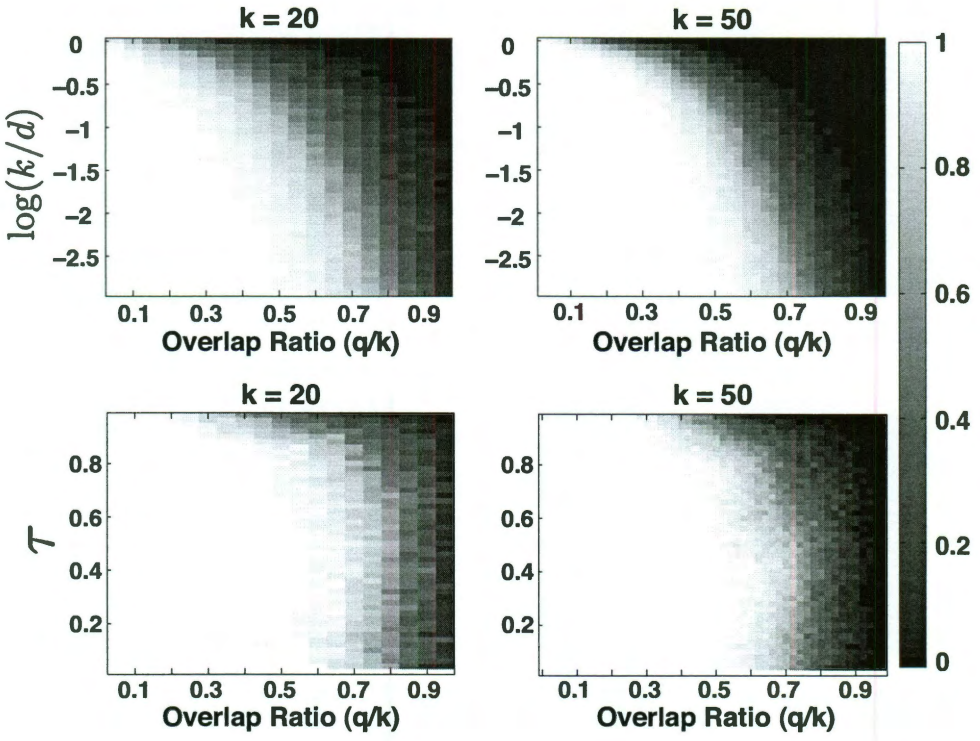


Figure 6.3: *Probability of EFS for Different Sampling Conditions.* We study the probability of EFS for unions of subspaces of dimension $k = 20$ (left column) and $k = 50$ (right column). On top, we show the probability of EFS as we vary the logarithm of the oversampling ratio $\log(k/d)$ versus the overlap ratio q/k. Below, we show the probability of EFS as a measure of the amount of energy contained within the intersection $\tau \in [0, 1)$ versus the overlap ratio $q/k \in [0, 1)$.

and then normalize the points such that the energy contained within the overlapping blocks is limited to a fixed value of τ. The energy in the remaining $k - q$ orthogonal blocks is set to $1 - \tau$. We set $m = 200$ for these experiments.

Remarks. Surprisingly, we find that while the amount of energy that points have in their intersection does play a role in EFS, the effect that this parameter has on EFS is much less pronounced than in the previous experiment. In particular, we observe a nearly constant phase boundary at $q/k \approx 0.9$ as we vary τ. It is not until the energy exceeds $\tau > 0.6$ that this parameter has a significant impact on the probability of

EFS. Even after points have more than 80% of their energy in their intersection, the phase boundary remains at around $q/k = 0.7$. This is quite surprising because even when points have nearly all of their energy in the intersection, we can still reliably obtain support sets that admit exact features.

6.3.3 Comparison of ℓ_0 and NN-graphs

In Figure 6.4, we compare the phase transitions for EFS for (left) ℓ_0-graphs and (right) nearest neighbor graphs as we vary the relative dimension of the intersection $q/k \in [0, 1]$ and the oversampling ratio $k/m \in [0, 1]$. For our simulations, we consider unions of orthoblock sparse signals for $k = 50$ and vary $k/m \in [0.2 \to 0.96]$ and $q/k \in [1/k, 1]$.

Remarks. An interesting result of this study is that there are regimes where EFS does not occur for NN-graphs but occurs with a non-trivial probability for ℓ_0-graphs. In particular, when subspaces exhibit high degrees of overlap for $q/k > 0.6$, the probability of EFS for nearest neighbor graphs quickly decays to zero. In contrast, ℓ_0-graphs provide feature sets with non-zero probability of EFS.

When the oversampling of the space is high, then the gap between ℓ_0-graphs and NN graphs shrinks. This implies that when we have a dense sampling of unions of orthoblock sparse signals, nearest neighbor graphs often provide similar estimates to that acquired from ℓ_0-graphs. On the other hand, when the sampling of the space is sparser, ℓ_0-graphs admit EFS with significantly higher proportion. Our study of EFS for different cross-spectra in Section 6.2 suggests that the gaps between nearest neighbor graphs and ℓ_0-graphs will be even more pronounced for subspaces with superlinear cross-spectral decay.

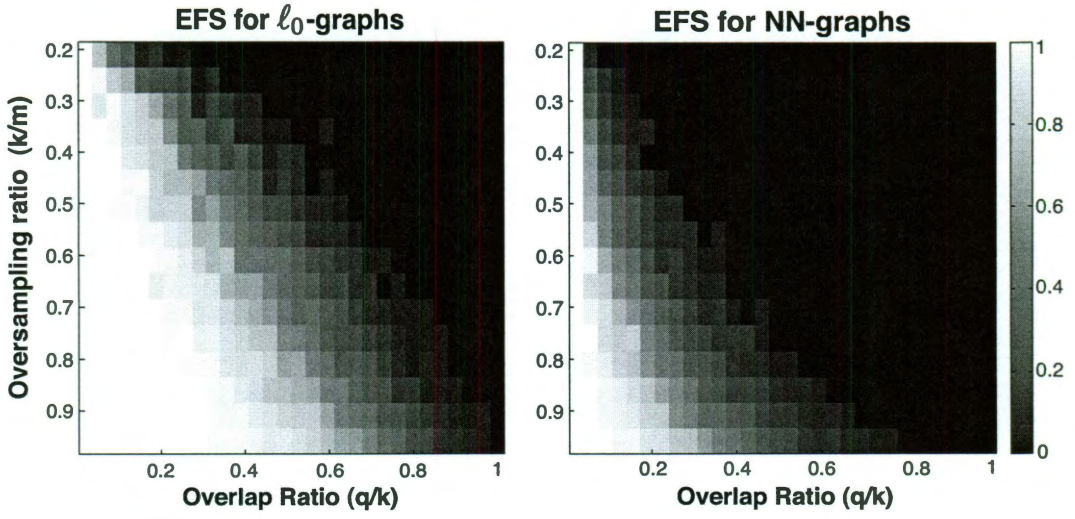


Figure 6.4: *Phase transitions for sparse recovery and NN-graphs.* We compare the probability of EFS for orthoblock sparse signals for the ℓ_0 (left) and NN (right) graphs as we vary the oversampling ratio k/m and the overlap ratio q/k.

CHAPTER 7

Methods for Learning Unions of Subspaces

Until now, we have primarily been concerned with studying EFS; this lead to the development of sufficient conditions for EFS and an empirical study of the probability of EFS for different subspace unions. Now, we will study the utility of greedy feature selection for finding solutions to subspace clustering problems when overlap exists between subspaces in the ensemble. Following this, we introduce a new algorithm for subspace consensus on the ℓ_0-graph and demonstrate that this method outperforms spectral clustering formulations in the presence of high degrees of overlap.

7.1 Methods for Subspace Learning

There are two main ingredients in most of the existing state-of-the-art methods for solving unsupervised subspace learning problems. First and foremost, we require an affinity matrix or some other structure that encodes the pairwise distances amongst points in the ensemble. We have already demonstrated that ℓ_0 and NN graphs are both attractive proposals for revealing the subspace connectivity between points; other affinities include the polar curvature of the ensemble used in spectral curvature clustering [10] and nearest-neighbors selected within beta-neighborhoods in [12].

The second ingredient that we require is a technique for forming an estimate of the subspaces present in the ensemble from our subspace affinity matrix.

Most techniques for finding estimates of the subspace structures present in a data ensemble fall into one of two classes. In the first class lie clustering-based approaches, where spectral clustering is performed on the graph Laplacian of our appropriately chosen subspace affinity matrix. The other class of estimation methods employ voting procedures or *consensus* to look for agreement across multiple subspace estimates to determine the most likely estimate.

In contrast to clustering approaches which view the graph as encoding the subspace connectivity between points in the set, in consensus approaches, the goal is to utilize the *geometric features* contained in the edges of the graph. To be precise, for each vertex we determine the set of vertices for which an edge exists and map this sample set onto the Grassmanian manifold (set of all k-dimensional subspaces in $\mathbb{R}^n$). By looking at the span of these points, we obtain an estimate of a subspace structure that may be present in the ensemble. The idea is that by looking at a number of such mappings for different vertices in the graph, we can quickly converge to a correct estimate of the subspaces in the ensemble by finding the mode in the mappings. We point the reader to Vidal's review in [38] on subspace clustering for a thorough description of the subspace clustering problem as well as methods for obtaining solutions to this problem.

7.2 Clustering or Consensus?

We would now like to explore the implications of EFS on the performance of subspace recovery algorithms that employ either spectral clustering or a consensus-based estimation procedure. EFS is intimately linked to the probability that we exactly recover the subspaces present in our ensemble. In particular, consensus methods are guaran-

teed to recover the subspaces present in the ensemble, as long there are a sufficient number of points that admit exact features across the dataset. Thus, the probability of EFS provides an explicit lower bound for the probability of recovering a sufficient number of correct local subspace estimates; this will in turn lead to accurate recovery of the subspaces in our union.

In contrast, even when all of the points in the set admit exact features, this is not sufficient to guarantee that spectral clustering based methods like sparse subspace clustering (SSC) [13] will recover the correct set of subspaces from the data. This is due to the fact that even for graphs with no links across subspaces, spectral clustering or graph cuts may still be unwieldy due to scaling and normalization issues. In practice, we find that in a number of settings, spectral clustering over the ℓ_0-graph will often recover small clusters containing less than k points from the same subspace. Even after removing these points, the same issues arise in subsequent iterations. This results in clusterings consisting of a large number of small clusters, from which the true union of subspaces underlying the data can not be ascertained. This issue is even more pronounced in real-world datasets.

In contrast to clustering-based approaches, consensus methods are designed in a way such that they to remain robust to this degradation in the probability of EFS. Thus, consensus methods lend themselves well to settings where the probability of selecting sets of points with exact features is bounded above zero but not equal to one. For this reason, consensus approaches provide a natural means by which we can obtain efficient solutions to subspace learning problems when high-degrees of overlap exist or in settings where each point in the set can not be guaranteed to admit exact features.

7.3 Consensus on the ℓ_0-graph

In our numerical studies of EFS, we found that there is a smooth transition in the probability of EFS as we vary the overlap between subspaces. To exploit this smooth phase transition in the probability of EFS, we propose the following subspace recovery algorithm which we refer to as *consensus on the ℓ_0-graph.* The main idea behind this method is to simply replace step (1) in standard consensus-based methods that use sets of near neighbors [12] with the feature sets selected via OMP. In contrast to clustering-based approaches where no guarantees can be made, when all of the points in the ensemble admit support sets with exact features, our consensus-based approach is guaranteed to recover the true subspaces underlying the data. This method is very similar in spirit to the iterative subspace identification approach proposed by Gowreesunker et al. in [34]. We detail our proposed method in Algorithm 2.

7.4 Experimental Results

In Figure 7.1, we compare the performance of our ℓ_0-consensus approach with the equivalent spectral clustering formulation on the ℓ_0-graph proposed in [13]. We also compare these methods with a slightly modified version of SSC, where instead of clustering the eigenvector corresponding to the smallest non-zero eigenvalue of the graph Laplacian, we select the set of k largest and k smallest entries in this vector. These sets corresponds to two sets of k points from each cluster that are most separated with respect to their edge weights on the ℓ_0-graph.

In Figure 7.1, we observe similar behavior in the subspace recovery from ℓ_0-graphs our slight modification to SSC and for ℓ_0-consensus. However, when $q/k \geq 0.8$, we see a quick drop in the probability of recovery for modified SSC and we maintain a non-zero probability of recovery with ℓ_0-consensus, even when $q/k = 0.9$. In contrast,

Algorithm 2 : Subspace Consensus on the ℓ_0-graph

Input: An ensemble of d data points $Y \in \mathbb{R}^{n \times d}$, subspace dimension k, number of points required for consensus s, threshold λ.

Output: A collection of ONBs $\{Q_i\}_{i=1}^{p}$, and the number of points that agree upon each of the p subspace estimates $\mathcal{N} = \{n_i\}_{i=1}^{p}$, where $n_i \geq s$ for all i.

Solve the support recovery problem in (1.2) for Y to obtain a collection of support sets $\mathcal{S}(\mathcal{Y}) = \{\Lambda^{(i)}\}_{i=1}^{d}$.

for $i = 1 \rightarrow d$ **do**

1. Compute an orthonormal basis Q_i for which $\text{range}(Y_{\Lambda^{(i)}}) = \text{range}(Q_i)$.
2. Compute the energy of points in the sub-dictionary $Y_{\Lambda^{(j)}}$ when projected onto the subspace spanned by Q_i

$$d(i,j) = \sum_{n \in \Lambda^{(j)}} \left(c(n)(I - Q_i Q_i^T) y_n\right)^2,$$

where $c_j(n)$ is the contribution of the the n^{th} point in $\Lambda^{(j)}$ to the representation of y_j.

3. Count the number of points that agree upon the i^{th} subspace estimate,

$$n_i = \sum_{j=1}^{d} = T_\lambda(d(i,j)),$$

where $T_\lambda(\cdot) = 1$ when its argument is less than λ and 0 otherwise.

end for

4. Place all unique projectors $Q_i Q_i^T$ for which $n_i \geq s$ into the set Γ.

return Subspace estimates $Q_{est} = \{Q_i\}_{i \in \Gamma}$ and the number of points that agree on each estimate, $\mathcal{N} = \{n_i\}_{i \in \Gamma}$.

when we perform standard spectral clustering, we observe a decrease in the probability of recovery when $q/k = 0.7$. These results suggest that ℓ_0-graphs provide reliable feature sets for both clustering and consensus, even for high degrees of overlap. However, consensus can be used in settings where high degrees of overlap exist to maintain reliable recovery performance even when spectral clustering methods begin to fail.

In Figure 7.2, we show the gap between modified SSC and ℓ_0-consensus when we vary s (the number of points that we require to form consensus). We see that the gap between these methods increases as we require less confidence in the estimate. However, if we require a large degree of confidence for our estimate $s > 10$, the

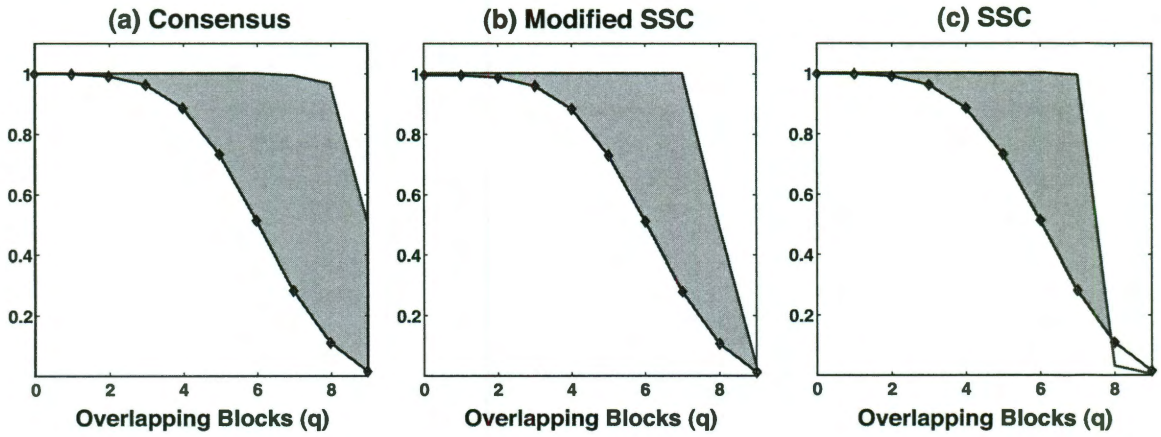


Figure 7.1: *Probability of subspace recovery.* The probability of recovery is shown for (a) ℓ_0-consensus (b) modified SSC, and (c) SSC. The empirical probability of EFS is displayed below these curves (dots) and the area between this curve and the probability of recovery is shaded. The results are averaged across 150 trials with $k = 10$, $d = 200$, $s = 5$, and $\lambda = 1e - 5$.

performance of ℓ_0-consensus is very similar to the performance of modified SSC as we vary the overlap between subspaces in the ensemble.

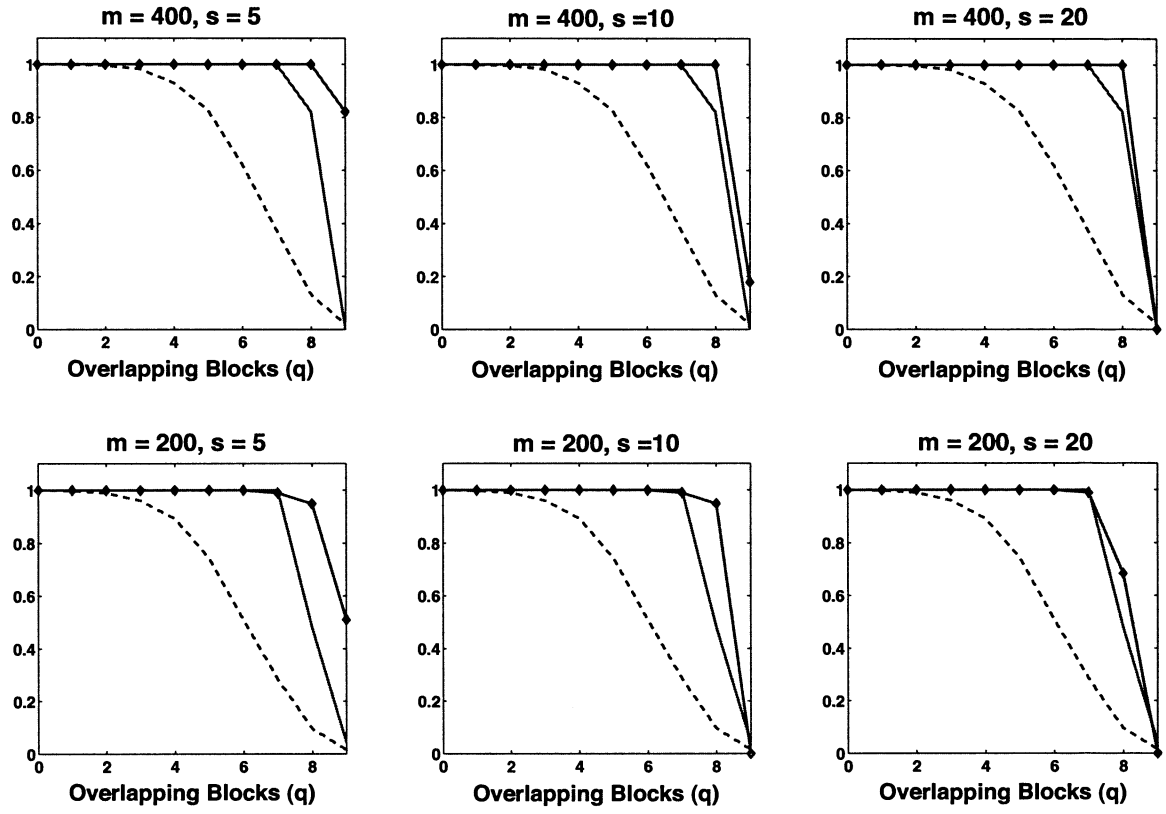


Figure 7.2: *Performance of subspace consensus.* In each plot, we overlay the probability of recovery for ($\diamond$) ℓ_0-consensus, (solid) modified SSC, as well as (dash) the empirical probability of EFS. The results are averaged across 50 trials with $k = 10$ and $\lambda = 1e-5$. On the (left) $s = 5$, (middle) $s = 10$, and (right) $s = 20$. The number of points in each subspace is set to $m = 200$ and $m = 400$ in the top and bottom rows respectively.

CHAPTER 8

Learning Unions of Subspaces from Image and Text Data

In this chapter, we will study the application on endogenous sparse recovery to real data and apply our ℓ_0-consensus algorithm to image and motion segmentation tasks.

8.1 Face Illumination Subspaces

We now compare the properties of ℓ_0 and NN-graphs for unions of 'illumination subspaces' arising from images of three different faces under various illumination conditions. If we fix the camera center and position of the persons face and capture multiple images under different lighting, the resulting images live on or are well-approximated by a 10-dimensional subspace [1].

In Figure 8.1, we show the affinity matrices obtained from the ℓ_0-graph and the NN graphs from a collection of 64 different images of 3 people that we have subsampled 4 times. All of the images are taken from the Yale Database B [39]. On the left, we show the NN graph obtained for $k = 10$ nearest neighbors, after projecting each subset of faces onto a 10D subspace. In the middle, we show the ℓ_0-graph obtained via OMP

on the raw data (no pre-processing). On the right, we show the ℓ_0-graph obtained via OMP after projecting the data onto a 10D subspace as in the NN graph on the left.

Remarks. Since there is actually a high-degree of overlap between each of the datasets, when no dimensionality reduction is performed, a proportion of points do not admit exact features on the ℓ_0-graph. However, once we project each collection of face images onto a 10D subspace with PCA, the resulting ℓ_0-graph has practically all of its energy concentrated in the correct block/cluster. In contrast to this drastic change that we observe in the ℓ_0-graph when dimensionality reduction is performed, the nearest neighbor graph admits the same probability of EFS after dimensionality reduction as the raw dataset; this is due to the fact that the nearest neighbors in the ensemble are effectively preserved after PCA.

This experiment provides a number of interesting insights into feature selection that we were not able to ascertain from synthetic experiments. In particular, we see hubs arise in our NN graph that link points from the wrong subspaces. Despite the fact that these highly structured intersections exist in the data, greedy feature selection with OMP manages to avoid these hubs and select points from the dataset that belong to the same subspace. In addition to avoiding hubs, we also observe that the representations formed for each point tend to be diverse and spread across the dataset (each cluster is more filled in).

8.2 Motion Segmentation Data

Motion segmentation is an important yet challenging problem in computer vision where one aims to segment different rigid body motion trajectories from one another directly from video sequences. Each trajectory may correspond to the motion of a different object or even the motion introduced from the camera. In can be shown that

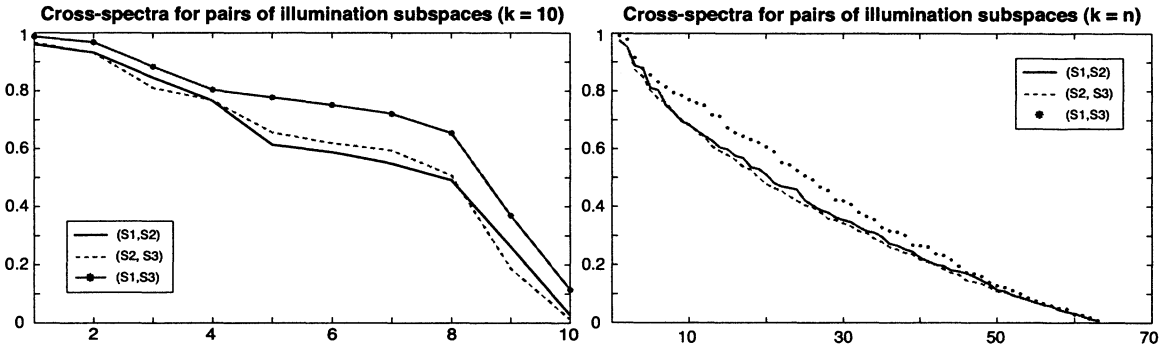


Figure 8.1: *Cross-spectra for illumination subspaces.* We display the cross-spectra for reduced data which has been projected onto a 10D subspace (left) and the cross-spectra obtained from the raw data (right). In each subplot, we overlay the cross-spectra between subspaces $(\mathcal{W}_1, \mathcal{W}_2)$ (solid) where $\|\sigma\|_2 = 2.10$, $(\mathcal{W}_2, \mathcal{W}_3)$ (dash) where $\|\sigma\|_2 = 2.12$, and between $(\mathcal{W}_1, \mathcal{W}_3)$ where $\|\sigma\|_2 = 2.30$ (star).

rigid body motion arising from point correspondences in multiple affine views live on a 5D affine hyperplane embedded in the ambient dimension. Thus, when multiple rigid body motions are combined within the field of view, the problem of motion segmentation boils down to learning subspaces from point correspondences and then segmenting the data in accordance with these learned hyperplanes.

In Figure 8.3 and 8.4, we compare the results obtained on the Hopkins155 database with ℓ_0-consensus (on the far right, labeled SSC-Grassman) to those obtained with other existing methods, including SSC. We note that both SSC and our method obtain state-of-the-art performance in comparison with other existing methods. In Figure 8.3, we show the classification performance obtained from segmenting video sequences with only two rigid body motions and in Figure 8.4 we show the results from segmenting three motions.

8.3 Multispectral Image Segmentation

In this section, we will apply endogenous sparse recovery to segment multispectral image data. Automated segmentation of multispectral image data is essential in many

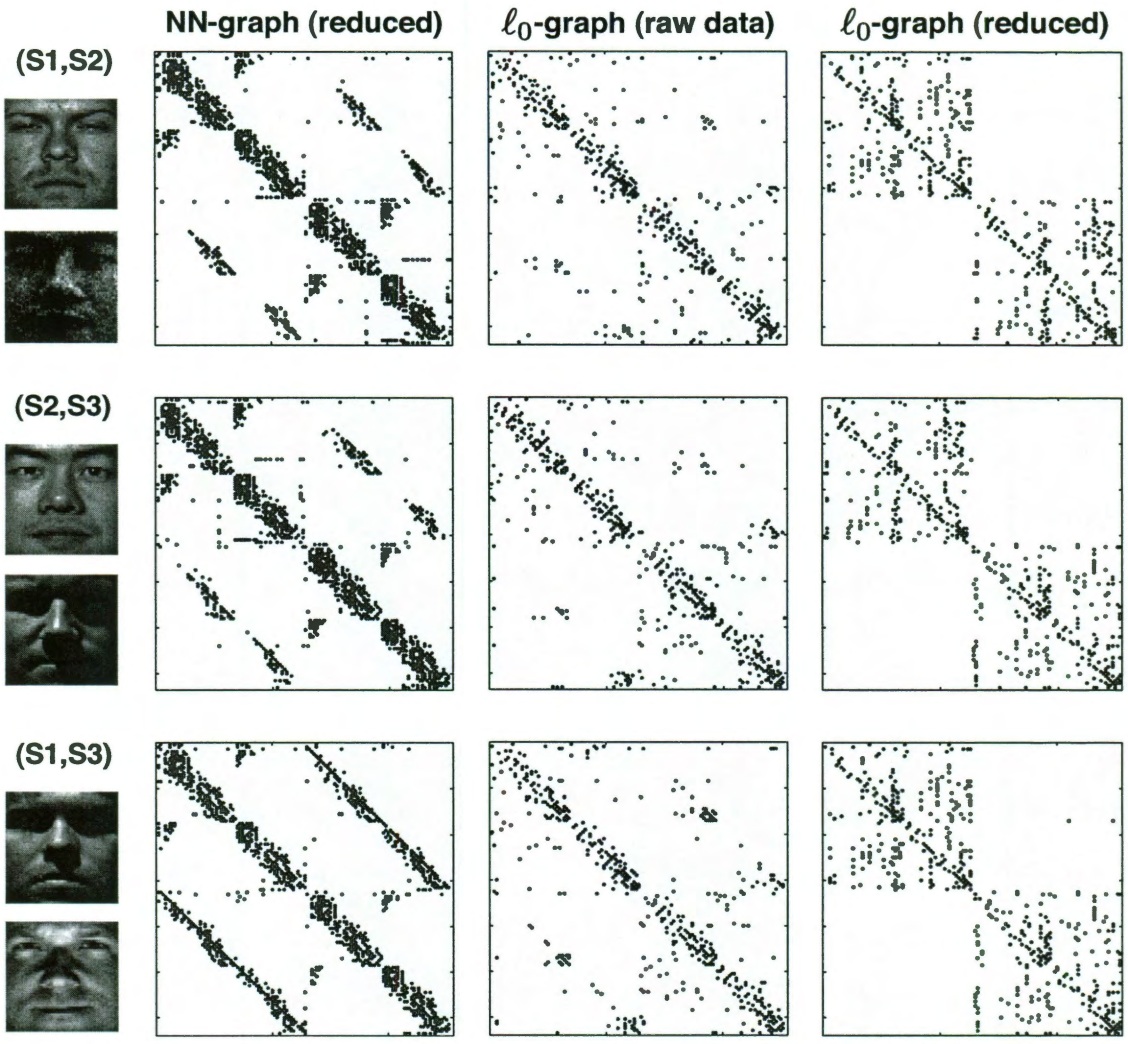


Figure 8.2: *Comparison of ℓ_0-graphs with NN-graphs.* In each row, we show affinity matrices for a different pair of subspaces in the dataset. In each row we show the (left) NN-graph, (middle) ℓ_0-graph for the raw dataset, and (right) ℓ_0-graph for the reduced dataset (each subspace is reduced to $k = 10$ dimensions with PCA prior to support recovery) for $(\mathcal{W}_1, \mathcal{W}_2)$ (top row), $(\mathcal{W}_2, \mathcal{W}_3)$(middle row) , and $(\mathcal{W}_1, \mathcal{W}_3)$ (bottom row), where $k = 10$. On the left of the affinity matrices are exemplar images from each illumination subspace.

applications where both spectraland spatial information can be jointly extracted from a sample. To be precise, for each point in space (pixel), we collect multispectral data which carries information about the absorption of a material at a particular wavelength of light in the visible range.

Checkerboard	GPCA	RANSAC	SSC-B	SSC-N	NS-k4	NS-k3	SC-Grass.
Average	6.09	6.52	0.83	1.12	0.22	0.35	0.92
Median	1.03	1.75	0	0	0	0	0
Traffic	GPCA	RANSAC	SSC-B	SSC-N	NS-k4	NS-k3	SC-Grass.
Average	1.41	2.55	0.23	0.02	0.92	2.24	0.33
Median	0	0.21	0	0	0.42	0.41	0
Articulated	GPCA	RANSAC	SSC-B	SSC-N	NS-k4	NS-k3	SC-Grass.
Average	2.88	7.25	1.63	0.62	2.33	2.36	1.64
Median	0	2.64	0	0	0.88	0.88	0
ALL	GPCA	RANSAC	SSC-B	SSC-N	NS-k4	NS-k3	SC-Grass.
Average	4.59	5.56	0.75	0.82	0.69	0.79	0.82
Median	0.38	1.18	0	0	0	0	0

Figure 8.3: *Classification performance for segmenting two motions.* We show the classification rates for the Hopkins155 Database for segmenting two rigid body motions from point correspondences.

Checkerboard	GPCA	RANSAC	SSC-B	SSC-N	NS-k4	NS-k3	SC-Grass.
Average	31.95	25.78	4.49	2.97	0.72	0.86	2.72
Median	32.93	26	0.54	0.27	0	0	0.98
Traffic	GPCA	RANSAC	SSC-B	SSC-N	NS-k4	NS-k3	SC-Grass.
Average	19.83	12.83	0.61	0.58	11.99	1.37	0.91
Median	19.55	11.45	0	0	4.07	1.46	0
Articulated	GPCA	RANSAC	SSC-B	SSC-N	NS-k4	NS-k3	SC-Grass.
Average	16.85	21.38	1.6	1.6	20.55	4.25	4.92
Median	16.85	21.38	1.6	1.6	20.55	4.25	6.38
ALL	GPCA	RANSAC	SSC-B	SSC-N	NS-k4	NS-k3	SC-Grass.
Average	28.66	22.94	3.55	2.45	4.11	1.15	2.46
Median	28.26	22.03	0.25	0.2	0.75	0.45	0.93

Figure 8.4: *Classification performance for segmenting three motions.* We show the classification rates for the Hopkins155 Database for segmenting three rigid body motions from point correspondences.

For our experiments, we study multispectral images [40], where each pixel in the image contains a 31-dimensional spectral representation in the visible light range. We show a single image from this database in the top row of Figure 8.5, for three different spectral bands. To be precise, the spectral bins range from 400nm to 700nm in 10nm increments. We select a random subset of 2500 pixels from the image for

training which corresponds to about 1% of the total $n = 512 \times 512$ pixels in the image. We treat each of these pixels (and its associated spectral vector) as a point in our ensemble, subtract the mean, and then normalize each vector. Following this, we apply endogenous sparse recovery to this set of exemplar spectra to learn a collection of two-dimensional subspaces. Following this, we segment the entire image based upon the nearest subspace to each pixel in the image.

We show these segmentation results for different number of classes in Figure 8.5. Interestingly enough, after including more than three subspaces in our representation, we are able to reliably segment the real human face from the photo of the person's face.

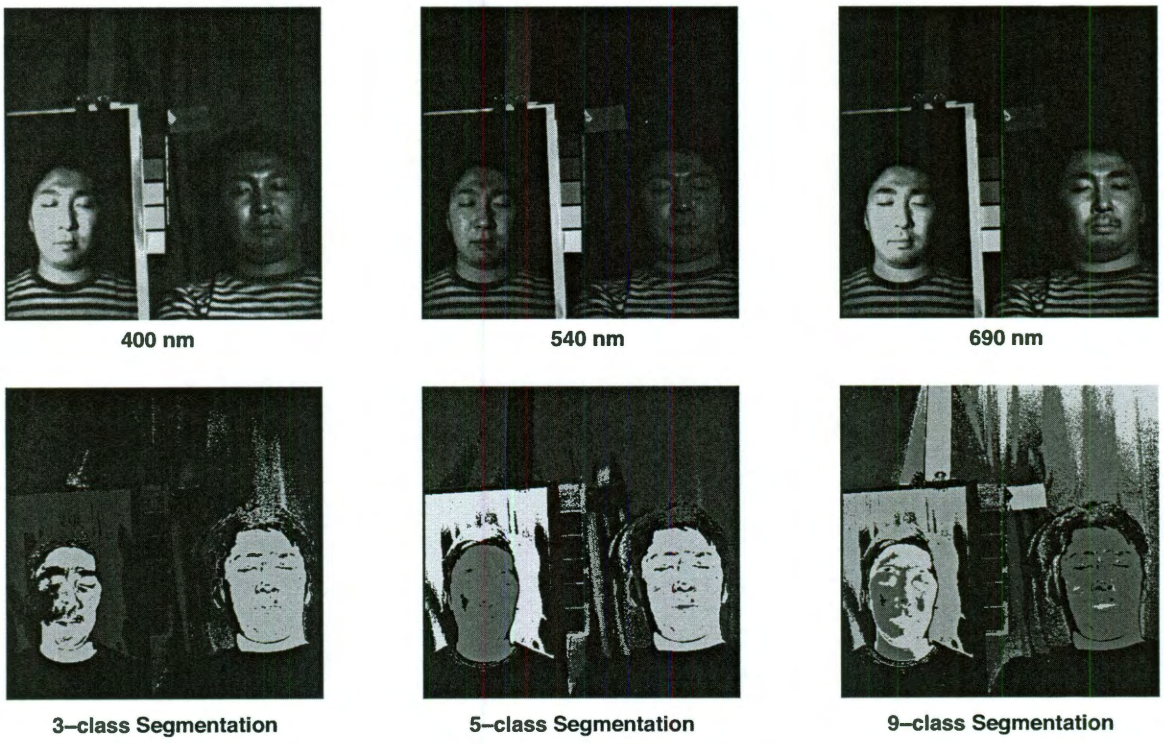


Figure 8.5: *Multispectral image segmentation.* Along the top row, we display multispectral image data for three spectral bands, (left) 400nm, (middle) 540nm, (right) 690nm. Along the bottom row, we show the segmentation results obtained via ℓ_0-consensus for (left) 3 classes, (middle) 5 classes, and (right) 9 classes.

8.4 Document Clustering

In this section, we will employ endogenous sparse recovery to study clustering within a corpora of documents. Each document in our collection is a different subsection from the textbook, "Fundamentals of Electrical Engineering" by Don Johnson [41].

8.4.1 Clustering Documents with ℓ_0-graphs

To study similarity across the subsections in the text, we treat each word that appears in the corpora as a separate coordinate and form a representation of each document with respect to the number of times a particular word in the global vocabulary set appear in the document. We study 92 subsections from the text over a reduced vocabulary (after removing stop-words and other uninformative words from the set) of 1952 words. By representing each document in terms of its word content, we can simply stack each document's word vector into a document matrix $Y \in \mathbb{R}^{n \times d}$, where $n = 1952$ and $d = 92$. We then normalize each vector such that it has unit ℓ_2 norm.

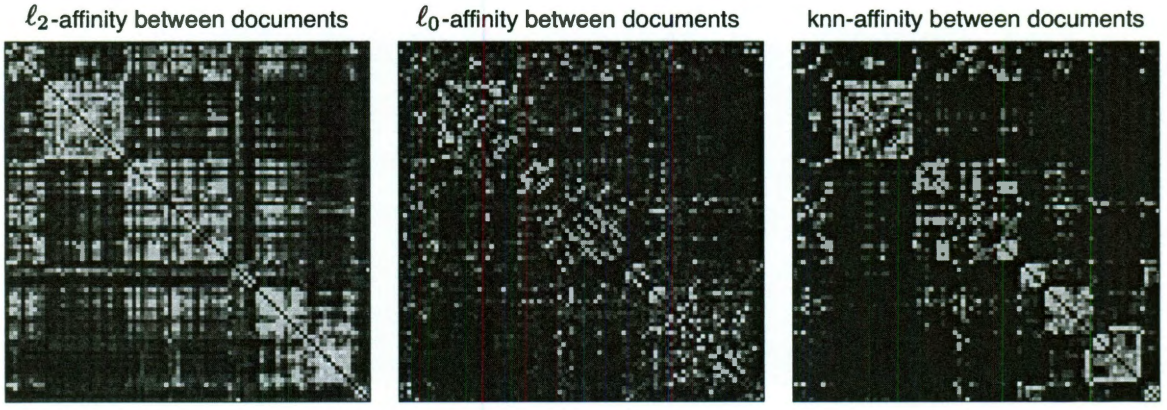


Figure 8.6: *Document affinity matrices.* On the left, we display the Gram matrix (ℓ_2 affinity) for each document in our corpora. In the middle, we display the affinity matrix for the ℓ_0-graph and on the right, we display the affinity for the NN graph, where $k = 7$.

The interpretation behind the ℓ_0 and ℓ_2-graphs formed across documents can be

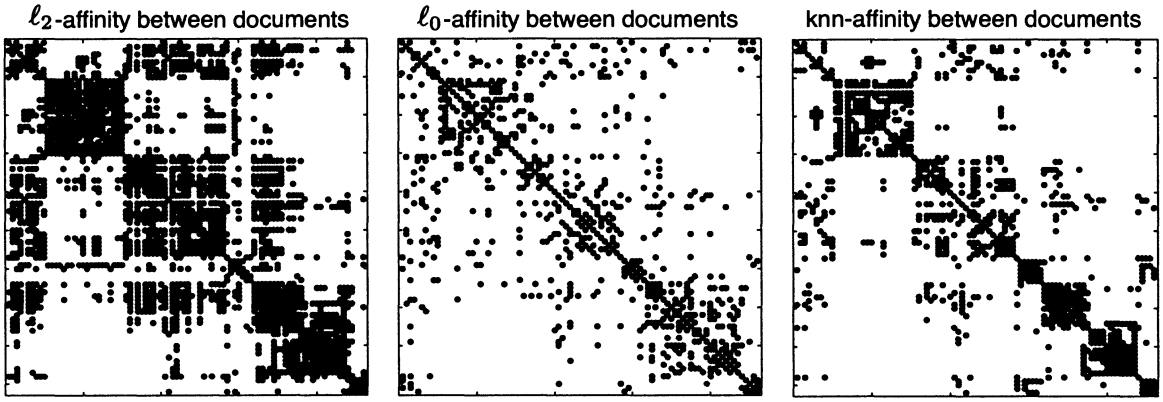


Figure 8.7: *Document support matrices.* On the left, we display the thresholded Gram matrix (ℓ_2 affinity) for each document in our corpora. In the middle, we display the thresholded affinity matrix for the ℓ_0-graph and on the right, we display the thresholded affinity for the NN graph, where $k = 7$ and the threshold is set to 0.3

thought of as follows. Whereas the nearest neighbor information between documents will tend to group two documents with similar word distributions (the intensity or proportion of a particular word), when forming sparse representations of a document, the absolute proportion of word counts is much less important. In particular, because we assume that each document can be written as a combination of other documents (relative to their normalized word counts), endogenous sparse recovery will tend to reveal subsets of documents that use the same vocabulary set (and thus have the same support in the word count space) rather than the same proportion of each word.

We show the image scaled version of the ℓ_0-graph from the ensemble in Figure 8.6 and show the structure of this affinity matrix when we threshold the graph such that only edge weights over $\tau = 0.3$ are displayed.

8.4.2 Visualizing Information Flow Across Documents

To visualize the ℓ_0 connectivity amongst documents in the corpora, we will now generate a graph that reveals the information flow across sections of the textbook. In

particular, we place all of the documents into different clusters based upon the chapter that they are contained in. In Figure 8.8, we separate documents (nodes) into clusters based upon the chapter they appear in chronological order. In the Figure, each of these clusters is separated along the horizontal axis. For each cluster, we scatter the document nodes about the chapter's centroid at random. By visualizing the ℓ_0-graph of the corpora in this way, we can more easily visualize information flow in the text—as well as visualize the concept map or connectivity between different chapters.

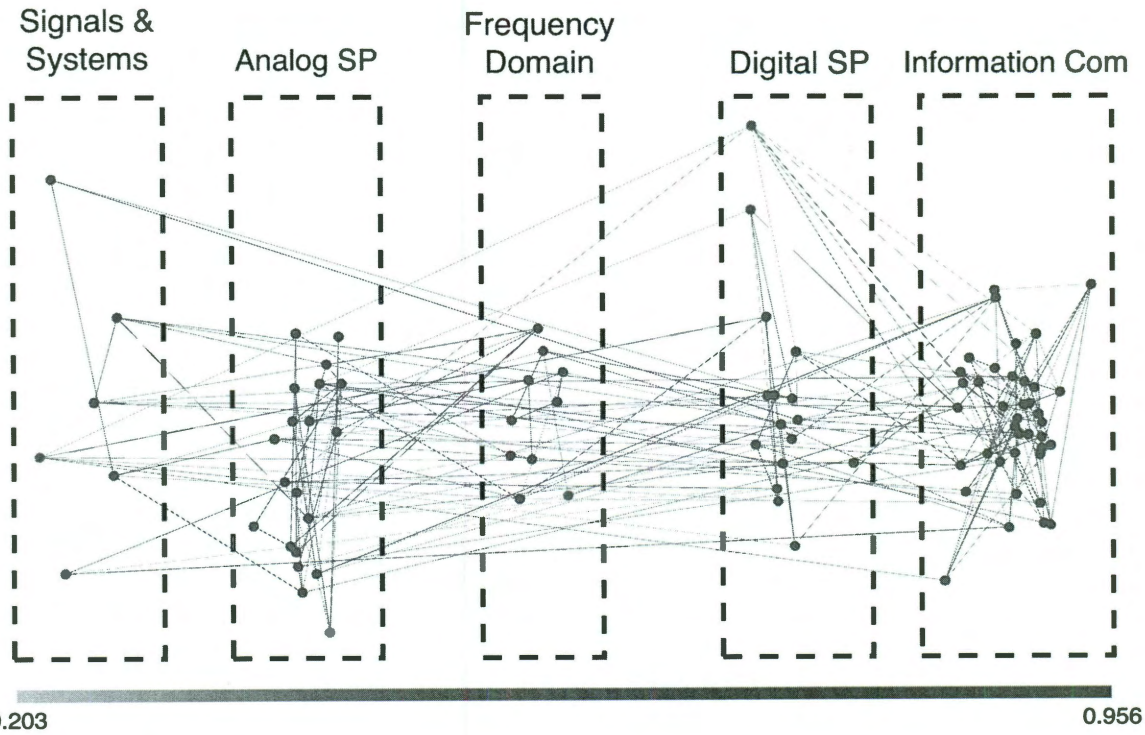


Figure 8.8: *Visualizing information flow across documents.* We display the edges in the ℓ_0-graph after partitioning the documents by chapter. The weight of each edge is given by the colorbar at the bottom and each cluster in the graph is labeled according to its corresponding chapter in the textbook.

We observe a number of interesting trends in this graph. In particular, we observe dense connectivity amongst documents in the last chapter of the book on Information Theory. In contrast to this clustered document connectivity, the documents contained

in the introduction on basic signals and systems has long range connections with concepts that arise throughout the text. Chapters 2-4 seem to exhibit a mixture of both chapter-specific clustering as well as links between each of these chapters and neighboring clusters.

CHAPTER 9

Discussion

9.1 Summary of Results

Let us now revisit the main contributions of this thesis which we outlined in Introduction; we are now equipped to summarize our results on each of these fronts.

1. *Theoretical analysis of greedy feature selection from unions of subspaces.* We developed sufficient conditions that describe when OMP will return feature sets that contain exact features. An interesting result of our analysis is a sufficient condition that highlights the tradeoff between the minimum principal angle between subspaces and the covering of each subspace. We provide an extension of these results to the case where we assume that we have a uniformly bounded union of subspaces. This enables us to reveal the connection between the sampling of each subspace and the entire distribution of the principal angles between subspaces in the ensemble required to guarantee EFS.

2. *Empirical study of EFS from unions of overlapping subspaces.* Following our analysis of greedy feature selection, we conducted an empirical study to explore the role that both the sampling and geometry of subspaces play in the proba-

bility of obtaining EFS. One of the most striking results of our empirical study is that the probability of EFS is strongly linked with the decay in the cross-spectra; in fact, the minimum principal angle (maximum value of cross-spectra) provides a very poor indicator of whether EFS will occur for an ensemble. We conjecture that the rate of decay of the cross-spectra may be the fundamental geometric quantity that governs whether EFS occurs for points in a given dataset. Thus an interesting question is whether we can accurately predict the phase transitions for EFS for a particular union of subspaces by studying another unions with the same cross-spectra. If this is indeed the case, we provide a simple way to create a wide-range of structured cross-spectral interactions from shift-invariant dictionaries which may be used in the future for large-scale studies of endogenous sparse recovery.

3. *Study and comparison of methods for learning unions of subspaces from local subspace estimates.* After studying EFS from unions of subspaces, we studied competing methods for learning unions of subspaces from local subspace estimates. We introduced a new algorithm for subspace consensus from local subspace estimates and provided theoretical justification for the utility of this method when high degrees of overlap exist between subspaces in the ensemble. We demonstrated that in the presence of overlap, consensus based approaches indeed outperform clustering-based formulations.

9.2 Implications of this Work

In this section, we discuss the implications of this work in a number of related areas, including discriminative dictionary learning, model-based CS, and sparse approximation.

9.2.1 Cross-spectral Minimization

In both theory and in practice, we find that the decay of the cross-spectra is strongly linked with the probability of EFS on the ℓ_0-graph. Since all of the unions that we have studied admit the same minimum principal angle, our study suggests that the spectral norm does not provide an adequate glimpse into the nature of the interactions between two collections of data living on unions of subspaces. Thus, in settings where we can manipulate the cross-spectral interaction between two collections of data, e.g., supervised classification [18] and discriminative dictionary learning [36], our analysis suggests that it is far more advantageous to reduce the ℓ_1-energy in the entire cross-spectra instead of simply minimizing the maximum coherence between points in different subspaces as in [36].

This finding opens up the possibility that instead of constraining dictionary learning and sensing matrix optimization in compressive sensing (CS) to minimize the maximum coherence between points in distinct classes, a superior strategy is to minimize the trace norm between the sub-matrices that correspond to points in each class. An interesting and relevant question is how one might impose such a constraint in discriminative dictionary learning methods.

9.2.2 'Data Driven' Sparse Approximation

The standard paradigm in signal processing and approximation theory is to compute a compact representation of a signal in a fixed and pre-specified basis or dictionary. In most cases, the dictionaries used to form these representations are designed according to some mathematical desiderata. A more recent approach has been to learn a dictionary from a collection of data that admit a sparse representation of all of the points in the ensemble.

The applicability and utility of endogenous sparse recovery in subspace learning

draws into question whether we can use endogenous sparse recovery in other tasks, including approximation and compression. The question that naturally arises is, "do we design, learn, or use the data directly?" Understanding the advantages and tradeoffs between each of these approaches is certainly an interesting and open question.

9.2.3 Learning Block Sparse Signal Models

Block-sparse signals and other structured sparse signals have received a great deal of attention over the past few years, especially in the context of compressive sampling from structured unions of subspaces [4, 5] and in model-based CS [6]. In all of these settings, one wishes to exploit the fact that signals admit structured support patterns to obtain improved recovery of sparse signals in noise and in the presence of undersampling.

However, to exploit such structure in sparse signals—especially in situations where the structure of signals or blocks of active atoms may be changing across different instances in time, space, etc.—the underlying subspaces must be learned directly from the data. Thus, the methods that we have described for learning union of subspaces from ensembles of data, can certainly be utilized in the context of learning block sparse and other structured sparse signal models.

9.3 Going Beyond Coherence

Our study is the first of its kind to uncover the connection between the principal angles between subspaces and the performance of sparse recovery methods from overcomplete dictionaries. In some cases, the principal angles between certain sub-dictionaries of atoms can resemble the cumulative coherence of the dictionary; however, the principal angles formed from pairs of sub-dictionaries provide an even richer description of the

geometric properties of a dictionary. A further exploration of the distribution of principal angles between sub-dictionaries could prove fruitful.

9.4 Open Questions

In this paper, we set out the understand some facets of the behavior of endogenous sparse recovery from unions of subspaces. In the end, we answered a number of these questions, in addition to uncovering a number of new questions that are likely to entertain us and (hopefully) other researchers for some time. Some of these questions and future lines of work include:

1. How can we characterize the average-case behavior of endogenous sparse recovery-based methods? How can we analytically characterize the phase transitions we observe empirically?

2. What are sufficient conditions for EFS from overlapping subspaces?

3. How does endogenous sparse recovery behave on noisy data? What about when ensembles are compressible or live near a union of subspaces, i.e., ℓ_p-balls for $p < 1$.?

4. How can we predict and characterize the "gap" between ℓ_0 and NN-graphs over unions of subspaces? As the cross-spectra varies? As the subspace dimension increases?

References

[1] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 2, pp. 218–233, February 2003. 1.1, 8.1

[2] K. Kanatani, "Motion segmentation by subspace separation and model selection," in *Proc. IEEE Int. Conf. Comp. Vis. (ICCV)*, 2001. 1.1

[3] K. Kanatani, "Evaluation and selection of models for motion segmentation," in *Proc. European Conf. Comp. Vision (ECCV)*, 2002. 1.1

[4] Y. M. Lu and M. N. Do, "Sampling signals from a union of subspaces," *IEEE Signal Processing Mag.*, vol. 25, no. 2, pp. 41-47, 2008. 1.1, 9.2.3

[5] T. Blumensath and M. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 4, pp. 1872–1882, 2009. 1.1, 9.2.3

[6] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp. 1982–2001, 2010. 1.1, 9.2.3

[7] B. Gowreesunker, A. Tewfik, V. Tadipatri, A. J., G. Pellize, and G. R., "A subspace approach to learning recurrent features from brain activity," *IEEE Trans. Neur. Sys. Reh*, vol. 19, no. 3, pp. 240–248, 2011. 1.1, 2.2

[8] E. Arias-Castro, G. Chen, and G. Lerman, "Spectral clustering based on local linear approximations," *arXiv:1001.1323v2 [stat.ML]*, September, 2010. 1.1

[9] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *Proc. European Conf. Comp. Vision (ECCV)*, 2006. 1.1

[10] G. Chen and G. Lerman, "Spectral curvature clustering," *Int. J. Computer Vision*, vol. 81, pp. 317–330, 2009. 1.1, 7.1

[11] T. Zhang, A. Szlam, Y. Wang, and G. Lerman, "Hybrid linear modeling via local best-fit flats," *arXiv:1010.3460v1 [cs.CV]*, October, 2010. 1.1

[12] T. Zhang, A. Szlam, Y. Wang, and G. Lerman, "Randomized hybrid linear modeling by local best-fit flats," in *Proc. IEEE Conf. Comp. Vis. Patt. Recog. (CVPR)*, June 2010. 1.1, 7.1, 7.3

[13] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Conf. Comp. Vis. Patt. Recog. (CVPR)*, June 2009. 1.2, 2.2, 7.2, 7.4

[14] E. Elhamifar and R. Vidal, "Clustering disjoint subspaces via sparse representation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, March 2010, pp. 1926–1929. 1.2, 1.2

[15] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1998. 1.2, 2.1.1, 2.1.2

[16] G. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency decompositions," *SPIE J. Opt. Engin.*, vol. 33, no. 7, pp. 2183–2191, 1994. 1.2, 2.1.1, 2.1.2

[17] R. Garg, H. Du, S. M. Seitz, and N. Snavely, "The dimensionality of scene appearance," in *Proc. IEEE Int. Conf. Comp. Vis. (ICCV)*, October 2009. 1.2

[18] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 2, pp. 210–227, 2009. 1.2, 2.2, 3.3, 9.2.1

[19] R. Jenatton, J. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *Technical Report*, 2009. 1.2

[20] G. Peyre and J. Fadili, "Group sparsity with overlapping partitions," in *Proc. Europ. Sig. Processing Conf. (EUSIPCO)*, 2011. 1.2

[21] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *J. Mach. Learn. Res.*, vol. 11, pp. 2487–2531, December 2010. 1

[22] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006. 2.1.1

[23] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006. 2.1.1

[24] E. Candès, "Compressive sampling," in *Proc. Int. Congress of Math.*, Madrid, Spain, Aug. 2006. 2.1.1

[25] J. Bobin, J. L. Starck, J. M. Fadili, Y. Moudden, and D. L. Donoho, "Morphological component analysis: An adaptive thresholding strategy," *IEEE Trans. Image Processing*, vol. 16, no. 11, pp. 2675–2681, 2007. 2.1.1

[26] D. L. Donoho and G. Kutyniok, "Microlocal analysis of the geometric separation problem," *Preprint*, 2010. 2.1.1

[27] K. Huang and S. Aviyente, "Sparse representation for signal classification," *Proc. Adv. in Neural Processing Systems (NIPS)*, 2006. 2.1.1

[28] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Statist. Soc B*, vol. 58, no. 1, pp. 267–288, 1996. 2.1.1, 2.1.2

[29] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993. 2.1.2

[30] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2009. 2.1.2

[31] J. Tropp, "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 1030–1051, 2006. 2.1.3

[32] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004. 2.1.3

[33] J. Tropp, "Topics in sparse approximation," *Ph.D. Dissertation, Computational and Applied Mathematics, Univ. Texas at Austin*, 2004. 2.1.3

[34] B. V. Gowreesunker and A. H. Tewfik, "Learning sparse representation using iterative subspace identification," *IEEE Trans. Signal Processing*, vol. 58, no. 6, pp. 3055–3065, 2010. 2.2, 2.2, 7.3

[35] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust Principal Component Analysis?," *Journal of the ACM*, vol. 58, no. 1, pp. 1–37, 2011. 2

[36] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conf. Comp. Vis. Patt. Recog. (CVPR)*, June 2008. 2.2, 9.2.1

[37] E. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *arXiv:0903.1476v1 [cs.IT]*, March 2009. 5.1

[38] R. Vidal, "Subspace clustering," *IEEE Signal Processing Mag.*, vol. 28, no. 2, pp. 52–68, 2011. 7.1

[39] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 6, pp. 643–660, 2001. 8.1

[40] F. Yasuma, T. Mitsunaga, D. Iso, and S. Nayar, "Generalized Assorted Pixel Camera: Post-Capture Control of Resolution, Dynamic Range and Spectrum," *Technical Report, Department of Computer Science, Columbia University CUCS-061-08*, November, 2008. 8.3

[41] D.H. Johnson, "Fundamentals of Electrical Engineering," *Available online at: 'http://cnx.org/content/col10040/latest/'*. 8.4