ABSTRACT

SKAU, ERIK WEST. Relaxations to Sparse Optimization Problems and Applications. (Under the direction of Agnes Szanto and Hamid Krim.)

Parsimony is a fundamental property that is applied to many characteristics in a variety of fields. Of particular interest are optimization problems that apply rank, dimensionality, or support in a parsimonious manner. In this thesis we study some optimization problems and their relaxations, and focus on properties and qualities of the solutions of these problems.

The Gramian tensor decomposition problem attempts to decompose a symmetric tensor as a sum of rank one tensors. We approach the Gramian tensor decomposition problem with a relaxation to a semidefinite program. We study conditions which ensure that the solution of the relaxed semidefinite problem gives the minimal Gramian rank decomposition.

Sparse representations with learned dictionaries are one of the leading image modeling techniques for image restoration. When learning these dictionaries from a set of training images, the sparsity parameter of the dictionary learning algorithm strongly influences the content of the dictionary atoms. We describe geometrically the content of trained dictionaries and how it changes with the sparsity parameter. We use statistical analysis to characterize how the different content is used in sparse representations. Finally, a method to control the structure of the dictionaries is demonstrated, allowing us to learn a dictionary which can later be tailored for specific applications.

Variations of dictionary learning can be broadly applied to a variety of applications. We explore a pansharpening problem with a triple factorization variant of coupled dictionary learning. Another application of dictionary learning is computer vision. Computer vision relies heavily on object detection, which we explore with a hierarchical convolutional dictionary learning model.

Data fusion of disparate modalities is a growing topic of interest. We do a case study to demonstrate the benefit of using social media data with satellite imagery to estimate hazard extents. In this case study analysis we apply a maximum entropy model, guided by the social media data, to estimate the flooded regions during a 2013 flood in Boulder, CO and show that the results are comparable to those obtained using expert information.

Relaxations to Sparse Optimization Problems
and Applications

by
Erik West Skau

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Applied Mathematics

Raleigh, North Carolina

2017

APPROVED BY:

Seth Sullivant	Dávid Papp
Agnes Szanto Co-chair of Advisory Committee	Hamid Krim Co-chair of Advisory Committee

DEDICATION

To my parents, John Skau and Judy West, and my partner, Alyssa Stubbs.

BIOGRAPHY

Erik Skau received B.S. degrees in Applied Mathematics as well as Physics (2010) in Raleigh, NC at North Carolina State University. In graduate school under the direction of Dr. Agnes Szanto and Dr. Hamid Krim, Erik pursued an applied mathematics degree.

Erik's research interests include machine learning techniques, especially when applied to designing efficient numerical algorithms for matrix factorizations and neural networks. Computer vision, image, and dataset analysis techniques for inverse problems such as denoising and super-resolution. Applications of computational geometry such as using Delaunay triangulations and Voronoi diagrams for meshing and geometric analysis.

ACKNOWLEDGEMENTS

I would first like to thank my advisors Dr. Agnes Szanto and Dr. Hamid Krim for their tremendous help and support in this process. They provided invaluable advice and guidance for both academic problems and with life's challenges. Additionally I would like to thank the rest of my committee, Dr. Edge Lobotan, Dr. Dávid Papp, and Dr. Seth Sullivant for both their suggestions and patience during my graduate career.

Additionally, I would like to thank Dr. Greg Bleckerham, Dr. Liyi Dai, Dr. Lek-Heng Lim, Dr. Bernard Mourrain, Dr. Jim Renegar, Dr. Michael Singer, and Dr. Brendt Wohlberg for their useful suggestions related to various sections of this thesis.

Finally, I would like to acknowledge my lab mates for both their support in shared struggles and insightful discussions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

Parsimony is a concept fundamental to many problems and applications. Parsimonious characteristics can emerge in a wide range of things from order, support, rank, and dimension. In practice one often formulates optimization problems to encourage parsimony, or strives for sparsity in some characteristic.

Optimization problems with a sparsity characteristic are notoriously hard to solve. Fortunately, relaxations of these problems often provide surrogate problems that are much more tractable. These tractable problem solutions are then taken as approximations to the ideal solutions. This relaxation procedure brings into question exactly how close the relaxed solution is to the ideal solution.

In this dissertation, we consider varying forms of relaxations to parsimonious optimization problems and applications. In some instances we consider properties of the original and not relaxed problems. In other instances, we consider the effect of sparsity on our models and propose modifications to the problems to yield a more desirable solution.

1.1 Outline

The rest of the thesis is organized as follows. In Chapter 2 we give an overview of ways to characterize and use sparsity in optimization problems. We describe a few applicable optimization problems with parsimonious characteristics. Typical parsimonious properties such as rank or support are measured with non-convex functions which present barriers for efficient optimization. Through

relaxation these non-convex problems can be replaced with convex surrogates. Some of these relaxations have properties suitable for algorithms to solve efficiently.

In Chapter 3 we look at the Gramian tensor decomposition problem. The tensor decomposition problem is formulated as a rank minimization optimization problem, which is relaxed to a nuclear norm minimization problem. Our interest in this research is to find cases where the relaxed nuclear norm solution is also of minimal rank. We then provide some specific cases where the optimal nuclear norm solution is also minimal rank and discuss the intricacy of the problem in other cases.

Chapter 4 introduces dictionary learning and some popular methods to solve dictionary learning problems. Dictionary learning has a range of applications in image restoration problems. We consider a geometric interpretation of dictionary learning and the role of sparsity in the dictionary learning model. We present some of these applications, and proposed modifications to dictionary learning.

Chapter 5 explores two applications using variations of dictionary learning. In this chapter, coupled dictionary learning is reformulated to include an additional factorization and is applied to a pansharpening problem. We also develop a proof of concept for a hierarchical convolutional dictionary learning method to construct more complex dictionary filters.

In Chapter 6 we consider a data fusion problem for hazard extent estimation. We explore the suitability of using social media data to estimate the extent of inundated areas using a regularized maximum entropy model. We compare the effectiveness of using social media data against expert knowledge and discuss the filtering required of the social media.

In Chapter 7 we conclude our discussion and suggest future research.

CHAPTER

2

SPARSITY AND OPTIMIZATION BACKGROUND

This chapter provides some of the main concepts and mathematical tools used to understand and formulate applicable optimization problems.

2.1 Principle of Sparsity

In a variety of fields, data is effectively of low order, support, rank, or dimension. In the last several decades, optimization problems applying these characteristics have become a widely used tool for a variety of applications [99, 126, 140]. The underlying principle of parsimony is common to all of these applications, simply applied to different characteristics. In this chapter we review some of the mathematical formulations of these notions for their incorporation into optimization problems. Since these notions are closely related, we concentrate on the notion of support size, or sparsity, and describe the relation between support and the other notions.

2.1.1 Characterizing Sparsity with Norms

One way to understand the notion of support or sparsity is through the analysis of norms.

Definition 2.1.1. *Given an n dimensional vector $\mathbf{x}$ in Euclidean space, $\mathbb{R}^n$, the p-norm is defined for $p > 0$ as*

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

When $p = 2$, it is called the ℓ_2 norm, or Euclidean norm, and simplifies to

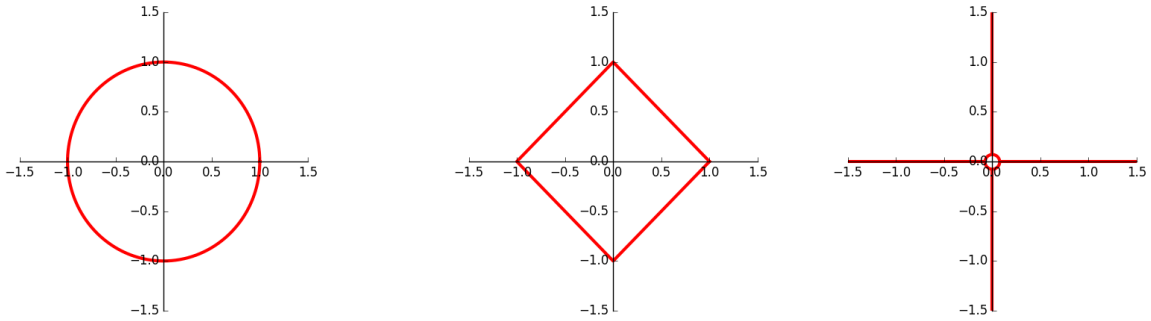
$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

The 2-norm is a convex function. A depiction of the ℓ_2 unit ball can be seen in Figure 2.1 (a).

When $p = 1$ the norm is called the ℓ_1 norm, taxicab norm, or Manhattan norm as it is representative of how a taxi measures distance in city blocks. The ℓ_1 norm simplifies to

$$\|\mathbf{x}\|_1 = \sum_{x=1}^{n} |x_i|$$

and is a convex function. A depiction of the ℓ_1 unit ball can be seen in Figure 2.1 (b).



(a) 2-norm unit level set (b) 1-norm unit level set (c) 0-norm unit level set

Figure 2.1 Level curves for a variety of norm functions.

The limit as p approaches 0 relates to the so called 0-norm which is not actually a norm, but measures the number of nonzero elements in a vector, or the size of its support. The 0-norm is denoted

$$\|\mathbf{x}\|_0 = \lim_{p \to 0} \|\mathbf{x}\|_p^p = \lim_{p \to 0} \sum |x_i|^p,$$

and a depiction of its level sets can be seen Figure 2.1 (c). The 0-norm is a non-convex function. The sparsity of a vector or matrix is related to the complement of the size of its support. Sparsity is

defined as the number of zero-valued elements divided by the total number of elements, or in the case of our n dimensional vector, $\frac{n-\|\mathbf{x}\|_0}{n}$. These mathematical functions can be used to characterize the fundamental properties that are pervasive in so many fields.

2.2 Optimization and Relaxations

Optimization problems are used to search a space for the best element with regard to some criterion. Within the field of optimization there are two principle categories, convex and non-convex optimizations. Convex optimization problems, which entail optimizing a convex function over a convex set, are in many senses much easier than non-convex optimization problems. The tools for non-convex problems include heuristic local methods that fail to guarantee global optimality such as gradient descent, and exact methods such as branch-and-bound that take exponential effort. Convex optimization problems, on the other hand, have the distinct benefit that local solutions are also globally optimal. This allows for global optimality with relatively quick gradient based methods.

Due to the relative ease of solving convex problems, non-convex problems are often approximated with a convex surrogate [16, 64, 98]. In applications where some characteristic should be used parsimoniously, it is natural to integrate this into the optimization problems. For instance, when sparsity is a desirable characteristic, it makes sense to insert this into the objective function. Though the 0-norm perfectly characterizes sparsity, the fundamental quality we often wish to promote, its non-convexity poses significant challenges in optimization. Consider the example minimization problem

$$\begin{aligned} \underset{\mathbf{x}}{\arg\min} \quad & \|\mathbf{x}\|_0 \\ \text{subject to} \quad & \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2 \end{aligned} \tag{2.1}$$

with two optimal solutions of $x = [0,1]^T$ and $x = [2,0]^T$. To find these solutions we must take a combinatorial approach, restricting the support of the vector $\mathbf{x}$ and solving the system. In such a small example this is not problematic, but as the problem size grows this quickly becomes computationally infeasible.

The convex relaxation of (2.1) to

$$\begin{aligned} \underset{\mathbf{x}}{\arg\min} \quad & \|\mathbf{x}\|_1 \\ \text{subject to} \quad & \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2 \end{aligned} \tag{2.2}$$

has a single optimal solution at $x = [0,1]^T$. The relaxed problem (2.2), is a linear programming problem that can be solved efficiently using a variety of algorithms. In this case, the solution of the

relaxation is one of the solutions to the original non-convex problem, while the other solution is lost in the relaxation. This type of relaxation technique is a powerful tool with both heuristic support [127, 140, 141] and theoretical merit [48, 49, 51, 143].

Another parsimonious characteristic in a significant number of applications is rank. The rank of the matrix, A, can be calculated from its singular values and the 0-norm.

Definition 2.2.1. *Given an n by m matrix A with singular values* $\sigma = \left[\sigma_1, \ldots, \sigma_{\min(m,n)}\right]$ *the rank can be defined as*

$$\text{rank}(A) = \|\sigma\|_0$$

In rank minimization problems, one searches the feasible space for the lowest rank solution which also requires a combinatorial approach. To approximate a solution one can replace rank with nuclear norm.

Definition 2.2.2. *Given an n by m matrix A with singular values* $\sigma = \left[\sigma_1, \ldots, \sigma_{\min(m,n)}\right]$ *the nuclear norm of A, denoted* $\|A\|_*$*, can be defined as*

$$\|A\|_* = \text{trace}(\sqrt{A^* A}) = \|\sigma\|_1$$

This provides a computationally feasible objective function first introduced in [54] that can be used to approximate minimal rank solutions in many instances [98].

The relaxations of non-convex problems into convex problems provides a drastic benefit to computational efficiency at the price of quality of solution. In some instances the solution to the relaxed problems will also be optimal in the original problem. Characterizing these cases is often a subject of study. In other instances the solution to the relaxed problem will not be a solution to the non-convex problem. In these cases it is worthwhile to study and understand the relation between the solutions to understand the effect of the relaxation.

CHAPTER

3

GRAMIAN TENSOR DECOMPOSITION

3.1 Introduction

In this chapter we examine a tensor decomposition problem, posed as a rank minimization problem. We study the relaxation of the problem and consider cases when the relaxed solution is a solution to the original problem. In some instances of tensor rank and order, we prove generically that the solution to the relaxation will be optimal in the original. In other cases we present interesting examples and approaches that demonstrate the complexities of this problem.

3.1.1 Background

Let $\mathscr{A} \in \mathbb{F}^{(n+1)\times\cdots\times(n+1)}$ be a D-way, or order D, symmetric tensor over a field $\mathbb{F}$ of size $(n+1)\times\cdots\times(n+1)$ (D-times). Let $R := \mathbb{F}[x_1,\ldots,x_n]$ and let R_D denote the set of polynomials of degree at most D in R. Then we can associate to $\mathscr{A}$ a polynomial

$$p = \sum_{\beta\in\mathbb{N}^n, |\beta|\leq D} \binom{D}{D-|\beta|,\beta_1,\ldots,\beta_n} p_\beta x^\beta \in R_D \tag{3.1}$$

by simply multiplying $\mathscr{A}$ by the vector $[1, x_1,\ldots,x_n]$ from all the D directions. This gives a bijection between symmetric D-way tensors over $\mathbb{F}$ and polynomials in R_D.

We define the *symmetric rank* of the tensor $\mathscr{A}$, and the *rank* of the polynomial p as follows:

Definition 3.1.1. *We say that $\mathscr{A} \in \mathbb{F}^{(n+1)\times\cdots\times(n+1)}$ has symmetric rank r if there exist distinct $\mathbf{v}_1 = (v_{1,0}, v_{1,1}, \ldots, v_{1,n}), \ldots, \mathbf{v}_r = (v_{r,0}, v_{r,1}, \ldots, v_{r,n}) \in \overline{\mathbb{F}}^{n+1}$ with coordinates from the algebraic closure $\overline{\mathbb{F}}$ of $\mathbb{F}$, and $\lambda_1, \ldots, \lambda_r \in \overline{\mathbb{F}}/\{0\}$ such that r is minimum and*

$$\mathscr{A} = \sum_{t=1}^{r} \lambda_t \mathbf{v_t}^{\otimes D} := \sum_{t=1}^{r} \lambda_t \left[v_{t,i_1} \cdots v_{t,i_D} \right]_{i_1,\ldots,i_D=0}^{n}. \tag{3.2}$$

Equivalently, we say that $p \in R_D$ has rank r if r is minimal and

$$p = \sum_{t=1}^{r} \lambda_t L_{\mathbf{v}_t}^{D}, \tag{3.3}$$

where $L_{\mathbf{v}_t}(x_1, \ldots, x_n) := v_{t,0} + v_{t,1}x_1 + \cdots + v_{t,n}x_n$ is the linear form associated to $\mathbf{v}_t = (v_{t,0}, v_{t,1}, \ldots, v_{t,n})$ for $t = 1, \ldots, r$. The expressions in (3.2) or (3.3) are called the rank r symmetric decompositions of $\mathscr{A}$ and p, respectively.

There are different, non-equivalent notions of tensor rank in the literature, such as the multilinear rank or non-symmetric rank, etc. (see [32]). Also, one can define the symmetric rank over non-algebraically closed fields, which unlike for matrices, may differ from the above defined symmetric rank for tensors of order > 2. If the field $\mathbb{F}$ is the set of real numbers and the order $D = 2d$ is even, we can define the *Gramian rank* as follows:

Definition 3.1.2. *Let $\mathscr{A} \in \mathbb{R}^{(n+1)\times\cdots\times(n+1)}$ be a real symmetric tensor of order $2d$ and $p \in R_{2d}$ be the corresponding real polynomial. We say that $\mathscr{A}$ and p is Gramian with Gramian rank r if there exist distinct $\mathbf{v}_1 = (v_{1,0}, v_{1,1}, \ldots, v_{1,n}), \ldots, \mathbf{v}_r = (v_{r,0}, v_{r,1}, \ldots, v_{r,n}) \in \mathbb{R}^{n+1}$ and $\lambda_1, \ldots, \lambda_r \in \mathbb{R}_{>0}$ positive real numbers such that r is minimal and (3.2) or (3.3) holds. The decompositions in (3.2) and (3.3) are called the Gramian decompositions of $\mathscr{A}$ and p, respectively.*

In this chapter we consider the problem of finding the Gramian rank and decomposition for a real symmetric tensor of order $2d$, or equivalently, for a polynomial of degree $2d$. Note that not all polynomials of degree $2d$ are Gramian, in particular, Gramian polynomials are a subset of sum of square polynomials. Hillar and Lim in [69] proved that deciding whether a tensor/polynomial is Gramian is NP-hard even for $d = 2$. Also note that even if a tensor is Gramian, its Gramian rank may be much higher than its symmetric rank.

We give an algorithm that finds the Gramian decomposition in the case when the Gramian rank is sufficiently small. Our approach is to use a relaxation of this problem to semidefinite programming and to show that for sufficiently small Gramian rank r the optimum of the relaxed problem gives a Gramian decomposition of length r. This work is a first step to attack the more general problem of finding the symmetric rank and decomposition via semidefinite relaxation. The general case is subject to future research.

The main results of this Chapter are as follows:

- We give a meaningful semidefinite relaxation of the problem of finding the Gramian rank and decomposition of a polynomial $p \in R_{2d}$, assuming that its Gramian rank is sufficiently small. The relaxation becomes a matrix completion problem of moment matrices with minimal trace.
- We simplify and interpret the condition that a given moment matrix is the optimum of our relaxed semidefinite program, using special properties of the dual of the semidefinite program.
- We analyze special cases when we can guarantee that a given moment matrix is the optimum of the relaxed semidefinite program. In these special cases we point to a connection to the theory of the regularity index of overdetermined polynomial systems. Using this theory we list triples (n, d, r) where we can prove that the optimum of the semidefinite relaxation corresponds to the Gramian decomposition of rank r of a polynomial of degree $2d$ in n variables.

3.1.2 Related Work

Motivation for looking at the tensor decomposition problem comes from its broad application areas. The earliest results on tensor decomposition were applications in mathematical physics ([70, 71]); psychometrics ([24, 25, 67, 145, 146]); algebraic complexity theory ([72, 81, 83, 85, 86, 138]); and in chemometrics ([6, 61, 135]). In higher order statistics, moments and cumulants are intrinsically tensors (cf. [109]). Symmetric tensor decomposition is proven to be useful in *blind source separation* techniques, which are capable of identifying a linear statistical model only from its outputs (cf. [33]). These blind identification techniques in turn are very popular in numerous applications, including telecommunication ([2, 66, 130]); radar ([26]); biomedical engineering ([41]); image and signal processing ([42, 63]) just to name a few. An excellent survey of more recent applications of tensor methods can be found in [69].

Despite the rich literature on the numerical aspects of the symmetric tensor decomposition problem, there are relatively small numbers of publications concerned with the symbolic computational aspects of computing the rank of symmetric and non-symmetric tensors. Even though the first algorithm solving the problem in the bivariate symmetric case goes back to [139], and several other symbolic algorithms exist in the literature for finding the rank of symmetric tensors (see for example [18, 31, 78, 88–90, 115, 122]) and non-symmetric tensors (see for example [2, 10, 11, 43, 44, 87, 115, 137]), they all have strong constrains on the degree d, dimension n and/or on the rank r. A list of all cases where we know the defining equations for (border)-rank r symmetric tensors can be found in [89]. Symmetric rank computation is NP-hard [69], and its approximation doesn't always

exist as the set of rank r tensors is not closed [132]

As we will see in the preliminaries below there is a close relationship between the so called truncated moment problem and the Gramian decomposition of tensors. Here we only mention work that is closest to our problem, namely when representing measures that are finitely atomic. The foundations of the theory and algorithms to study this truncated moment problem were laid down in a sequence of work by Curto and Fialkow in [35, 36], including the so called stopping criteria that we use in this chapter. In a series of papers [91–94] the moment problem is connected to polynomial optimization and the solution of polynomial systems over the reals, and our approach is based on this work. The direct relationship between symmetric tensor decomposition and the truncated moment problem was described in the works [11, 18]; our approach strongly relies on these results. As we mentioned earlier, in [69] they prove that detecting if a symmetric tensor is Gramian is NP-hard, and they also discussed the relationship between Gramian, non-negative definite tensors, and completely positive matrices. Reznick in [123] proved that the cone of tensors and of Gramian tensors are dual. It is also proved here that the set of Gramian rank r tensors is closed. In [97] they deduce a computationally feasible condition for uniqueness using the notion of coherence. In [96] they study nonnegative approximations of nonnegative tensors, where they use a generalization of the notion of completely positive matrices, which is different from Gramian and nonnegative-definite tensors.

Relaxations of matrix rank minimization problems using the nuclear norm of matrices was first introduced in [54, 55]. There is a rich literature on results about the accuracy of the relaxation of a low rank optimization problem using the nuclear norm. The low rank matrix completion approach assumes that a linear image of the underlying low rank matrix M is known and attempts to recover the full matrix M. The motivation and justification for this relaxation is that the nuclear norm of matrices is the convex envelope of the rank function (cf. [121]). The main results in [19, 20, 121] give general assumptions which guarantee both the rank minimization problem and its relaxation to have M as its unique solution (with high probability). One of these assumptions in [19] is the existence of a bound on the so called *coherence* of the column and row spaces of the output M. Another such assumption is given for the input. In [121] they show that if a certain *restricted isometry property* holds for the linear transformation defining the constraints, the minimum-rank solution can be recovered by the nuclear norm relaxation. Similar ideas were explored in [59, 65, 101, 112, 131] to recover low *multilinear rank* tensors. Here the objective function is the sum of the ranks of the flattenings of the tensor which is subject to linear constrains. This is relaxed by using the sum of the nuclear norms of the flattenings instead. Recently, [163] gave a general SDP approximation scheme with penalty parameters to find the minimal rank solution. They proved that with some mild restrictions on the constraints the optimum of the relaxations converge to the minimal rank

solution as the penalty parameters approach 0. As we mentioned above, our approach is closest to the one in [128].

3.2 Preliminaries

Before describing our results, let us give a brief summary of the main results in the theory of flat extensions of moment matrices (see [18] for more details). Assume that we have a Gramian decomposition as in (3.2) or (3.3) for some $\mathbf{v}_1 = (v_{1,0}, v_{1,1}, \dots, v_{1,n}), \dots, \mathbf{v}_r = (v_{r,0}, v_{r,1}, \dots, v_{r,n}) \in \mathbb{R}^{n+1}$ and $\lambda_1, \dots, \lambda_r \in \mathbb{R}_{>0}$. We assume that

$$v_{1,0} = 1, \dots, v_{r,0} = 1$$

and denote by

$$\mathbf{z}_i = (v_{i,1}, \dots, v_{i,n}) \in \mathbb{R}^n \text{ for } i = 1, \dots, r.$$

Consider the infinite matrix M and its truncation $M_{i,j}$ for some $i, j \in \mathbb{N}$ defined by

$$M := \left[m_{\beta+\beta'}\right]_{\beta,\beta' \in \mathbb{N}^n} \text{ and } M_{i,j} := \left[m_{\beta+\beta'}\right]_{|\beta| \le i, |\beta'| \le j}, \tag{3.4}$$

where for $\alpha \in \mathbb{N}^n$

$$m_\alpha = \sum_{t=1}^{r} \lambda_t \mathbf{z}_i^\alpha,$$

denotes the *moments* corresponding to the points $\{\mathbf{z}_1, \dots, \mathbf{z}_r\}$. If $i = j$ we will denote $M_{i,i}$ by simply M_i. These matrices have so called quasi-Hankel structure (see [111]), and called *moment matrices*, i.e. they are matrices whose rows and columns are indexed by monomials and the entries depend only on the product of the indexing monomials.

Let $V := [\mathbf{z}_i^\beta]_{i=1,..r,\beta \in \mathbb{N}^n}$ be the Vandermonde matrix with infinitely many columns, its truncation $V_i := [\mathbf{z}_i^\beta]_{i=1,..r,\beta \in \mathbb{N}^n, |\beta| \le i}$, and let $\Lambda := \operatorname{diag}(\lambda_1, \dots, \lambda_r)$. Then we have

$$M = V^T \Lambda V \quad M_{i,j} = V_i^T \Lambda V_j. \tag{3.5}$$

When we only know the tensor $\mathscr{A}$ or the polynomial p as in (3.1) for $D = 2d$, but not the decomposition, from (3.3) it is easy to see that for $|\beta + \beta'| \le 2d$ we have

$$[M]_{\beta,\beta'} = m_{\beta+\beta'} = \binom{2d}{\beta+\beta'}^{-1} p_{\beta+\beta'}. \tag{3.6}$$

The truncations $M_{i,j}$ for $i + j \le 2d$, are called *catalecticant matrices*, and its theory goes back to

Sylvester in [139]. Note that for $i = j = d$, M_d is a symmetric matrix of size $\binom{n+d}{d} = \dim R_d$.

Next we define the notion of flat extensions of moment matrices:

Definition 3.2.1. *Given M_D a moment matrix for some degree $D \geq 0$ as in (3.4). We call an infinite moment matrix M an extension of M_D if*

$$[M]_{\beta,\beta'} = [M_D]_{\beta,\beta'} \text{ for } |\beta + \beta'| \leq D.$$

If, in addition,

$$\operatorname{rank}(M) = \operatorname{rank}(M_D),$$

then we say that M is a flat extension of M_D. Furthermore, if M is positive semidefinite, we call M a Gramian flat extension of M_D.

Clearly, if $p \in R_{2d}$ has symmetric rank r, then there exists at least one infinite moment matrix M of rank r that extends M_d. Similarly, if p has Gramian rank r then there exists some positive semidefinite moment matrix M of rank r that extends M_d. If, in addition, M_d also has rank r, then M is a Gramian flat extension of M_d. Note that if the decomposition of p is not unique, then the flat extensions of M_d may not be unique either. The converse is not entirely true: if M_d has an infinite flat extension M of rank r , then p has a so called *generalized decomposition*, where the points $\{\mathbf{z}_t\}_{t=1}^r$ may be repeated (see [12] for more details). However, for a positive semidefinite flat extension the corresponding points in the decomposition are always distinct. Thus, these positive semidefinite flat extensions always correspond to a Gramian decomposition of the tensor [35].

In [11, 18, 37] they give conditions for the existence of a (Gramian) flat extension in terms of finite truncations of M:

Theorem 3.2.2 (STOPPING CRITERION FOR FLAT EXTENSION)**.** *Let M_d be a moment matrix as above. Let M be an infinite extension of M_d as above. M has rank r if and only if there exist $D \geq 0$ such that*

$$\operatorname{rank}(M_D) = \operatorname{rank}(M_{D+1}) = r.$$

If, in addition, M_{D+1} is positive semidefinite, then M is also positive semidefinite. We call M_{D+1} a truncated (Gramian) flat extension of M_D.

Note that once the above stopping criterion is satisfied, one can compute a system of multiplication matrices from the kernel of M_{D+1}, and the coordinates of the points $\mathbf{z}_i$ for $i = 1, \ldots, r$ can be read out from the eigenvalues of these multiplication matrices.

In the present chapter we assume that $p \in R_{2d}$ has Gramian rank r satisfying

$$\operatorname{size}(M_{d-1}) = \binom{n+d-1}{n} < r \leq \binom{n+d}{n} = \operatorname{size}(M_d)$$

and M_d has a truncated Gramian flat extension M_{d+1} of rank

$$\text{rank}(M_d) = \text{rank}(M_{d+1}) = r,$$

i.e. $D = d$ in the stopping criterion above.

However, given $p \in R_{2d}$, we only know the entries of M_d, so we want to find a truncated Gramian flat extension M_{d+1}. Note that if $r \leq \text{size}(M_{d-1})$ then by the stopping criterion we do not need to extend the matrix M_d to find the Gramian rank. So the *truncated Gramian flat extension problem* that we attempt to solve in this chapter is the following:

Definition 3.2.3 (TRUNCATED GRAMIAN FLAT EXTENSION PROBLEM)**.** *Given $p \in R_{2d}$ as in (3.1) with non-zero constant term. Assume that the corresponding truncated moment matrix M_d has rank r and is positive semidefinite. Find a positive semidefinite moment matrix extension M_{d+1} of M_d which has rank r, if one exists. Equivalently, find a minimal rank positive semidefinite extension M_{d+1} of M_d.*

Unfortunately, the minimal rank optimization problem is NP-hard, and all known algorithms which provide exact solutions are double exponential in the dimension of the matrix (cf. [19]). However, relaxation techniques were successfully applied for "low rank matrix completion" or "affine rank minimization" problems that are very similar in structure to our problem. Namely, the constraints on the extension matrix M_{d+1} are all linear equalities. These relaxation techniques replace rank minimization by the minimization of the nuclear norm of the matrix. Recall that the nuclear norm of a matrix M is defined by

$$\|M\|_* := \sum_{i=1}^{r} \sigma_i,$$

where $\sigma_1 > \sigma_2 > \cdots > \sigma_r > 0$ are the non-zero singular values of M. The advantage is that the nuclear norm is a convex function and can be optimized efficiently using semidefinite programming. Note that when M is positive semidefinite then

$$\|M\|_* = \text{trace}(M).$$

Definition 3.2.4 (RELAXATION OF TRUNCATED GRAMIAN FLAT EXTENSION)**.** *Given $p \in R_{2d}$ with non-zero constant term, find a positive semidefinite moment matrix M_{d+1} satisfying $[M_{d+1}]_{\beta,\beta'} = \binom{2d}{\beta+\beta'}^{-1} p_{\beta+\beta'}$ for $|\beta+\beta'| \leq 2d$, and* trace(M_{d+1}) *is minimal.*

The purpose of this chapter is to prove that for sufficiently low Gramian rank r, the optimum of the relaxation is the minimal rank solution.

In [19, 20, 121] the goal of the low rank matrix completion and affine rank minimization problems is to give conditions on the matrix and on the linear constraints so that the optimum of the minimal rank problem is unique and equal to the optimum of the nuclear norm relaxation. In our case uniqueness cannot always be expected, since symmetric tensors can have many minimal decompositions, resulting in different flat extensions of the same rank. For example, if r is the generic rank as in [4], [110] conjectures that the solution is never unique, except for three cases. The lack of uniqueness is a significant obstacle for the relaxation to find the minimal rank solution as the set of minimal rank decompositions may be a non-convex object. For this reason we cannot expect to find the minimal rank decomposition via semidefinite optimization. To address this obstacle we constrain ourselves to cases where the minimal decomposition of the symmetric tensor is essentially unique (up to unimodulus scaling).

For symmetric decompositions rather strong uniqueness results were proved in [30, 78, 110]. Namely, for a decomposition as in (3.3), if $d \geq 2$, and

$$r \leq \binom{d+n}{d} - n + 1 = \dim R_d - n + 1 \tag{3.7}$$

then the decomposition is essentially unique, as long as the points $\{\mathbf{z}_i\}_{i=1}^r$ are in general position (cf. [78, Th.2.6]). Our ultimate goal would be to prove that in the cases of unique decomposition, the semidefinite relaxation gives the minimal rank solution. At this point we could only prove a small portion of these cases, however, in the process we uncovered some interesting connections of this problem to the theory of the regularity index of polynomial systems, which is an active research area in mathematics.

A difference between our problem and the ones considered in [19, 20, 121] is that the linear constraints on the extension M_{d+1} are not given at random, and we cannot expect that the corresponding linear map would satisfy either the restricted isometry conditions of [121] or the injectivity when restricted to the tangent space of rank r matrices at the optimum as in [19]. Thus to tackle our problem we needed new ideas. Our approach is closest to the work in [128], where they study conditions when the semidefinite relaxation solves the minimal rank matrix diagonal completion problem. They reinterpret the dual of the semidefinite relaxation problem in several different ways and connect their original problem to other well-studied problems in statistics and geometry. We follow a similar approach, but leading to very different results.

3.3 Relaxation and Dual Problem

Given $d \in \mathbb{N}$, $n \geq 1$, and $p \in R_{2d}$ as in (3.1), and let

$$M_d = \left[m_{\beta+\beta'} \right]_{\substack{\beta,\beta' \in \mathbb{N}^n \\ |\beta+\beta'| \leq 2d}}$$

be the corresponding truncated moment matrix as in (3.6) with moments $m_\alpha = \binom{2d}{\alpha}^{-1} p_\alpha$ for $|\alpha| \leq 2d$. Denote by

$$N := \binom{n+d+1}{n} = \dim R_{d+1},$$

and by $\mathscr{S}_N$ the space of real symmetric matrices of size N. The truncated Gramian flat extension problem in Definition 3.2.3 is finding a symmetric matrix $X \in \mathscr{S}_N$, with columns and rows indexed by $\alpha, \beta \in \mathbb{N}^n$, such that

$$\begin{array}{rl} \min & \operatorname{rank}(X) \\ \text{Subject To} & \begin{cases} [X]_{\beta,\beta'} = m_{\beta+\beta'} & \text{for } |\beta+\beta'| \leq 2d \\ [X]_{\beta,\beta'} - [X]_{\gamma,\gamma'} = 0 & \text{if } \beta+\beta' = \gamma+\gamma' \\ X \succeq 0 \end{cases} \end{array}$$

Using the bilinear form

$$< A, B >:= Tr(A \cdot B),$$

we choose an orthonormal basis for the space of symmetric matrices $\mathscr{S}_N$ as specified in Definition 3.3.1.

Definition 3.3.1 (CHOICE OF ORTHOGONAL BASIS FOR $\mathscr{S}_N$)**.** *For each $\alpha \in \mathbb{N}^n$ such that $|\alpha| \leq 2d+2$, we define the subspace $\mathscr{S}_\alpha \subset \mathscr{S}_N$ of symmetric matrices with support indexed by the set of pairs $\{(\gamma,\delta) : \gamma+\delta = \alpha\}$. Fix $Y_\alpha \in \mathscr{S}_\alpha$ to be the moment matrix which has 1 at each entry in its support. Then choose an arbitrary orthonormal basis $\{Z_{\alpha,i} : 1 \leq i \leq \dim \mathscr{S}_\alpha - 1\} \subset \mathscr{S}_\alpha$ for the subspace of $\mathscr{S}_\alpha$ orthogonal to Y_α.*

Example 3.3.1. *For example, in the univariate case with a monomial basis of $1, x, x^2$ we define an orthogonal decomposition of $\mathscr{S}_3$:*

$$Y_0 = \begin{bmatrix} 1 & & \\ & & \\ & & \end{bmatrix}, Y_1 = \begin{bmatrix} & 1 & \\ 1 & & \\ & & \end{bmatrix}, Y_2 = \begin{bmatrix} & & 1 \\ & 1 & \\ 1 & & \end{bmatrix}, Y_3 = \begin{bmatrix} & & \\ & & 1 \\ & 1 & \end{bmatrix}, Y_4 = \begin{bmatrix} & & \\ & & \\ & & 1 \end{bmatrix}$$

One can easily see that this set is a basis for all 3 by 3 Hankel matrices. We also see that with our

monomial list there are two ways to obtain $x^2 = x^2 \cdot 1 = x \cdot x$ *so we then define one matrix orthogonal to* Y_2 *with respect to our inner product and with the same support* $Z_2 = \begin{bmatrix} & & -1 \\ & 2 & \\ -1 & & \end{bmatrix}$.

Using this notation we rewrite the truncated Gramian flat extension problem as follows:

$$\begin{array}{ll} \min & \operatorname{rank}(X) \\ \text{Subject To} & \begin{cases} < Y_\alpha, X >= m_\alpha & |\alpha| \leq 2d \\ < Z_{\alpha,i}, X >= 0 & |\alpha| \leq 2d+2, \quad 1 \leq i \leq \dim(\mathscr{S}_\alpha)-1 \\ X \succeq 0 \end{cases} \end{array},$$

This we relax to a semidefinite program:

$$\begin{array}{ll} \min & < I, X > \\ \text{Subject To} & \begin{cases} < Y_\alpha, X >= m_\alpha & |\alpha| \leq 2d \\ < Z_{\alpha,i}, X >= 0 & |\alpha| \leq 2d+2, \quad 1 \leq i \leq \dim(\mathscr{S}_\alpha)-1 \\ X \succeq 0 \end{cases} \end{array},$$

Thus we get the following primal and dual semidefinite optimization problems (in standard form):

$Primal$	$Dual$
$\min_X < I, X >$	$\max_{(\mathbf{y},\mathbf{z},S)} \sum m_\alpha y_\alpha$
$\text{Subject To} \begin{cases} < Y_\alpha, X >= m_\alpha \\ < Z_{\alpha,i}, X >= 0 \\ X \succeq 0 \end{cases}$	$\text{Subject To} \begin{cases} S = I - \sum y_\alpha Y_\alpha - \sum z_{\alpha,i} Z_{\alpha,i} \\ S \succeq 0 \end{cases}$

,

where the indices of y_α and Y_α run through $|\alpha| \leq 2d$, while the indices of $z_{\alpha,i}$ and $Z_{\alpha,i}$ run through $|\alpha| \leq 2d+2$ and $1 \leq i \leq \dim(\mathscr{S}_\alpha)-1$, using the notation of Definition 3.3.1.

In the rest of this chapter we will use the following notation for the above semidefinite programs:

$$
\begin{aligned}
(\mathscr{P}) &: \text{primal problem in standard form;}\\
(\mathscr{D}) &: \text{dual problem in standard form;}\\
\mathscr{P} &: \text{feasible set of problem } (\mathscr{P});\\
\mathscr{D} &: \text{feasible set of problem } (\mathscr{D});\\
\mathscr{P}^* &: \text{optimal set of problem } (\mathscr{P});\\
\mathscr{D}^* &: \text{optimal set of problem } (\mathscr{D}).
\end{aligned}
$$

3.4 Certificate of Optimality

Assume that we are given a Gramian decomposition of $p \in R_{2d}$

$$p = \sum_{i=1}^{r} \lambda_i (1 + v_{i,1} x_1 + \cdots + v_{i,n} x_n)^{2d},$$

corresponding to the points $\mathbf{z}_i = (v_{i,1}, \ldots, v_{i,n}) \in \mathbb{R}^n$ and $\lambda_i > 0$ for $i = 1, \ldots, r$. Using the Vandermonde matrix V_{d+1} of the points $\{\mathbf{z}_1, \ldots, \mathbf{z}_r\}$ and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ as in (3.5), it is clear that $M_{d+1} = V_{d+1}^T \Lambda V_{d+1}$ is in the feasible set, $\mathscr{P}$. Our goal is to give conditions that guarantee that M_{d+1} is in the set of optimal solutions, $\mathscr{P}^*$. To get such conditions we use both $(\mathscr{P})$ and $(\mathscr{D})$ defined above.

One can see that $(\mathscr{D})$ is strictly feasible with $S = I$ and its optimum is bounded above by trace(M_{d+1}) since $M_{d+1} = V_{d+1}^T \Lambda V_{d+1}$ as a feasible solution for $(\mathscr{P})$. This implies that there is no duality gap between the optimal values of $(\mathscr{P})$ and $(\mathscr{D})$, although $(\mathscr{D})$ might not attain its optimum [147]. However, if we can construct a feasible pair $X \in \mathscr{P}$ and $(y, z, S) \in \mathscr{D}$ such that

$$< X, S >= 0$$

then we must have $X \in \mathscr{P}^*$ and $(y, z, S) \in \mathscr{D}^*$ since

$$0 =< X, S >=< I, X > -\mathbf{m}^T \mathbf{y},$$

which implies optimum by weak duality. Note that for positive semidefinite matrices X and S we have

$$< X, S >= 0 \iff XS = 0.$$

Thus we get the following theorem:

Theorem 3.4.1. *The moment matrix $M_{d+1} = V_{d+1}^T \Lambda V_{d+1}$ is optimal for $(\mathscr{P})$, or $M_{d+1} \in \mathscr{P}^*$, if there exists $S \in \mathscr{S}_N$ such that:*

$$\begin{aligned} & M_{d+1} S = 0, \\ & S = I - \sum_{|\alpha| \le 2d} y_\alpha Y_\alpha - \sum_{\substack{|\alpha| \le 2d+2, \\ 1 \le i \le \dim(\mathscr{S}_\alpha)-1}} z_{\alpha,i} Z_{\alpha,i}, \\ & S \succeq 0 \end{aligned}$$

Using Theorem 3.4.1 we study when the optimal solution of $(\mathscr{P})$ is unique and $\mathscr{P}^* = \{M_{d+1}\}$. We are only concerned with cases where the rank r symmetric decomposition of the associated polynomial p is unique, and it is Gramian. In this case Proposition 3.4.2 gives sufficient conditions to show $\mathscr{P}^* = \{M_{d+1}\}$:

Proposition 3.4.2. *Assume that p has Gramian rank r and the rank r symmetric decomposition of p is unique. If $\exists S$ satisfying Theorem 3.4.1 of rank $N-r$, then $\mathscr{P}^* = \{M_{d+1}\}$.*

Proof. Suppose p has a unique Gramian rank r decomposition and let S be a matrix satisfying Theorem 3.4.1 of rank $N-r$. Let $M \in \mathscr{P}^*$. Since $MS = 0$ and $\operatorname{rank}(S) = N-r$, we have $\operatorname{rank}(M) \le r$. But by the stopping criteria in Theorem 3.2.2, M defines a rank $\le r$ symmetric decomposition for p, so the uniqueness of the symmetric decomposition implies that $M = M_{d+1}$. □

Additionally we note the following about the set of matrices satisfying Theorem 3.4.1.

Proposition 3.4.3. *If $\exists S$ satisfying Theorem 3.4.1, then $\exists \bar{S}$ satisfying Theorem 3.4.1 with* $\operatorname{rank}(\bar{S}) \le \binom{n+d}{d+1}$.

Proof. Suppose $\exists S$ satisfying Theorem 3.4.1. By zeroing the Schur compliment of the submatrix indexed by degree $d+1$ monomials, we can produce $\bar{S}$ with $\operatorname{rank}(\bar{S}) \le \binom{n+d}{d+1}$. □

To better aid our analysis of the problem we reformulate Theorem 3.4.1 into a sum of squares decomposition problem. Corollary 3.4.4 gives an alternative formulation of Theorem 3.4.1 by noticing that the polynomial $\mathbf{x}^T S \mathbf{x}$ do not depend on the $z_{\alpha,i}$ variables and interpreting the problem as a sum of squares decomposition.

Corollary 3.4.4. *The moment matrix $M_{d+1} = V_{d+1}^T \Lambda V_{d+1} \in \mathscr{P}^*$ corresponding to the points $\mathbf{z}_1, \ldots, \mathbf{z}_r$*

is optimal if there exists $q \in R_{2d+2}$ and $q_\alpha \in R_{d+1}$ for $|\alpha| = d+1$ such that:

$$\begin{aligned} q &= \sum_{|\alpha|=d+1} q_\alpha^2 \\ q_\alpha(\mathbf{z}_i) &= 0, \text{ for all } 1 \le i \le r, |\alpha| = d+1 \\ \text{coeff}(q, x^\beta) &= \delta_{2|\beta} \text{ for } |\beta| = 2d+1, 2d+2, \end{aligned}$$

where

$$\delta_{2|\beta} = \begin{cases} 1 & \text{if } \exists \gamma \in \mathbb{N}^n \text{ such that } 2\gamma = \beta \\ 0 & \text{otherwise.} \end{cases}$$

Proof. We prove the equivalence of the criteria of Theorem 3.4.1 and Corollary 3.4.4. First we prove the conditions of Theorem 3.4.1 implies the condition of Corollary 3.4.4. Assume there exists S such that $M_{d+1}S = 0$, $S = I - \sum y_\alpha Y_\alpha - \sum z_{\alpha,i} Z_{\alpha,i}$, and $S \succeq 0$ as in Theorem 3.4.1. Without loss of generality from Proposition 3.4.3 we assume $\text{rank}(S) \le \binom{n+d}{d+1}$ with Cholesky factorization $S = LL^T$. With $\mathbf{x} = [x^\beta]_{|\beta| \le d+1}$, we let $q = \mathbf{x}^T S \mathbf{x}$ and let the collection q_α consist of the polynomials $L^T \mathbf{x}$. Then $q = \mathbf{x}^T S \mathbf{x} = \mathbf{x}^T LL^T \mathbf{x} = \sum_\alpha q_\alpha^2$, and each q_α vanishes on $\mathbf{z_i}$ since $M_{d+1}S = 0 \implies V_{d+1}^T \Lambda V_{d+1} LL^T = 0 \implies V_{d+1} L = 0$. Using the observations that

$$< \mathbf{x}\mathbf{x}^T, I >= \mathbf{x}^T \mathbf{x}, \quad < \mathbf{x}\mathbf{x}^T, Y_\alpha >= x^\alpha, \quad < \mathbf{x}\mathbf{x}^T, Z_{\alpha,i} >= 0,$$

we conclude that for $|\beta| = 2d+1, 2d+2$, we have,

$$\begin{aligned} \text{coeff}(q, x^\beta) &= \text{coeff}(\mathbf{x}^T (I - \sum_{|\alpha| \le 2d} y_\alpha Y_\alpha + \sum_{\substack{|\alpha| \le 2d+2, \\ 1 \le i \le \dim(\mathscr{S}_\alpha)-1}} z_{\alpha,i} Z_{\alpha,i}) \mathbf{x}, x^\beta) \\ &= \text{coeff}(\mathbf{x}^T I \mathbf{x}, x^\beta) \\ &= \delta_{2|\beta}. \end{aligned}$$

Now we prove the conditions of Corollary 3.4.4 implies the conditions of Theorem 3.4.1. Assume there exists q and q_α as in Corollary 3.4.4. Then we form a coefficient matrix, L, from the coefficient vectors of q_α and let $S = LL^T$ so $S \succeq 0$. Also $q_\alpha(\mathbf{z}_i) = 0$ for $1 \le i \le r \implies V_{d+1} L = 0 \implies M_{d+1} S = 0$. To conclude, it is sufficient to show that the two sets

$$\left\{ S \in \mathscr{S}_N : < Y_\beta, S >= \delta_{2|\beta} \text{ for } |\beta| = 2d+1, 2d+2 \right\}$$

and

$$\left\{ S \in \mathscr{S}_N \,:\, S = I - \sum_{|\alpha| \le 2d} y_\alpha Y_\alpha - \sum_{\substack{|\alpha| \le 2d+2, \\ 1 \le i \le \dim(\mathscr{S}_\alpha)-1}} z_{\alpha,i} Z_{\alpha,i},\ y_\alpha, z_{\alpha,i} \in \mathbb{R} \right\}$$

are equal. Above we proved the "$\supseteq$" direction. Since both of these sets are affine spaces, it is enough to prove that the vector spaces

$$\{S \in \mathscr{S}_N \,:< Y_\beta, S >= 0 \text{ for } |\beta| = 2d+1, 2d+2\}$$

and

$$\left\{ S \in \mathscr{S}_N \,:\, S = \sum_{|\alpha| \le 2d} y_\alpha Y_\alpha + \sum_{\substack{|\alpha| \le 2d+2, \\ 1 \le i \le \dim(\mathscr{S}_\alpha)-1}} z_{\alpha,i} Z_{\alpha,i},\ y_\alpha, z_{\alpha,i} \in \mathbb{R} \right\}$$

have the same dimension. By construction, we have that $\{Y_\alpha, Y_\beta, Z_{\gamma,i} \,:\, |\alpha| \le 2d, |\beta| = 2d+1, 2d+2, |\gamma| \le 2d+2, 1 \le i \le \dim(\mathscr{S}_\gamma)-1\}$ is a basis for $\mathscr{S}_N$, which proves the claim. □

3.5 Sufficient Conditions for Optimality

In this section we demonstrate that in some special cases M_{d+1} will generically be optimal in $(\mathscr{P})$ by imposing an assumption on the polynomials q_α in Corollary 3.4.4.

Corollary 3.5.1. *The moment matrix $M_{d+1} = V_{d+1}^T \Lambda V_{d+1} \in \mathscr{P}^*$ corresponding to the points $\mathbf{z}_1, \ldots, \mathbf{z}_r$ is optimal if there exists $q \in R_{2d+2}$ and $q_\alpha = x^\alpha + l.d.t. \in R_{d+1}$ for $|\alpha| = d+1$ such that:*

$$q = \sum_{|\alpha| = d+1} q_\alpha^2$$
$$q_\alpha(\mathbf{z}_i) = 0, \text{ for all } 1 \le i \le r, |\alpha| = d+1$$
$$\text{coeff}(q, x^\beta) = 0 \text{ for } |\beta| = 2d+1.$$

Proof. Suppose there exists q and $q_\alpha = x^\alpha + l.d.t.$ satisfying Corollary 3.5.1, then $\text{coeff}(q, x^\beta) = \delta_{2|\beta}$ for $|\beta| = 2d+2$ because degree $2d+2$ terms only depend on the squares of the degree $d+1$ terms in q_α. □

The assumption on q_α simplifies the criteria sufficient to prove optimality of M_{d+1} into the solvability of a linear system. We note here that $\text{Ker}(V_{d+1}) = \text{Ker}(M_{d+1})$, so we can use the two interchangeably.

Proposition 3.5.2. *Let V_d, V_{d+1} be the Vandermonde matrices of r points. Let K_d be a matrix with columns that form a basis for* $\text{Ker}(V_d)$*, and let F be the matrix of coefficients of the normal forms*

of monomials of degree $d+1$ modulo the vanishing ideal of our r points. If $\operatorname{rank}(V_d) = r$ *then the columns of the matrix* $K_{d+1} = \begin{bmatrix} K_d & -F \\ 0 & I \end{bmatrix}$ *form a basis for* $\operatorname{Ker}(V_{d+1})$.

Using K_{d+1} we can look at a matrix existence formulation of Corollary 3.5.1.

Corollary 3.5.3. *The moment matrix $M_{d+1} = V_{d+1}^T \Lambda V_{d+1} \in \mathscr{P}^*$ if there exists $G \in \mathscr{S}_{N-r}$ such that:*

$$G = \begin{bmatrix} gg^T & g \\ g^T & I \end{bmatrix}$$
$$\operatorname{coeff}(\mathbf{x}^T K_{d+1} G K_{d+1}^T \mathbf{x}, x^\beta) = 0 \text{ for } |\beta| = 2d+1.$$

where g is a real matrix of size $\binom{n+d}{n} - r$ by $\binom{n+d}{d+1}$ and I is the identity matrix of size $\binom{n+d}{d+1}$.

Proof. G is clearly positive semidefinite with the decomposition $G = \begin{bmatrix} g \\ I \end{bmatrix} \begin{bmatrix} g^T & I \end{bmatrix}$. Using G we let $q = \mathbf{x}^T K_{d+1} G K_{d+1}^T \mathbf{x}$ and associate each q_α with the corresponding element of the vector $\mathbf{x}^T K_{d+1} \begin{bmatrix} g \\ I \end{bmatrix}$. Then $q = \sum_\alpha q_\alpha^2$ by construction. Since K_{d+1} is in the null space of V_{d+1} we also conclude that $q_\alpha(\mathbf{z}_i) = 0$, for all $1 \le i \le r, |\alpha| = d+1$. Lastly, $\operatorname{coeff}(\mathbf{x}^T K_{d+1} G K_{d+1}^T \mathbf{x}, x^\beta) = 0$ for $|\beta| = 2d+1 \implies \operatorname{coeff}(q, x^\beta) = 0$ for $|\beta| = 2d+1$. □

Proposition 3.5.4. *The values of g satisfying Corollary 3.5.3 are the solution of an inhomogeneous linear system of equations.*

Proof. Let $q = \mathbf{x}^T \begin{bmatrix} K_d g - F \\ I \end{bmatrix} \begin{bmatrix} g^T K_d^T - F^T & I \end{bmatrix} \mathbf{x}$ and consider the degree $d+1$ polynomials in the row vector, $\mathbf{x}^T \begin{bmatrix} K_d g - F \\ I \end{bmatrix}$. The degree $d+1$ components of these polynomials consist of a single monomial that is independent of $g_{i,j}$. The degree d coefficients of these polynomials are inhomogeneous but linear in $g_{i,j}$. Because degree $2d+1$ coefficients of q rely only on the product of degree d and degree $d+1$ coefficients of the polynomials the values of g satisfy an inhomogeneous linear system. □

In order for the inhomogeneous linear system of Proposition 3.5.4 to have a solution, it is sufficient that the corresponding homogeneous equations are linearly independent. Thus we should try to understand what is the coefficient matrix of this linear system and determine when it is full row rank. For this we first define a subresultant matrix.

Definition 3.5.5. *The degree Δ subresultant matrix of t homogeneous polynomials $h_1,\dots,h_t$ of degree $d \le \Delta$ in n variables is the matrix whose columns are the coefficient vectors of the multiples of each h_i with all monomials of degree $\Delta - d$. For example, if $\Delta - d = d+1$ and the monomials of degree $d+1$ are $\{x^{\alpha_i}\}_{i=1}^{s}$ as above, then*

$$\mathrm{Sres}_{\Delta}(h_1,\dots,h_t) := \left|\begin{array}{c|c|c||c||c|c|c} x^{\alpha_1}h_1 & \dots & x^{\alpha_s}h_1 & \cdots & x^{\alpha_1}h_t & \dots & x^{\alpha_s}h_t \end{array}\right|.$$

Subresultant matrices play an important role in studying the homogeneous parts of the ideal $\langle h_1,\dots,h_t\rangle$.

Theorem 3.5.6. *Let G be a matrix satisfying Corollary 3.5.3 and denote the entries of K_d by $k_{i,\beta}$ for $i = 1,\dots,t$ and $|\beta| \le d$. We define the homogeneous degree d polynomials:*

$$h_i := \sum_{|\beta|=d} k_{i,\beta}\, x^{\beta} \quad i = 1,\dots,t.$$

Then the coefficient matrix of the linear system in Proposition (3.5.4) in the variables $\{g_{i,j}\}$ is $\mathrm{Sres}_{2d+1}(h_1,\dots,h_t)$.

Proof. First note that the normal form coefficients only appear in the constant terms, so do not appear in the coefficient matrix. The rows of the coefficient matrix correspond to monomials x^{β} of degree $|\beta| = 2d+1$. For each $j \in \{1,\dots,\binom{n+d}{d+1}\}$ associate with it a unique monomial of degree $d+1$, α_j. For fixed $i \in \{1,\dots,t\}$ and $j \in \{1,\dots,\binom{n+d}{d+1}\}$, the column corresponding to the variable $g_{i,j}$ has zero entry in the row corresponding to x^{β} unless x^{α_j} divides x^{β}. If $x^{\alpha_j} | x^{\beta}$ then the entry is $k_{i,\beta-\alpha_j}$, which shows that the column of $g_{i,j}$ is the coefficient vector of $x^{\alpha_j} h_i$. □

Corollary 3.5.7. *Let $M_{d+1} = V_{d+1} \Lambda V_{d+1}^T$, with V_{d+1} the Vandermonde matrix, and V_d^T has full column rank. Define the homogeneous degree d polynomials $h_1,\dots,h_t$ from $\mathrm{Ker}(V_d^T)$ as in Theorem 3.5.6. Then the matrix $M_{d+1} \in \mathscr{P}^*$ if $Sres_{2d+1}(h_1,\dots,h_t)$ has full row rank.*

In the rest of this subsection we study when the rows of the subresultant matrix are independent. Note that the rows are independent if and only if

$$\langle h_1,\dots,h_t\rangle_{2d+1} = R_{=2d+1}, \tag{3.8}$$

where the left hand side denotes the homogeneous part of degree $2d+1$ of the ideal generated by $h_1,\dots,h_t$, and the right hand side denotes the space of homogeneous polynomials of degree $2d+1$. Thus (3.8) is satisfied only if $2d+1$ is greater or equal than the *regularity index* of $\langle h_1,\dots,h_t\rangle$, i.e. the

smallest degree where the Hilbert function of the ideal agrees with its Hilbert polynomial. Note that if $h_1,\dots,h_t$ has common roots in the projective space over $\mathbb{C}$ then (3.8) can never be satisfied, which implies that we need to have $t \geq n$.

For the rest of the section we assume that $h_1,\dots,h_t$ is a system such that the dimension of $\langle h_1,\dots,h_t\rangle_{2d+1}$ is the maximum possible. In the results below we give specific constructions of particular real systems $h_1^*,\dots,h_t^*$ and study when we have $\langle h_1^*,\dots,h_t^*\rangle_{2d+1} = R_{=2d+1}$. Therefore, if we assume that our $\langle h_1,\dots,h_t\rangle_{2d+1}$ is maximal, then it will also imply that $\langle h_1^*,\dots,h_t^*\rangle_{2d+1} = R_{=2d+1}$.

Remark 3.5.8. *We cannot prove that the assumption on $h_1,\dots,h_t$ will ever be satisfied in our case. In fact, our polynomials $h_1,\dots,h_t$ are not generic, they are real, and they are the highest degree parts of degree d polynomials vanishing on some generic real points. In [117] it was shown that systems $h_1,\dots,h_t$ for which $\langle h_1,\dots,h_t\rangle_{2d+1}$ is not maximal are defined by non-trivial polynomial equations, so over $\mathbb{C}$ they form a Zariski closed subset. Furthermore, even for the "generic" case over $\mathbb{C}$ the behavior of $\langle h_1,\dots,h_t\rangle_{2d+1}$ is not well understood. In [57] they give a conjecture about the Hilbert series of generic systems over $\mathbb{C}$.*

The regularity index of $n \times n$ homogeneous systems were widely studied in the literature, but for highly overdetermined systems that has Hilbert series as in Fröberg's conjecture in [57] only the asymptotic behavior of the regularity index is known as $n \to \infty$ (c.f. [7, 8]).

The next theorem gives all values of d and n when (3.8) is satisfied in the cases when $t = n$ and $t = n+1$. The analysis of the cases when $t > n+1$ is still ongoing. Since $r = \binom{n+d}{n} - t$, we can easily translate these results in terms of the Gramian rank r. Finally, we want to note that on the other end of the spectrum, when $t = \binom{n+d}{n-1} = \dim R_{=d}$ and $h_1,\dots,h_t$ are generic, then the coefficient vectors of $h_1,\dots,h_t$ form a square full rank matrix, thus (3.8) is satisfied for all n and d. However in this case $r = \binom{n+d-1}{n} = \dim R_{d-1}$, and the matrices M_{d-1} and M_d already satisfy the stopping criterion for flat extension, so we do not need an extension to M_{d+1}.

Proposition 3.5.9. *Let $h_1,\dots,h_t$ be homogeneous polynomials of degree d in n variables, and assume that $\langle h_1,\dots,h_t\rangle_{2d+1}$ is maximal. Then $\langle h_1,\dots,h_t\rangle_{2d+1} = R_{=2d+1}$ if*

1. *in the case of $t = n$*

$$n = 2 \text{ for arbitrary } d,$$
$$n = 3 \text{ and } d \leq 3,$$
$$n = 4 \text{ and } d \leq 2,$$
$$n \geq 5 \text{ and } d = 1.$$

2. *in the case of* $t = n+1$

$$\begin{aligned}
&n = 2 \text{ or } 3 \text{ for arbitrary } d,\\
&n = 4 \text{ and } d \le 6,\\
&n = 5 \text{ and } d \le 3,\\
&n = 6,7,8 \text{ and } d \le 2,\\
&n \ge 9 \text{ and } d = 1.
\end{aligned}$$

Proof. First note that if we find a particular system $h_1^*,\dots,h_t^*$ of degree d that satisfy $\langle h_1^*,\dots,h_t^*\rangle_{2d+1} = R_{=2d+1}$, then any generic $h_1,\dots,h_t$ will also satisfy it. For $t = n$ the standard theory of subresultants uses the system

$$h_1^* := x_1^d,\dots,h_n^* := x_n^d.$$

Then one can define

$$\delta := n(d-1),$$

and it is easy to see that if $\Delta \ge \delta+1$ then the matrix $\mathrm{Sres}_\Delta(h_1^*,\dots,h_n^*)$ has more columns than rows and contains the identity matrix, so it has full row rank. Thus we need that $2d+1 \ge \delta+1$ and that is only satisfied in the cases listed in the claim.

For $t = n+1$ we will use the system

$$h_1^* := x_1^d,\dots,h_n^* := x_n^d,\ h_{n+1}^* := (x_1+\ldots+x_n)^d.$$

Let

$$H_d(\nu) := \mathrm{span}\{x^\gamma \,:\, |\gamma| = \nu,\ \forall i\ \gamma_i < d\}$$

and denote by $\mathscr{H}_d(\nu) := \dim H_d(\nu)$. Clearly, the monomials not in $H_d(\nu)$ generate $\langle x_1^d,\dots,x_n^d\rangle_\nu$. Define the linear map

$$\begin{aligned}
\psi_{h_{n+1}^*} : H_d(d+1) &\to H_d(2d+1)\\
x^\alpha &\mapsto x^\alpha \cdot h_{n+1}^* \mod \langle x_1^d,\dots,x_n^d\rangle_{2d+1}.
\end{aligned}$$

By [153, Corollary 3.5 and Theorem 3.8.(0)], the matrix of the map $\psi_{h_{n+1}^*}$ has full rank. So if

$$\mathscr{H}_d(2d+1) \le \mathscr{H}_d(d+1) \tag{3.9}$$

then $\psi_{h_{n+1}^*}$ is surjective, and $\mathrm{Sres}_\Delta(h_1^*,\dots,h_n^*)$ has full row rank. Using the fact that $\mathscr{H}_d(\nu) = \mathscr{H}_d(\delta-\nu)$ and that $\mathscr{H}_d(\nu)$ is monotonically decreasing in $[\lceil\frac{\delta}{2}\rceil,\delta]$, we get that (3.9) is satisfied when either $\lceil\frac{\delta}{2}\rceil \le d+1 \le 2d+1$ or $\lceil\frac{\delta}{2}\rceil \le \delta-(d+1) \le 2d+1$. This is always satisfied if $n \le 3$ and for $n \ge 4$ it reduces

to $d \leq \frac{n+2}{n-3}$, resulting in the values in the claim. □

A different approach was presented in [58, Theorem 6], where they studied the minimal number t such that a generic homogeneous form in n variables of degree kd is a sum of the k-th powers of t forms of degree d over $\mathbb{C}$. For the case of $k=2$ they prove that for

$$t = 2^{n-1}$$

and generic $h_1, \ldots, h_t \in R_{=d}$ we have

$$\langle h_1, \ldots, h_t \rangle_{2d} = R_{=2d},$$

which is slightly stronger than what we need in (3.8). Moreover, their construction for $k=2$ works over the reals, in particular, they show that the following 2^{n-1} real polynomials

$$h_I^* := \left(x_1 + \sum_{i \in I} x_i - \sum_{j \notin I} x_j \right)^d \quad \text{for all } I \subseteq \{2, \ldots, n\}$$

will generate $R_{=2d}$ in degree $2d$. Moreover, they show that there is an open subset of all real polynomials of degree $2d$ where the "typical rank" is 2^{n-1}, but there might be other "typical ranks" too (see also [34] on typical ranks over $\mathbb{R}$). They also show that for large enough d the $t = 2^{n-1}$ upper bound is sharp, but for small d this bound is not always sharp.

3.6 Cases When M_{d+1} is Never Optimal

In the previous section we explored cases of triplets (n, d, r) where we can generically prove that M_{d+1} is optimal for $(\mathscr{P})$ and list these cases. In this section we describe cases where we expect not to be able to find any S satisfying Theorem 3.4.1 by counting the degrees of freedom and number of linear constraints in Corollary 3.4.4.

The linear constraints can be indexed by the monomials of degree $2d+1$. So the number of linear constraints is $\binom{n+2d}{2d+1}$. In these linear constraints there are $\left(\binom{n+d}{d} - r\right)\binom{n+d}{d+1}$, $g_{i,j}$ variables and $\frac{\left(\binom{n+d}{d+1}\right)\left(\binom{n+d}{d+1}\right)}{2} - \binom{n+2d+1}{2d+2}$, $z_{\alpha,i}$ variables. The linear system is overdetermined if

$$\binom{n+2d}{2d+1} > \left(\binom{n+d}{d} - r\right)\binom{n+d}{d+1} + \frac{\left(\binom{n+d}{d+1}\right)\left(\binom{n+d}{d+1}\right)}{2} - \binom{n+2d+1}{2d+2}$$

or identically,

$$r > \frac{\frac{\left(\binom{n+d}{d+1}\right)\left(\binom{n+d}{d+1}\right)}{2} - \binom{n+2d+1}{2d+2} - \binom{n+2d}{2d+1}}{\binom{n+d}{d+1}} + \binom{n+d}{d}$$

.

Asymptotically, these bounds are not applicable due to the limitation that r is less than the size of M_d, but there are instances where this bound is applicable.

One instance is when $n = 2$ and $d = 3$, the bound indicates that M_{d+1} will generally not be optimal when $r = 10$ as the linear system is overdetermined. This triplet of (n, d, r) is a case where the corresponding decompositions are unique, and rank$(M_d) = r$, but M_{d+1} will generically not be optimal for $\mathscr{P}$.

3.7 Uncertain Cases

Outside of the cases listed in Sections 3.5 and 3.6, the possibility of M_{d+1} being optimal in P may depend on more than just the triplet (n, d, r). Instances may depend fundamentally on the sets of points $\{\mathbf{z}_i\}$. To demonstrate this we present two examples in the same triplet (n, d, r) where one example has M_{d+1} optimal, and one does not.

Let us consider the case when $n = 2, d = 3$. In this case, size$(M_d) = 10$, and size$(M_{d+1}) = 15$. A discussion of the extreme rays in this case can be found in [13]. Gramian rank 10 decompositions will generally not be optimal solutions in $(\mathscr{P})$ as the linear system in Corollary 3.4.4 is overcomplete. Gramian rank 8 decompositions will generically be optimal in $(\mathscr{P})$ from Corollary 3.5.7 and Proposition 3.5.9. Between these two ranks we wish to understand what happens. Here we present two examples of Gramian rank 9 decompositions, one where $\exists S$ satisfying Theorem 3.4.1, and one where $\nexists S$ satisfying Theorem 3.4.1.

Example 3.7.1. *Let* $\{\mathbf{z}_i\} = \{(78,87),(-45,78),(-38,32),(91,-76),(-18,94),(-22,-22),(27,99),(52,-16),(-58,-87)\}$ *be the set of* $r = 9$ *points, and let* $\lambda_i = 1$ *for* $i = \{1,\dots,9\}$. *In this case* $\exists S$ *satisfying Theorem 3.4.1 and* M_{d+1} *is optimal in* $(\mathscr{P})$.

Example 3.7.2. *Let* $\{\mathbf{z}_i\} = \{(-43,-34),(-18,-10),(-19,23),(52,72),(-66,-76),(48,-15),(35,45),(-83,-72),(51,22)\}$ *be the set of* $r = 9$ *points, and let* $\lambda_i = 1$ *for* $i = \{1,\dots,9\}$. *In this case* $\nexists S$ *satisfying Theorem 3.4.1 and* M_{d+1} *is not optimal in* $(\mathscr{P})$. *In this instance, the optimal solution is rank 11.*

These examples demonstrate the complexity of the cases where Proposition 3.5.9 does not hold, as the solution to the relaxed problem may or may not be optimal in the original problem. In these cases the triplet (n, d, r) is not sufficient to determine if M_{d+1} is optimal in $(\mathscr{P})$ and specific information of the points is necessary.

3.7.1 Future Work

Some of the methods used to search for certificates of optimality also suggest future research avenues. Given an instance of a specific (n, d, r) and $\{\mathbf{z}_i\}$ the standard approach to search for a

certificate of optimality using Theorem 3.4.1 involves two steps. First, we solve the under-determined linear system coeff($\mathbf{x}^T S\mathbf{x}$, x^β) $= 0$ for $|\beta| = 2d+1$. With the resulting affine solution, we look for an intersection with the positive semidefinite cone.

Let us consider for a moment the set $T = \{A \mid A \succeq 0,\ A = I - \sum_{\substack{|\alpha|=2d+2,\\ 1\le i\le \dim(\mathscr{S}_\alpha)-1}} z_{\alpha,i}\tilde{Z}_{\alpha,i}\}$. We know that $I \in T$ with $z_{\alpha,i} = 0$ for $|\alpha| = 2d+2$ and $1 \le i \le \dim(\mathscr{S}_\alpha)-1$, so the set is nonempty and has an interior. Another interesting observation is that this set is bounded.

Theorem 3.7.1. *Fix n and d and an orthonormal basis $\{\tilde{Y}_\alpha\} \cup \{\tilde{Z}_{\alpha,i}\}$ for $|\alpha| = 2d+2$ and $1 \le i \le \dim(\mathscr{S}_\alpha)-1$. Then the set*

$$I - \sum_{\substack{|\alpha|=2d+2,\\ 1\le i\le \dim(\mathscr{S}_\alpha)-1}} z_{\alpha,i}\tilde{Z}_{\alpha,i} \succeq 0$$

is bounded.

Proof. Choose $Y = \sum_{|\alpha|=2d+2} y_\alpha \tilde{Y}_\alpha \succ 0$ to be a full rank positive definite matrix. A generic collection of $\binom{n+d+1}{d+1}$ points will produce such a matrix. Suppose that there exists a matrix $Z = -\sum_{\substack{|\alpha|=2d+2,\\ 1\le i\le \dim(\mathscr{S}_\alpha)-1}} z_{\alpha,i}\tilde{Z}_{\alpha,i}$ such that $I + sZ \succeq 0$ for $s > 0$, then $Z \succeq 0$. But $< Y, Z >= 0$ by construction, therefore $Z = 0$ and the set is bounded. □

Using these observations of T, we can make some conclusions from the solution of our linear system. For instance, if the solution to the linear system can be solved independent of the $z_{\alpha,i}$, then Corollary 3.5.3 applies and M_{d+1} is optimal. Alternatively if the $z_{\alpha,i}$ variables are necessary, but yield a solution such that $\sum z_{\alpha,i}^2 < 1$, then M_{d+1} will be optimal since $I - \sum_{\substack{|\alpha|=2d+2,\\ 1\le i\le \dim(\mathscr{S}_\alpha)-1}} z_{\alpha,i}\tilde{Z}_{\alpha,i} \succ 0$. Lastly, if the diameter of the set in Theorem 3.7.1 is diam(n,d), then if the solution of the linear system closest to the origin has $\sum z_{\alpha,i}^2 >$ diam$(n,d)^2$ then Theorem 3.4.1 cannot apply.

Studying the sets of $\{\mathbf{z}_i\}$ that provide instances of each of these cases will be a topic for future exploration. Future research may also extend the idea further with a linear programming relaxation. The additional constraint $\sum |z_{\alpha,i}| < 1$ is linear and such a solution also guarantees the optimality of M_{d+1} in $(\mathscr{P})$.

Additionally, we are interested in the rank of the optimal solutions in the cases when M_{d+1} is not optimal in the minimal nuclear norm problem. One approach to address this question may be to examine the extremal rays in the feasible set of $\mathscr{P}$. This may provide a meaningful upper bound for the rank of the optimal solution, as $\mathscr{P}^*$ must contain an extremal ray.

3.8 Conclusion

In this Chapter we study the Gramian decomposition of tensors and polynomials by posing a rank optimization problem. Through relaxation of the optimization problem, we pose a convex

optimization problem to approximate the minimal rank solution. Our analysis of the relaxed problem reveals a relation between the Gramian decomposition problem and the theory of subresultants. Our research further provides specific cases where the optimal solution to our relaxation is also minimum rank. Lastly we provide some interesting cases demonstrating the complexity of the problem and discuss future work.

CHAPTER

4

GEOMETRIC ANALYSIS OF SPARSITY PARAMETER SELECTION IN DICTIONARY LEARNING

4.1 Introduction

Image restoration is one of the fundamental problems in image processing. Numerous perspectives on this problem have led to a variety of proposed approaches, including Bayesian filtering [124], total variation minimization [125], level sets [149], and block matching filtering [38], each with rich associated literature. Sparse representation and dictionary learning, which is of central interest here, is one such approach [105]. In this approach, one seeks to sparsely represent a broad class of signals with an overcomplete dictionary. The properties of dictionaries and their use in sparse representation are the subject of this chapter.

4.1.1 Sparse Representation and Dictionary Learning Background

One interpretation of the sparse coding problem is a change of basis problem, where the signal representation in the new basis is sparse. Given a signal s, a dictionary D, and tolerance ϵ, solve:

$$\begin{aligned} \underset{x}{\arg\min} \quad & \|x\|_0 \\ \text{subject to} \quad & \|Dx - s\|_2 \leq \epsilon, \end{aligned} \tag{4.1}$$

where $\|x\|_0$ is the number of nonzero elements of a vector x. This is a combinatorial problem in the subset selection of the columns of the dictionary, or the atoms, to use in the sparse representation of the signal. The sparse coding problem fits the data to a sparse linear subspace model where each signal lies in a relatively low dimensional subspace of ambient space.

Wavelets have been heavily invoked as a good dictionary D in image processing for their sparse representation properties [107]. Later improvements have led to sparse coding by optimizing over a class of dictionaries. Given a matrix of signals S, and tolerance ϵ, to find a dictionary of n atoms solve:

$$\begin{aligned} \underset{D,X}{\arg\min} \quad & \|X\|_0 \\ \text{subject to} \quad & \begin{cases} \|DX - S\|_2 \leq \epsilon, \\ \|D_i\|_2^2 \leq 1, \text{ for } i = 1, \ldots, n, \end{cases} \end{aligned} \tag{4.2}$$

where D_i is the i^{th} column of D. This allows the dictionary to adapt to signals of interest, thus allowing for sparser representations.

Computationally, problem 4.2 is an NP-hard minimization problem [142]. The problem is non-convex for two reasons, the product, DX, as well as the ℓ_0 norm. There are two common approaches to approximate solutions to the problem, both of which use an iterative alternating update procedure to handle the non-convexity of DX. To handle the ℓ_0 minimization one approach uses a greedy algorithm, the other uses a relaxation.

A greedy algorithm to approximate (4.2) pairs Orthogonal Matching Pursuit (OMP) [119] with K-SVD [1]. OMP is an algorithm that iteratively calculates a sparse representation. Each iteration incorporates an additional dictionary atom into the representation to minimize the residual. The K-SVD algorithm uses the support from the sparse representations to update individual atoms with a rank one singular value approximation. Algorithm 1 summarizes the steps of using OMP with K-SVD to learn a dictionary.

Data: matrix of signals: $S \in \mathbb{R}^{m \times N}$
sparsity parameter: $\epsilon > 0$
Result: dictionary: $D \in \mathbb{R}^{m \times n}$
sparse representation: $X \in \mathbb{R}^{n \times N}$

Initialize D to an arbitrary matrix
while *not converged* **do**
  Approximate

$$X = \arg\min_X \quad \|X\|_0 \qquad \text{subject to} \quad \|DX - S\|_2 \leq \epsilon,$$

  using OMP
  for $i = 1$ *to* n **do**
    Define the set of signals that use D_i, $\omega_i = \{j | X_{i,j} \neq 0\}$
    Compute the representation error matrix, $E_i = S - \sum_{j \neq i} D_j X_j$
    Restrict E_i to ω_i and obtain E_i^R
    Compute the rank one truncated SVD decomposition, $u s v^T = E_i^R$
    Update $D_i = u$ and $(X_i^T)^R = s v$
  end
end

Algorithm 1: Dictionary learning algorithm using OMP with K-SVD

A different approach relaxes $\|x\|_0$ to $\|x\|_1$ and solves the relaxed problem (4.3). In this relaxation one alternates solving the minimization problem for X and D until convergence is achieved. The minimization over X is a convex problem called Basis Pursuit DeNoising (BPDN) in signal processing [29] and Least Absolute Shrinkage and Selection Operator (lasso) [141] in statistics. This problem can be efficiently solved with the Alternating Direction Method of Multipliers (ADMM) algorithm [17]. The minimization over D is also convex and can be efficiently solved with the Method of Optimal Direction (MOD) algorithm [53] which entails solving an over-complete linear system and a projection. Algorithm 2 summarizes the steps of using lasso with MOD to perform dictionary learning.

$$\begin{aligned} &\arg\min_{D,X} && \|DX - S\|_F^2 + \lambda \|X\|_1 \\ &\text{subject to} && \|D_i\|_2^2 = 1, \text{ for } i \in \{1, \ldots, n\}. \end{aligned} \tag{4.3}$$

Data: matrix of signals: $S \in \mathbb{R}^{m \times N}$
sparsity parameter: $\lambda > 0$

Result: dictionary: $D \in \mathbb{R}^{m \times m}$
sparse representation: $X \in \mathbb{R}^{m \times N}$

Initialize D to an arbitrary matrix
while *not converged* **do**
Solve

$$X = \arg\min_{X} \quad ||DX - S||_F^2 + \lambda ||X||_1$$

using ADMM
Solve

$$D = \arg\min_{D} \quad ||DX - S||_F^2$$
$$\text{subject to} \quad ||D_i||_2^2 = 1, \text{ for } i = 1, \ldots, n.$$

using MOD
end

Algorithm 2: Dictionary learning algorithm using lasso with MOD

The two dictionary learning algorithms have different sparsity parameters, ϵ and λ, but with a known relation. In practice, the ϵ sparsity parameter is best seen as a tolerance parameter, indicating the required fidelity in the optimization. The λ sparsity parameter is a regularization coefficient, this parameter will also correlate with the fidelity term in the optimization, but does not directly guarantee any tolerance bounds. Both sparsity parameters share the trend that, the larger the parameter, the sparser the optimal solution.

4.1.2 Image Restoration

Image restoration is fundamentally an inverse problem that attempts to recover an unknown signal s_0, from a corrupted observation s [50]. There are numerous types of corruption in images, each with unique sources, properties, and corresponding solutions. We focus here on two common types of noise, additive white noise, and salt and pepper impulse noise.

Classicly additive noise models the observed signal as a sum of an unknown uncorrupted signal and noise or, $s = s_0 + \eta$. For white noise, η is zero mean homogeneous Gaussian noise. This noise model implies an inter-pixel independence of the corruption and an independence of the signal intensity. Common sources of white noise are thermal variations and signal transmission. Additive white noise is one of the most well studied with a range of potential solutions, some notable

approaches to white noise denoising are linear filtering [80] and block matching [38].

Salt and pepper noise is impulsive noise where a random subset of pixels are corrupted to assume either a minimum or maximum value. Similarly to additive white noise, salt and pepper corruption is spatially independent as well as independent of signal intensity. Common sources of salt and pepper noise include artifacts in analog to digital conversions as well as bit errors [14]. Approaches well suited to handle salt and pepper noise include interpolation methods [144], median filtering [76], and total variation denoising [156].

Dictionary learning is one technique that can be used to solve image restoration problems. In recent years different adaptations of dictionary learning have incorporated dictionary learning into the target application for online task-driven dictionary learning [73, 106]. For our experiments we break the process of image restoration into several steps. The first step is to train a dictionary on an uncorrupted set of image patches. With the resulting dictionary, the second step is to solve a form of the sparse coding problem on corrupted image patches. The final step is to restore the image from the sparse representation. In both denoising applications, the ideal sparsity parameter generally correlates with the sparsity level. It is rationalized that the noisier the signal, the more regularization is required to return the corrupted signal to the image patch manifold.

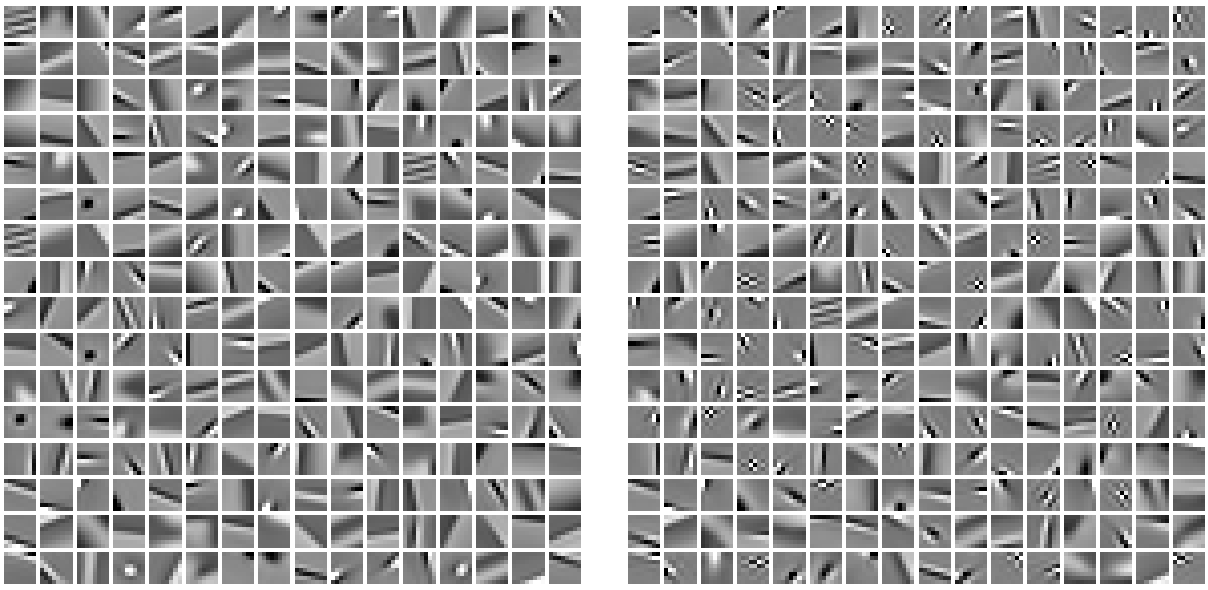
The main results of this Chapter are as follows:

- We observe the effect of sparsity parameters in dictionary learning and interpret their geometric use in a manifold model and their efficacy in image reconstruction.
- We analyze statistical properties of dictionary atoms and interpret the results geometrically.
- We propose a modification to dictionary learning thatexploits the statistical properties to make dictionaries that are structured and more can be tailored to application after training.

4.2 Sparsity Parameter: Effect and Analysis

We may intuitively expect that when denoising with a dictionary the best regularization would correlate with both the noise level as well as the regularization used during training. To test this premise, we train two dictionaries on a set of natural image patches using Algorithm 2 with two different regularizations. Given the resulting dictionaries, we proceed to solve image denoising problems, with a variety of regularizations.

Figure 4.1 depicts two dictionaries initialized randomly and trained on the same noiseless image patches. It is visually evident that the two different dictionaries contain different information. The dictionary trained with $\lambda = 0.1$ displays appealing atoms with stronger natural image features such as lines. The dictionary trained with $\lambda = 0.01$ has on the other hand, numerous small oscillatory components that are not visually dominant in natural images.



(a) $\lambda = 0.1$ (b) $\lambda = 0.01$

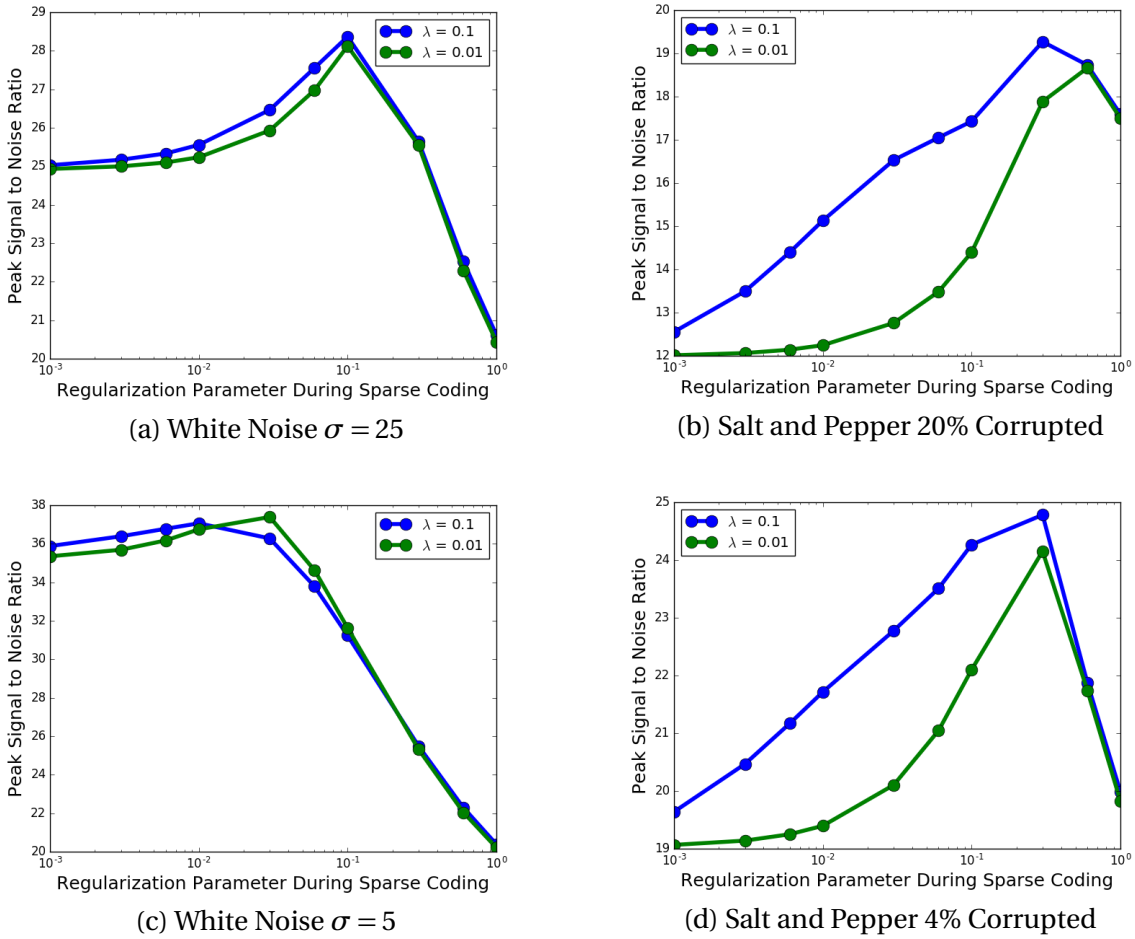
Figure 4.1 Dictionaries trained on natural image patches data using lasso with MOD.

Figure 4.2 displays the denoising performance comparisons of the two dictionaries for additive white noise and salt and pepper noise. For the denoising experiments a testing image was corrupted, broken into blocks and denoised with the lasso problem, and restored as an image which was used to calculate the peak signal to noise ratio.

In analyzing the white noise denoising case one may observe that the dictionaries are perhaps best used with a similar regularization with which they were trained. The peak performance when $\sigma = 25$ occurs when the dictionary trained with $\lambda = 0.1$ is also used with $\lambda = 0.1$ regularization. In the $\sigma = 5$ case, the peak performance occurs when the dictionary trained with $\lambda = 0.01$ is used with $\lambda = 0.01$ regularization.

The salt and pepper denoising results are a little different. The dictionary trained with $\lambda = 0.01$ regularization level is consistently outperformed when denoising salt and pepper. This result is not surprising, as a significant number of atoms in the dictionary have a relatively small support and may be delta function-like. The dictionary trained with $\lambda = 0.1$ has fewer of these atoms, and generally appear to have larger supports.

These experiments point to a certain relative dependence of the performance of dictionaries on both the type of noise and the noise level. Neither dictionary is consistently better at denoising, indicating that the two dictionaries contain different information that may or may not be desirable depending on the noise. This experiment demonstrates that what information is desirable depends on both the properties and magnitude of the noise. Our goal is to provide a geometric explanation



(a) White Noise $\sigma = 25$

(b) Salt and Pepper 20% Corrupted

(c) White Noise $\sigma = 5$

(d) Salt and Pepper 4% Corrupted

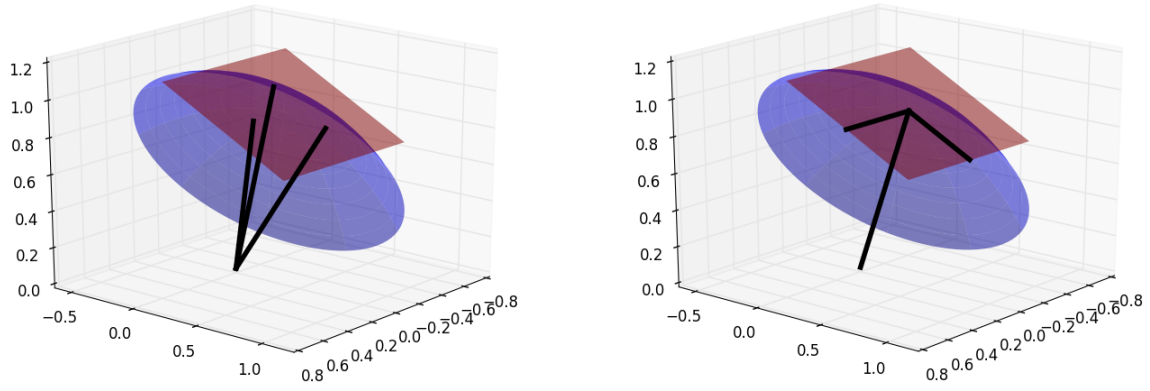
Figure 4.2 Denoising performance of two dictionaries at a variety of noise levels and regularizations.

of the different types of information, as well as propose a method that improves control over the information.

4.2.1 Geometric Interpretation

The assumptions of sparse representation theory typically are that each signal lies in a low dimensional subspace of the ambient space. The sparse coding problem picks the basis that minimizes the dimensionality of the subspace. We consider an alternative perspective, one where each signal lies on a low dimensional manifold. Previous research on the natural image manifold suggests that 3 by 3 image patches belong on a topological Klein bottle [23]. Other groups have considered the statistics of natural image patches without the imposition of a sparse representation model [95]. We develop a manifold interpretation of sparse representation and dictionary learning by describing

geometrically the relationship between the dictionary atoms and the manifold and show statistically the different usage patterns of different information.



(a) Principal Atoms

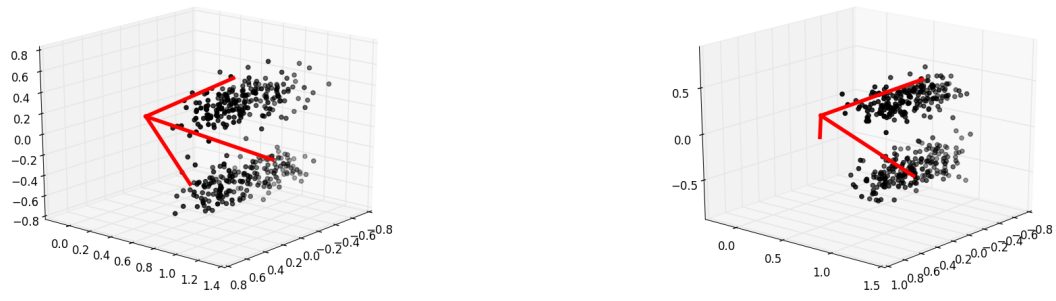
(b) Principal and Tangential Atoms

Figure 4.3 Example dictionaries to represent a 2-dimensional affine region.

Geometrically there are several different ways that a dictionary can be used to represent a manifold. To represent a k dimensional locally affine region of a manifold at least $k+1$ atoms are required. The distribution and orientation of the atoms relative to the manifold is the matter of interest. Figure 4.3 depicts two different ways to represent a two dimensional affine region in three dimensions. One potential distribution holds $k+1$ non-degenerate dictionary atoms pointing directly to the affine region. Linear combinations of these atoms can represent all other points on the affine plane, but not necessarily with a small ℓ_1 norm. An alternative distribution admits a single atom to be the affine offset, while k atoms span the tangent space. Throughout our discussion we will refer to atoms pointing directly to the manifold as principal atoms, and atoms lying in tangent spaces as tangential atoms.

4.2.2 Geometric Effect of Sparsity Parameter on Synthetic Data

To study how the sparsity parameters affect the atoms' distributions in the geometric interpretation, we first turn to synthetic data. Consider a one dimensional manifold consisting of two parallel line segments lying on an affine plane. The objective is to learn three atoms to represent this manifold and study how the optimal dictionaries change with sparsity parameters. Figure 4.4 depicts the manifold as well as the optimal dictionaries trained with different sparsity parameters using lasso with MOD.



(a) $\lambda = 0.1$ (b) $\lambda = 0.01$

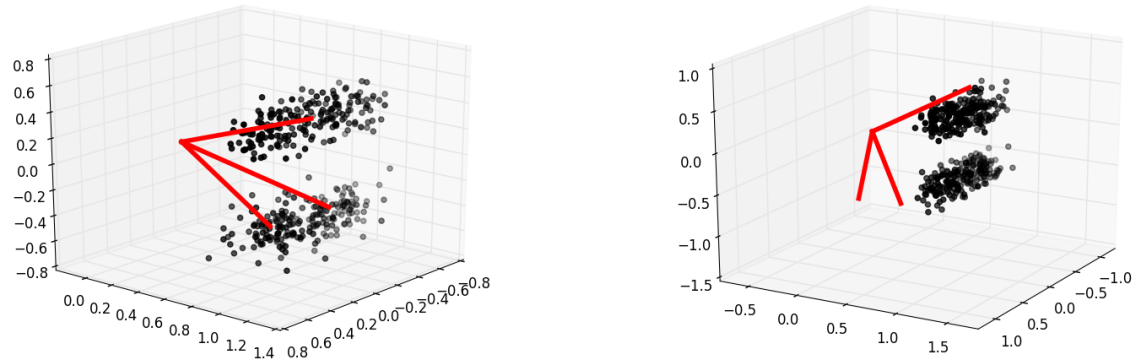
Figure 4.4 Dictionaries trained on synthetic data using lasso with MOD.

With the higher sparsity parameters, all three atoms are principal atoms, pointing approximately to the manifold. Each of these atoms acts as primary contributor of the data and has little tangential information. A significant number of signals use just two of the three atoms in their sparse representations. Such a dictionary achieves a good denoising performance, as each atom does point to the manifold.

With a lower sparsity parameter, two atoms function as principal atoms. The remaining atom points in between the two segments, but is not parallel to the line segments. The third atom does not fit into either role of principal or tangential. With this sparsity parameter the vast majority of the signals require use of all three atoms in their sparse representation. The atom that does not lie on the manifold, nor in the tangent space is an amalgamation of principal and tangential information. The cross contamination makes the denoising performance of this dictionary unsatisfactory.

To observe the effect of the sparsity parameter on dictionary learning when using OMP with K-SVD, a similar experiment was executed. Figure 4.5 depicts the manifold as well as the optimal dictionaries trained with different sparsity parameters using OMP with K-SVD. The sparsity parameters $\epsilon = 0.5$ and $\epsilon = 0.1$ were chosen to give a similar average fidelity as that obtained with $\lambda = 0.1$ and $\lambda = 0.01$.

With the higher sparsity parameter, all three atoms are principal atoms, almost identical to the dictionary learned with $\lambda = 0.1$. With the lower sparsity parameter, the dictionary is significantly different. None of the atoms point to the manifold, nor are they tangential. With the lower parameter, all three atoms are used in the representation of virtually all of the signals. As neither OMP nor K-SVD penalize large coefficients in the sparse representation, almost any basis is equally good with the sparsity parameter $\epsilon = 0.1$.



(a) $\epsilon = 0.5$ (b) $\epsilon = 0.1$

Figure 4.5 Dictionaries trained on synthetic data using OMP with K-SVD.

4.2.3 Geometric Effect of Sparsity Parameter on Image Patches

Figure 4.6 again depicts the effect of different sparsity parameters when training dictionaries using lasso with MOD. One approach to understand the different types of atoms in natural image dictionaries is to analyze some statistics of the dictionary atoms trained in Figure 4.6 (b). Visual inspection of the dictionary atoms suggest there are two types of atoms, one class consisting of lines and gradientes, another of checkboards and spots. We will call the two groups "Class One" and "Class Two". Manual classification of these atoms into these two groups is depicted in Figure 4.7.

To understand the differences between these two classes of atoms, we consider a few measures to characterize how the atoms are used in training. Let D, X be the optimal solution using Algorithm 2 trained on a collection of N signals, S. Then $(X^T)_i$ is the i^{th} row of X corresponding to the use of the i^{th} dictionary atom, D_i. Let

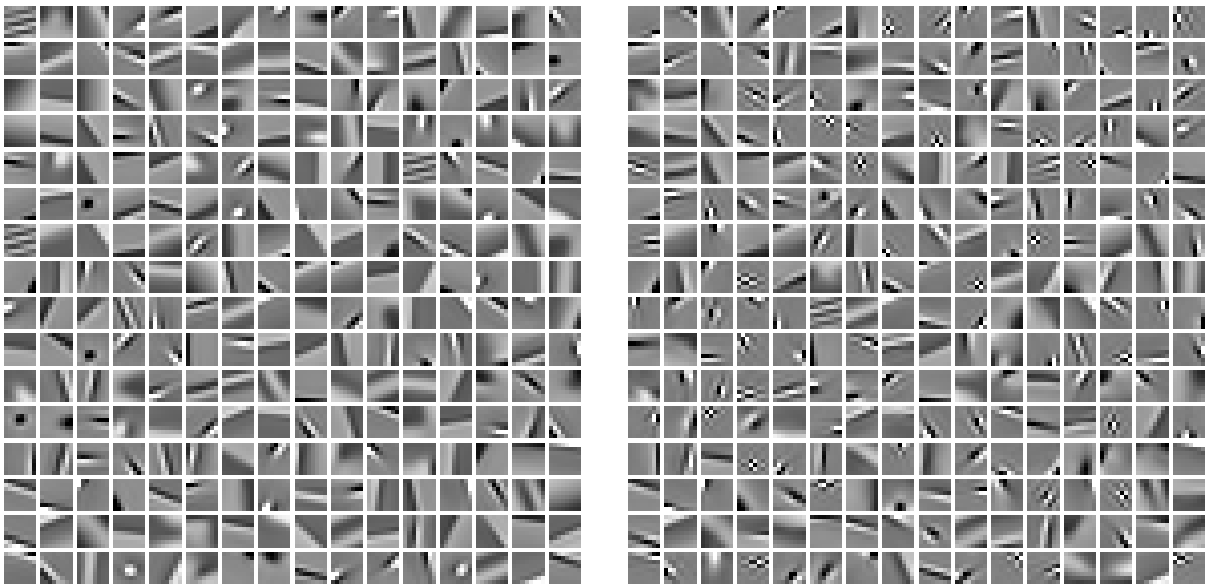
$$\theta(i) = \min_{j=1}^{N} \quad \arccos\left(\frac{D_i \cdot S_j}{S_j \cdot S_j}\right)$$

be the minimal angle between the i^{th} atom and all N signals. The probability that the i^{th} atom is used in any given sparse representation can be calculated

$$p(i) = \frac{||x_i||_0}{N}.$$

When the i^{th} atom is used,

$$m(i) = \frac{||x_i||_1}{||x_i||_0}$$



(a) $\lambda = 0.1$ (b) $\lambda = 0.01$

Figure 4.6 Dictionaries trained on natural image patches data using lasso with MOD.

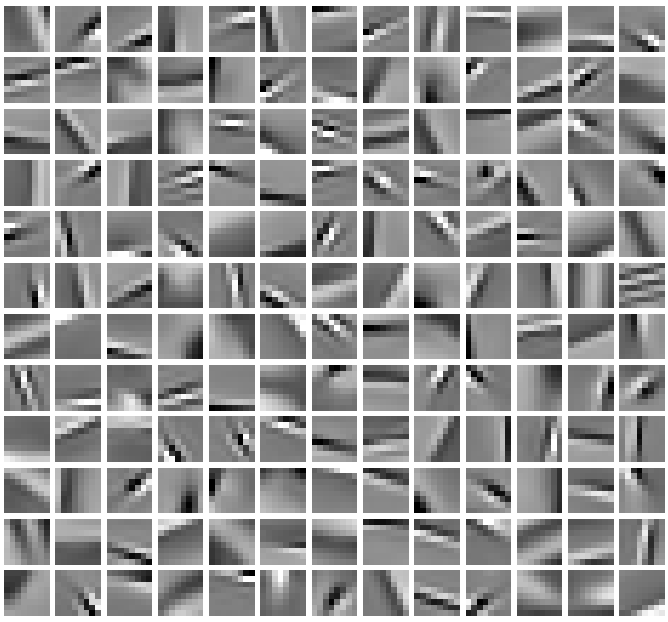
Table 4.1 Usage statistics of two different classes of atoms trained using lasso with MOD.

Class	$\theta(i)$		$p(i)$		$m(i)$	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
One	0.4919	0.2085	0.1411	0.0225	0.0466	0.0199
Two	0.9198	0.1727	0.1917	0.0295	0.0244	0.0075

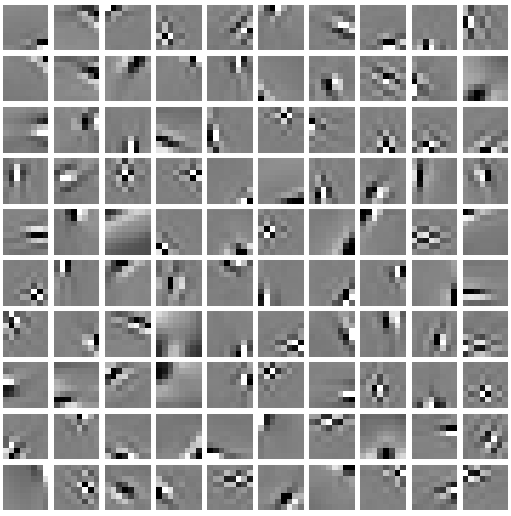
is the average magnitude of use. Table 4.1 shows the means and standard deviations of these values over the two classes.

The average minimal angles between the two classes, approximately 25 degrees for class one and 45 degrees for class two, suggest that the class one atoms are much more similar to principal atoms than class two. With the average angle being approximately 20 degrees closer to that of signals than class two, the class one atoms appear to point more directly to the manifold. From $p(i)$ and $m(i)$ there are two important observations from Table 4.1. First, the atoms in class one are approximately 5% less likely to be used in sparse representations than the atoms in class two. Second, when the atoms in class one are used, they are used with significantly higher magnitudes than those in class two. In other words, tangential atoms are used more often, but with a smaller magnitude, as modifiers rather than principal contributors. The geometric interpretation of this suggests that the radius of curvature is small relative to the affine offset.

Figure 4.8 depicts the effect of training a dictionary using OMP with K-SVD on the same set of signals and with analogous sparsity parameters. The sparsity parameters $\epsilon = 0.17$ and $\epsilon = 0.02$ were



(a) Class One



(b) Class Two

Figure 4.7 Dictionary atoms trained with $\lambda = .01$ regularization using lasso with MOD and visually split into two classes.

Table 4.2 Usage statistics of two different classes of atoms trained using OMP with K-SVD.

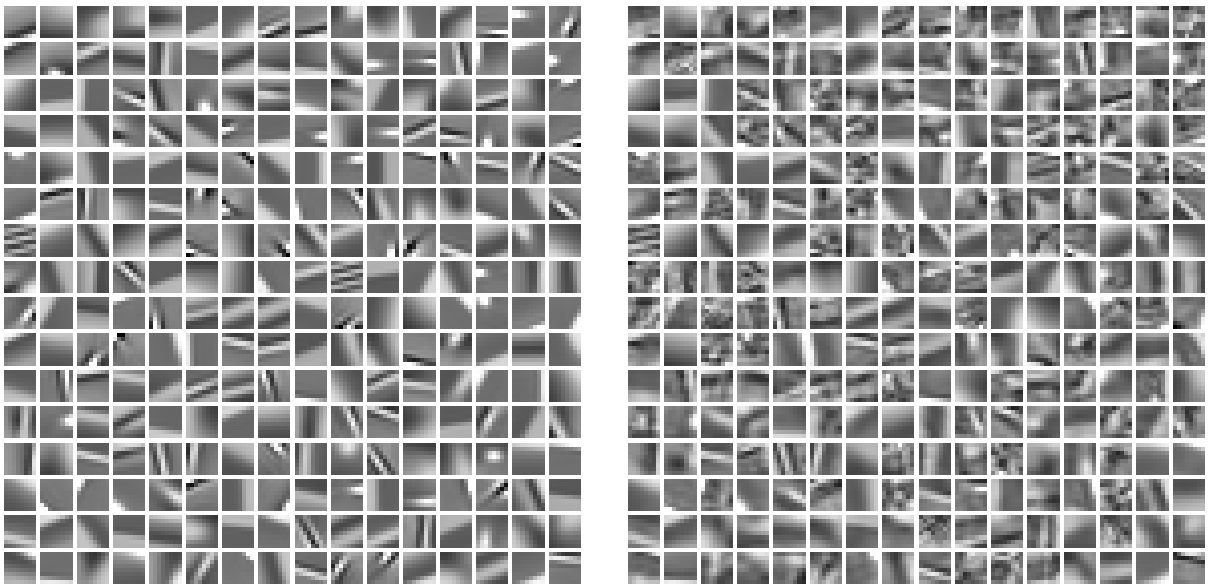
Class	$\theta(i)$		$p(i)$		$m(i)$	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
One	0.3317	0.1474	0.0418	0.0602	0.3198	0.1607
Two	0.5977	0.1143	0.1169	0.0782	0.1618	0.0323

chosen to give a similar average fidelity as that obtained with $\lambda = 0.1$ and $\lambda = 0.01$.

All of the atoms in the dictionary with the $\epsilon = 0.17$ sparsity parameter are representative of natural image patches. These atoms consist of lines, gradients, and other visually appealing structures. Relative to its counterpart trained with lasso with MOD, the lines in the dictionary appear to be more like step functions than ridges.

The dictionary trained with the lower regularization also has several visually appealing atoms that have counterparts in the dictionary trained with $\epsilon = 0.17$, but in addition has several atoms that appear noisy. Visual classification of these atoms is depicted in Figure 4.9. The Class Two atoms trained using OMP with K-SVD appear very different from their lasso with MOD counterparts. These atoms do not have a small support and appear to be much closer to white noise. The statistics of minimal angle, probability of use, and magnitude of use for these classes can be seen in Table 4.2.

The relative trends of class one and two are identical to those classified from the dictionary



(a) $\epsilon = 0.15$ (b) $\epsilon = 0.008$

Figure 4.8 Dictionaries trained on natural image patches data using OMP with K-SVD.

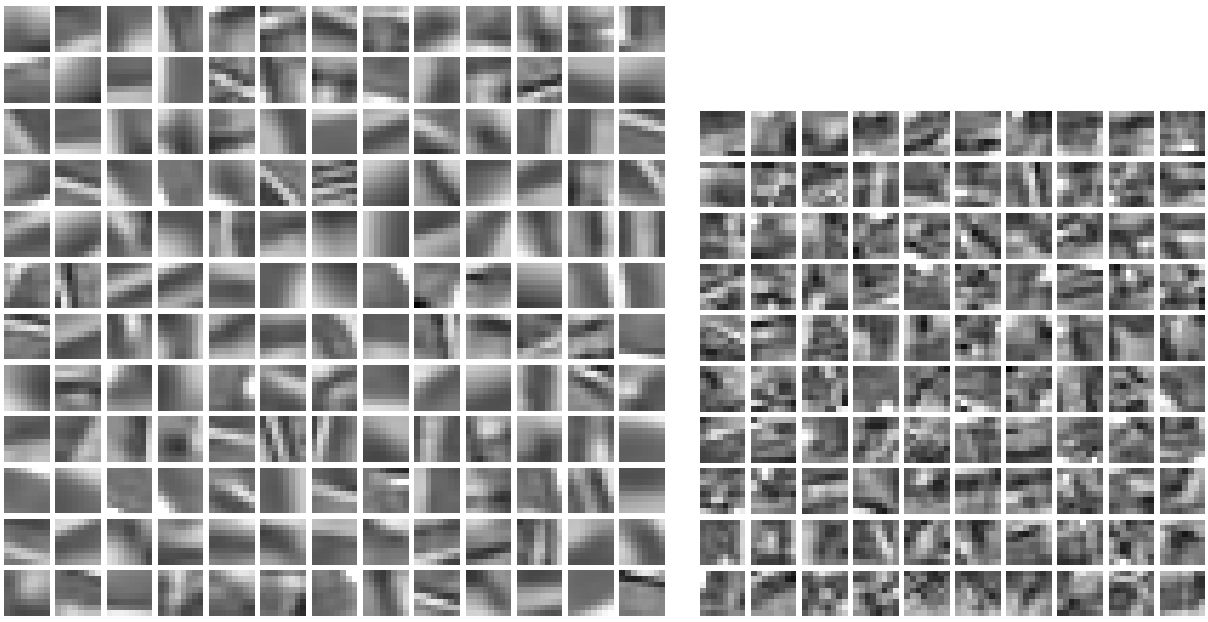
trained using lasso with MOD. The class one atoms have a much smaller angle than class two. Also, the class one atoms are used less often, but with higher magnitudes than class two atoms.

Comparing the lasso with MOD and OMP with K-SVD counterparts also provides some interesting insights. The OMP with K-SVD atoms have much smaller minimal angles with the signals than those trained with lasso and MOD, suggesting that in some ways they are much more like principal atoms. The probability of use for the OMP with K-SVD is also significantly lower than their lasso with MOD counterparts. Intuitively this can be attributed to the minimal angles, since the atoms are much more aligned with the signals, fewer atoms are needed to point to the manifold. This also explains the differences in average magnitude between lasso with MOD and OMP with K-SVD.

In both the lasso with MOD and OMP with K-SVD formulations there is no structure separating atoms into different roles. Despite all atoms being treated identically in their respective formulations, we see that atoms take on specialized roles. The different roles result in different usage patterns both in the ℓ_0 and ℓ_1 sense.

4.2.4 Single Step Update

Another method to probe the differences in dictionaries trained with different sparsity parameters is to alter the regularization during training and observe the effect. To this end, we train a dictionary until convergence with one regularization, then adjust the regularization and do a single interleaved cycle to update $D_{(i)}$,



(a) Class One (b) Class Two

Figure 4.9 Dictionary atoms trained with $\lambda = .01$ regularization using OMP with K-SVD, visually split into two classes.

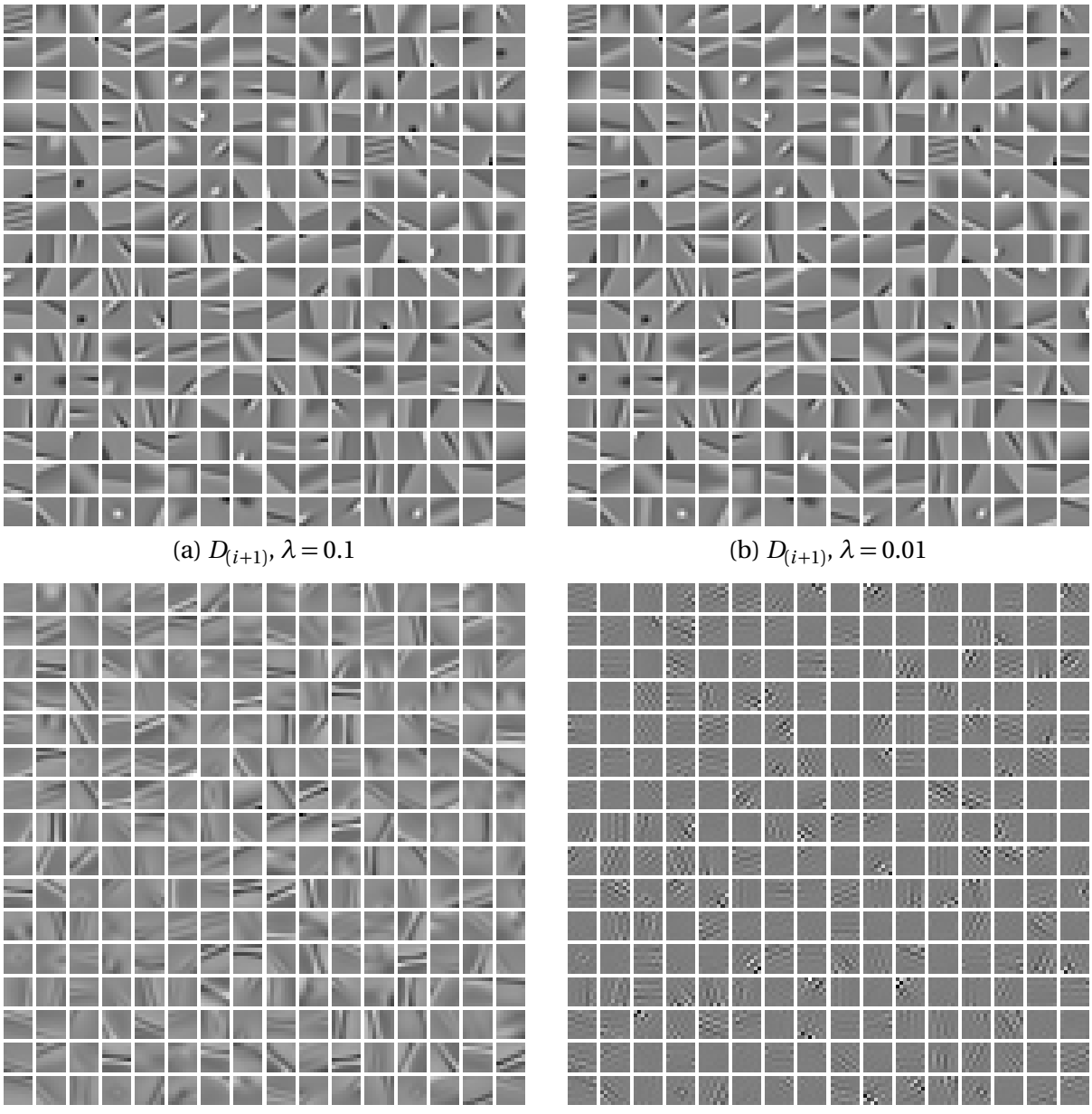
$$\begin{aligned} X_{(i+1)} = \quad & \arg\min_X \quad \left|\left|D_{(i)}X - S\right|\right|_F^2 + \lambda ||X||_1 \\ D_{(i+1)} = \quad & \arg\min_D \quad \left|\left|DX_{(i+1)} - S\right|\right|_F^2 \\ & \text{subject to} \quad \left|\left|D_j\right|\right|_2^2 \leq 1, \text{ for } j = 1, \ldots, n. \end{aligned}$$

For a controlled comparison we also consider the result of a single interleaved cycle without changing the regularization.

Using our trained dictionary shown in Figure 4.1 (a), we can observe what a single update step looks like with sparsity parameters $\lambda = 0.1$ and $\lambda = 0.01$. Figure 4.10 shows the updated dictionaries with two different regularizations as well as their change from the initializing dictionary.

Visually we observe that the two different dictionaries are essentially indistinguishable. The differences with the original dictionary provide more insight into what effect different regularizations have on a single step update. Considering the differences to the original dictionary, we can see that the dictionary learning continued with $\lambda = 0.1$ seems to be reinforcing the structure already present. On the other hand, the dictionary learning with $\lambda = 0.01$ seems to be fundamentally changing their structure.

To numerically measure the effect of the update, we measure the minimal angle between each



(a) $D_{(i+1)}$, $\lambda = 0.1$

(b) $D_{(i+1)}$, $\lambda = 0.01$

(c) $D_{(i+1)} - D_{(i)}$, $\lambda = 0.1$

(d) $D_{(i+1)} - D_{(i)}$, $\lambda = 0.01$

Figure 4.10 Single update step from learned dictionary with different regularizations, and the differences in steps.

Table 4.3 Statistics of minimal angles after a single X and D update.

Minimal Angle	Original	$\lambda = 0.1$	$\lambda = 0.01$
Mean	0.3904	0.3902	0.3925
S.D.	0.1868	0.1860	0.1877

atom and all of the signals, and calculate the averages and standard deviations over the atoms. These results are tabulated in Table 4.3. Interestingly the average minimal angle between atoms and signals goes slightly down when updating with $\lambda = .1$, but goes up when updating with $\lambda = 0.01$. This indicates that as the regularization is lowered, a significant portion of the atoms are deviating from the role of primary contributor to a signal. When updating with $\lambda = 0.01$ the atoms are being adjusted to leave the manifold and to lie in the tangent space.

4.3 Proposed Modification

Despite no design being built into the dictionary learning formulations to encourage atoms to fill specific roles, they split into a variety of roles from principal to tangential atoms. These different classes of atoms have different statistical characteristics in their sparse representations. We propose to employ the statistical differences to encourage atoms to fill specific roles in the lasso with MOD formulation. By introducing the product with a diagonal weight matrix into the ℓ_1 regularization, we can encourage specific properties in different atoms.

$$\begin{aligned} &\text{argmin}_{D,X} \ \frac{1}{2}||DX - S||_2^2 + \lambda \left|\left|\text{diag}(w)X\right|\right|_1, \\ &\text{Subject To } ||D_i||_2 \leq 1 \text{ for } 1 \leq i \leq k. \end{aligned} \tag{4.4}$$

In natural image patch dictionaries, the atoms with the lowest weights will be encouraged to be used more in the ℓ_1 sense. Since the principal atoms are used more in the ℓ_1 sense relative to the tangential atoms, the atoms with the lowest weights will become principal atoms. Using regularization to segment the atoms within the dictionary will reduce the cross contamination of principal and tangential information potentially resulting in better denoising properties. An additional advantage of encouraging atoms to fill specific roles is versatility. When solving problems where some information is less desirable, the corresponding atoms can be weighed more heavily or removed completely from the optimization. Overall, having a variety of weights for different atoms will produce higher quality and more flexible dictionaries. A geometric interpretation of the weights in formulation (4.4) effectively changes the lengths of the atoms. Weights smaller than 1.0 correspond to an effective lengthening of the corresponding atoms while atoms with lengths greater than 1.0, of an effective shortening. The longer effective atoms are more effective in representing information where a

significant component is necessary, encouraging their use to represent some information.

4.3.1 Proposed Modification on Synthetic Data

Consider the synthetic data with our proposed solution. We designate two of the atoms to represent principal information, and the third to represent the tangent space. Figure 4.11 depicts dictionaries trained where this structure was encouraged with weights using the formulation in (4.4).

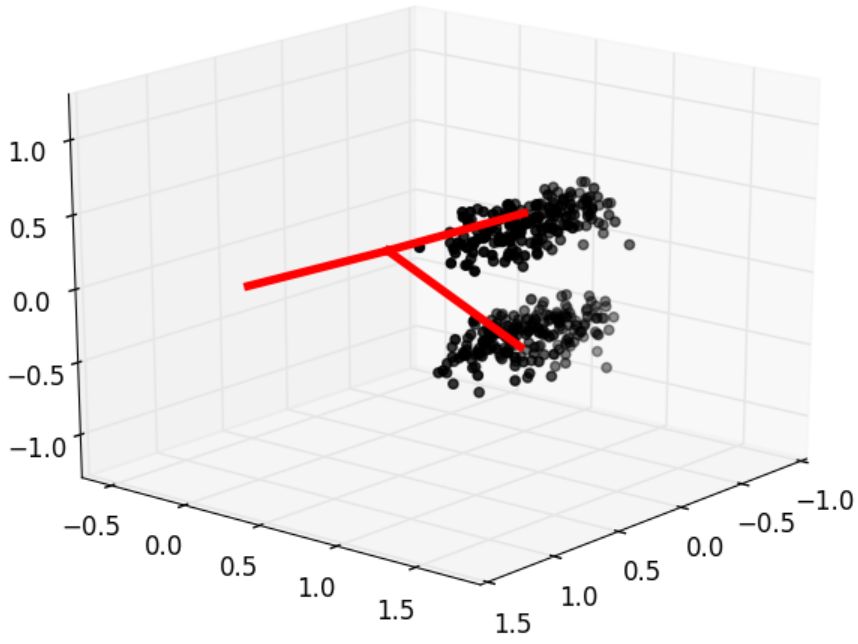


Figure 4.11 Dictionary trained on synthetic data by (4.4) with effective regularizations of 0.009, 0.009, and 0.012.

In this dictionary, the two atoms with the lower effective regularization lie approximately on the manifold, one on each line segment. The third atom, with the higher effective regularization, lies parallel to the tangent space of the manifold. This atom is used in almost every sparse representation, but with a much smaller magnitude. The diversity in the effective regularizations encourages specific

Table 4.4 Usage statistics of a dictionary trained with a variety of effective regularizations.

Weight Parameter	$\theta(i)$		$p(i)$		$m(i)$	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
0.9	0.5173	0.2668	0.1930	0.0362	0.0471	0.0214
1.0	0.7614	0.2359	0.1594	0.0272	0.0266	0.0088
1.2	1.0111	0.1770	0.0930	0.0122	0.0170	0.0035

atoms to play certain roles. Through this diversity we help eliminate cross-contamination of different classes of information in dictionary atoms, resulting in a higher quality dictionary for denoising.

4.3.2 Proposed Modification on Image Patches

Returning to natural image patches, our proposed modification is to train a dictionary with an array of weights. Consider a dictionary, which we denote D3, trained with sparsity parameter $\lambda = 0.01$, and a weight vector segmented into three groups. The first group of 128 atoms have a weight parameter of 0.9, and these atoms will be most representative of principal atoms. The second group of 64 atoms have a weight parameter of 1.0, and the last 64 atoms have a weight parameter of 1.2. These weight values average to 1.0 and were chosen heuristically to produce a slight spread in effective regularization. Both of these groups of atoms will be more tangential in nature. Figure 4.12 depicts the dictionary trained with this weight scheme. With these weights, initialization becomes increasingly important to ensure the use of all dictionary atoms.

The visual difference between the three groups of atoms is quite striking. The first 128 atoms trained with the the smaller weight are representative of low frequency filters. The second group of 64 atoms with the middle regularization are visually similar to moderate frequency filters. The last 64 atoms are most similar to high frequency impulse filters. The statistics in Table 4.4 indicate how likely and with what magnitude the different groups of atoms are used during training. Notably, amongst the first two groups of atoms, the atoms with the lower regularization are used less often in the ℓ_0 sense and more often in the ℓ_1 sense.

Segmenting the dictionary into groups provides the additional benefit of versatility in sparse coding problems. In problems where tangential information is ill-suited, the corresponding dictionary atoms can be excluded from the sparse representation. In addition to the full dictionary of 256 atoms, D3, we also consider the dictionaries consisting of only the first group of 128 atoms, denoted D1, and the first two groups of 192 atoms, D2.

Applying these dictionaries to our image denoising problems we can see how they compare to our previous dictionaries in Figure 4.13. The dictionary D3 behaves very similarly to the dictionary trained with $\lambda = 0.01$ as they both contain both tangential and primary information, while the dictionary D1 performs most similarly to the dictionary trained with $\lambda = 0.1$ as they both contain

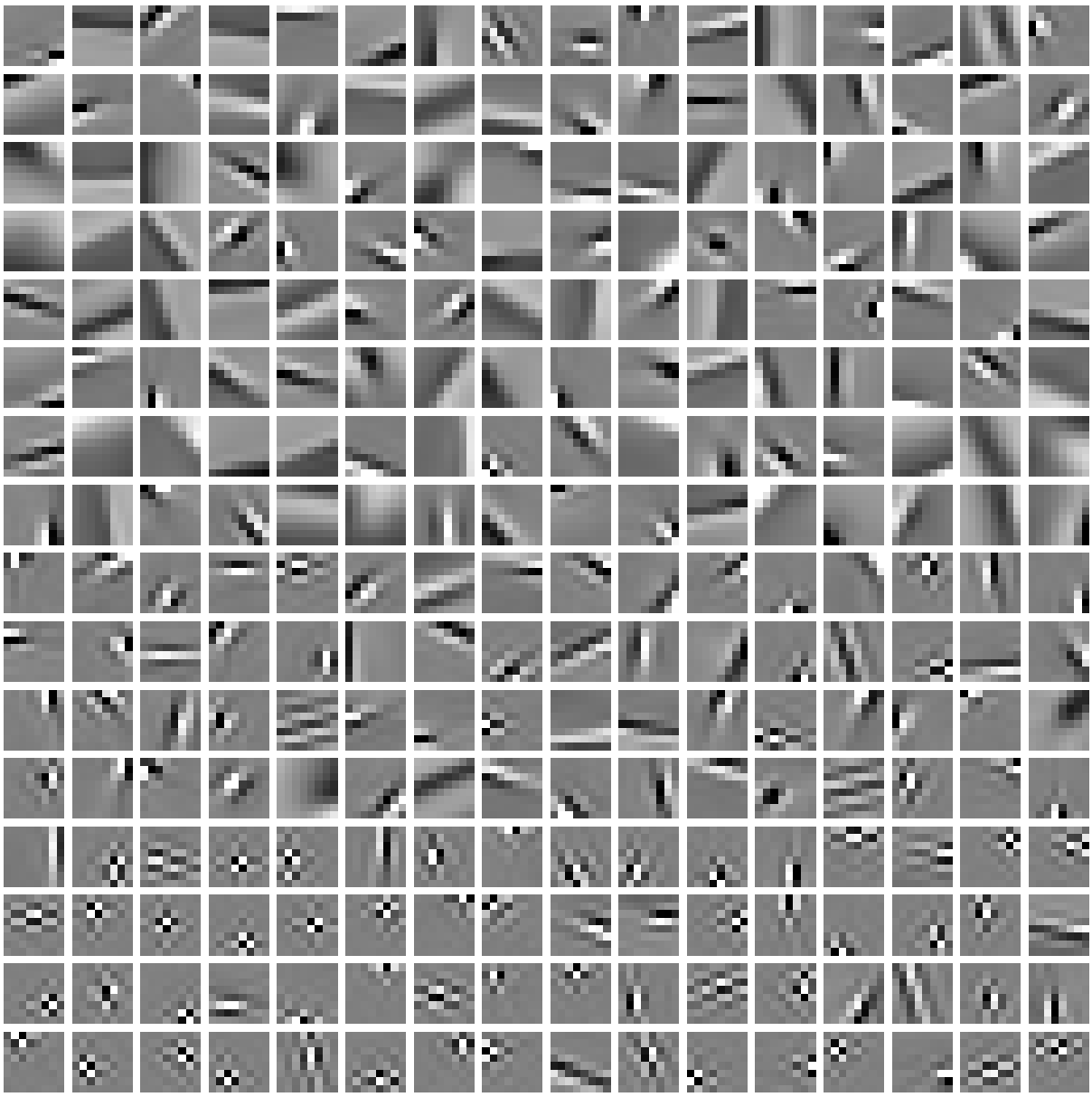
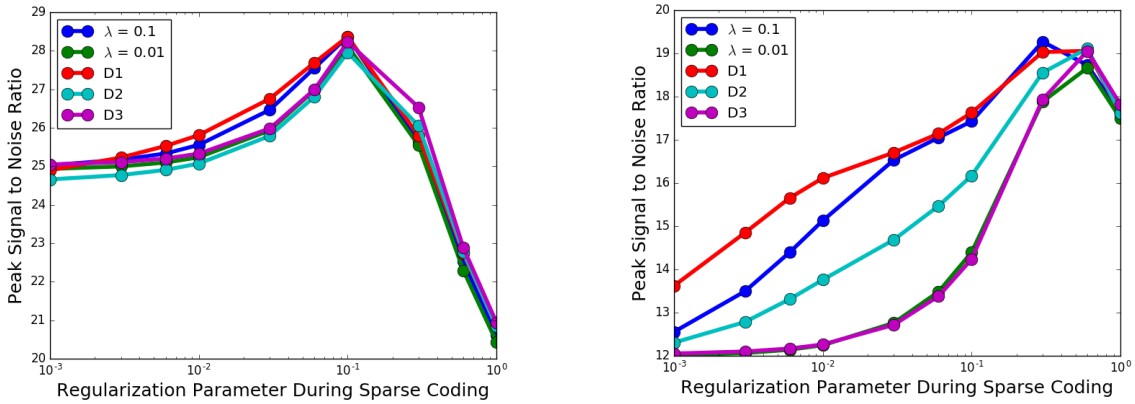


Figure 4.12 Dictionary, D3, trained on natural image patches with different sparsity weights on different atoms using formulation 4.4.

mostly primary information.

4.4 Conclusion

This chapter addresses the differences in dictionaries trained with different regularizations and their subsequent use. Depending on the type and intensity of noise, some dictionary qualities may or may



(a) White Noise $\sigma = 25$

(b) Salt and Pepper 20% Corrupted

(c) White Noise $\sigma = 5$

(d) Salt and Pepper 4% Corrupted

Figure 4.13 Comparison of dictionaries at a variety of noise levels and regularizations.

not be desirable. Different features are explained in a geometric sense and a method is proposed and shown to allow for control of feature presence. Dividing dictionaries into groups with different effective regularizations does two things: it helps prevent the contamination of different types of information, and it structures the dictionaries making them more versatile.

Future work in this field will include the statistical analysis of how atoms are used in conjunction with one another. One possible avenue of improvement is the inclusion of a dynamic group norm that captures the statistical relationship between different types of atoms. Because the tangent space is different at every point on the manifold, encouraging a group norm such as those in [75, 103, 114, 161] may provide additional control to the geometric structure of dictionary atoms.

Another avenue of research is the possibility of directly encouraging different structures in the atoms. Notably, additional regularization on the atoms can promote characteristics such as

smoothness or patterns by different regularizations such as total variation regularization in spatial domain [126] or sparsity regularization in the frequency domain. Also, alternatives to the unit length constraint in the dictionary learning and their effect on the geometry of the atoms are of significant interest.

CHAPTER

5

VARIATIONS OF DICTIONARY LEARNING AND APPLICATIONS

5.1 Pansharpening via Triple Factorization Dictionary Learning

5.1.1 Introduction

The research in this section is published and can be seen in [133]. Data fusion methods integrate different sources of information to construct a single consistent representation. Pansharpening is a data fusion problem which aims to combine high spatial and high spectral resolution data sets. Specifically, given a high resolution panchromatic, or black and white, image, the goal is to sharpen corresponding low resolution multispectral imagery to obtain high resolution multispectral images.

One approach to solving the pansharpening problem is dictionary learning. After training dictionaries to represent the different bands of a multispectral image, the trained model can be used to sharpen the low resolution images. The proposed dictionary learning model for data fusion can be applied to the pansharpening problem. Training and validating the model on real world data demonstrates the data fusion formulation's viability for pansharpening.

5.1.2 Background

Dictionary learning is a method that finds a set of vectors with the goal of sparsely representing a family of signals. Typically, one wishes to represent a set of signals S, with a dictionary D, whose columns are called atoms, and sparse assignment vectors X. A classic approach to this is the minimization problem

$$\begin{aligned} \underset{D,X}{\arg\min} \quad & \frac{1}{2}\|DX-S\|_F^2+\lambda\|X\|_1, \\ \text{subject to} \quad & \|D\|_{2,\infty}\leq 1, \end{aligned}$$

where λ is referred to as the regularization parameter that controls sparsity, and the constraint on D keeps the dictionary bounded [1]. More precisely the constraint

$$\|D\|_{2,\infty}=\max_i \sqrt{\sum_j |D_{j,i}|^2}\leq 1,$$

keeps the Euclidean norm of each atom bounded above by 1. After learning a dictionary, one can use that dictionary to sparsely represent signals since $DX\approx S$. Dictionary learning has had numerous uses, of which two notable applications are image denoising and inpainting.

When dictionary learning is applied to images, the signals are constructed by vectorizing overlapping patches or subimages of a fixed size. Several preprocessing techniques can be applied to the vectorized patches such as centering, normalizing, and whitening [77]. If one wishes to reconstruct the image after sparsely encoding the overlapping patches it is typical to resolve the multivalued pixels by averaging.

Coupled dictionary learning trains two different dictionaries to represent two different but related signals. A typical formulation for coupled dictionary learning is

$$\begin{aligned} \underset{D_i,X_i,W}{\arg\min} \quad & \frac{\sigma_1}{2}\|D_1X_1-S_1\|_F^2+\frac{\sigma_2}{2}\|D_2X_2-S_2\|_F^2+ \\ & \frac{\gamma}{2}\|WX_1-X_2\|_F^2+\lambda\|X\|_1, \\ \text{subject to} \quad & \|D_i\|_{2,\infty}\leq 1 \text{ for } i=1,2. \end{aligned}$$

Coupled dictionary learning has had significant success in image superresolution as well as image analogies [21, 152].

5.1.3 Related Works and Our Contribution

There are a variety of approaches to the pansharpening problem [5, 102]. Two of the more basic approaches to pansharpening that we will compare against are Gram-Schmidt (GS) [84] and principal component analysis (PCA) [27]. GS and PCA provide a good benchmark for validation of

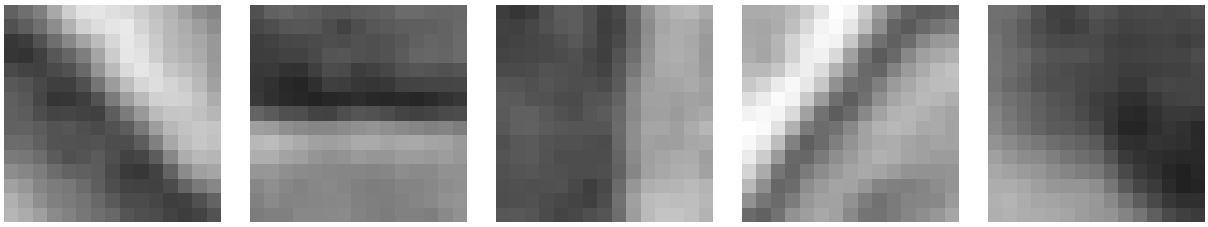


Figure 5.1 Sample dictionary atoms transformed to the reference image space.

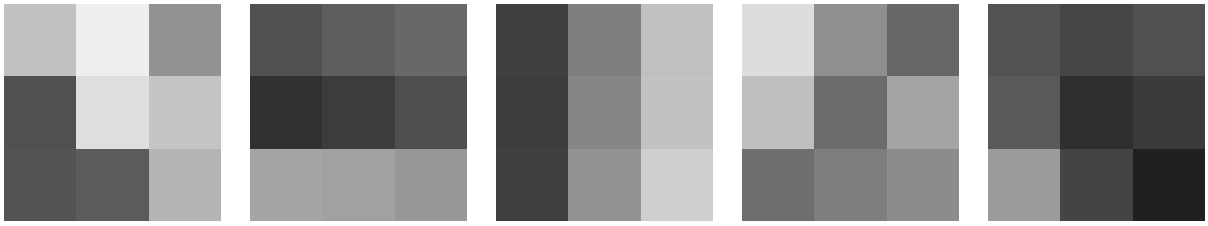


Figure 5.2 Sample dictionary atoms transformed to the degraded input image space.

problem formulation. There have also been a few previous works using variations of dictionary learning or sparse coding for pansharpening [28, 100]. Our investigation entails adding an additional factorization to coupled dictionary learning to solve the pansharpening problem.

We propose a different and novel variation of coupled dictionary learning with the goal of demonstrating that an additional factorization provides some new flexibility that can be suitable for pansharpening. Variations of our formulation should be applicable to a variety of problems. The goal of this section is to demonstrate that this formulation is applicable to the pansharpening problem.

5.1.4 Problem Formulation

One may observe that in coupled dictionary learning there is often very similar information stored in the different dictionaries. This is especially true when dealing with multispectral images, as there is a lot of overlapping information in different spectral bands. Since there is a lot of redundant information in the different bands, it is natural to construct a model to exploit this. This is the motivation for our proposed method, to find one dictionary that contains information from all bands, as well as a set of linear transformations taking the cumulative information to a specific band. Specifically, our formulation for the pansharpening problem

$$\begin{aligned} \underset{P_i,D,X}{\arg\min} \quad & \sum_i \frac{\sigma_i}{2}\|P_i D X - S_i\|_F^2 + \sum_i \lambda_i \|P_i\|_1 + \lambda \|X\|_1, \\ \text{subject to} \quad & \|D\|_{2,\infty} \leq 1, \end{aligned}$$

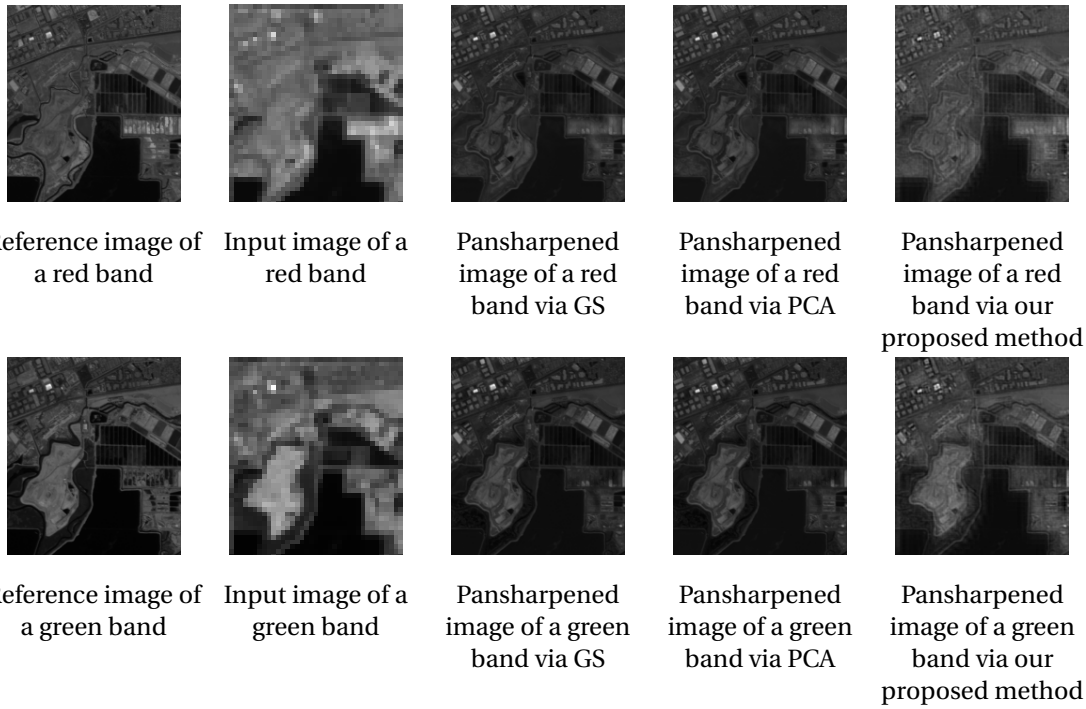


Figure 5.3 Some sample reference and degraded images to be sharpened from Moffett data set

trains a dictionary containing the information from every band D, as well as a set of sparse linear transformations P_i's from the general dictionary to the space of a particular band. Our proposed scheme produces an approximate triple factorization of the signal, $S_i \approx P_i D X$. The sparsity regularization on the linear transformations is appropriate when dealing with images because the value of any given pixel is most strongly correlated with its neighbors. One can imagine other applications where different regularization may be more appropriate.

One potential advantage of this formulation can be seen by considering the dimensionality of our dictionary. In classic dictionary learning, each atom needs to be the dimensionality of the signal. In coupled dictionary learning there are multiple signals, each signal with its own dictionary of corresponding dimensionality. In our proposed formulation there is a single dictionary whose dimensionality is adjustable, so it can be chosen appropriately to be able to hold the cumulative information from each signal. In addition to the single dictionary in our method, there are several sparse linear transformation matrices that map the generic dictionary to the space of a specific signal.

5.1.5 Algorithm

To perform the optimization, we used a classic alternating minimization algorithm consisting of fixing all but one variable at a time. For each P_i update and X update we utilized alternating direction method of multipliers (ADMM) [15], and a variation of method of optimal directions (MOD) was used for the D update [53].

$$\begin{aligned} X^* = &\quad \arg\min_X \quad \sum_i \frac{\sigma_i}{2}\|P_i D X - S_i\|_F^2 + \lambda\|X\|_1, \\ D^* = &\quad \arg\min_D \quad \sum_i \frac{\sigma_i}{2}\|P_i D X - S_i\|_F^2, \\ &\quad \text{subject to} \quad \|D\|_{2,\infty} \le 1, \\ P_i^* = &\quad \arg\min_{P_i} \quad \frac{\sigma_i}{2}\|P_i D X - S_i\|_F^2 + \lambda_i\|P_i\|_1. \end{aligned}$$

5.1.6 Training Procedure

To train our model we require both reference images for each band as well as the degraded input images that are to be processed during validation. For each band in the data set, our model trains one linear transformation to the reference images and one linear transformation to the input images. Along with the pairs of reference and input images, the single high resolution panchromatic image is also used during training. One half of the data was used for training the model, the other half was used as a validation set.

5.1.7 Validation Procedure

To test the trained model we applied it to the validation half of the data set. To compare our method to others we used three methods to measure the quality of the sharpened image. Cross correlation (CC) measures the spatial similarity of the sharpened image and the reference image. Spectral angle mapper (SAM) measures the spectral similarity between hyperspectral images. Root mean squared error (RMSE) measures the ℓ_2 distance between the reference and sharpened image. Numerous pansharpening methods involve interpolating a low resolution image to a high resolution and use the interpolated image as an approximation in which to fill in details. Our method is also suitable for this, but for clarity we skipped this to demonstrate the flexibility of our proposed technique. For a comparative study, we referred to two of the more basic pansharpening techniques, GS and PCA.

5.1.8 Moffett Dataset Results

The publicly available Moffett data set is a hyperspectral image of 224 bands that contains both urban and vegetative regions taken by Airborne Visible Infrared Imaging Spectrometer (AVIRIS).

Table 5.1 Comparison of methods.

	GS	PCA	Proposed Method
CC	0.917	0.906	**0.935**
SAM	12.95	13.4	**10.6**
RMSE	420.5	445.1	**364.6**

Our method was compared using the framework developed in [102]. In summary, the reference image was degraded using Wald's protocol by a factor of 5 to construct the low resolution image to be sharpened [151]. The patches for the high resolution and low resolution images were chosen to be 15 by 15 and 3 by 3 respectively. The dictionary was chosen to be 225 by 300. The only image preprocessing steps taken were centering of the signal. Samples of the trained dictionary atoms transformed to some bands are shown in figure's 5.1 and 5.2. Applying GS, PCA, and our proposed method to this data set we obtain the numeric results shown in Table 5.1.

Figure 5.3 depicts the reference images, input images, and results from the different methods. Visually one can see that the proposed method provides the best fidelity to the relative intensities of the different bands. Unfortunately our method also shows evidence of blocking artifacts, most notable around the high contrast areas at the bottom. These artifacts are a known disadvantage to dictionary learning techniques that originate from centering the data. The artifacts cause the detail in the GS and PCA methods to be visually more precise. Numerically from the three measures of image quality, our method is superior to both GS and PCA as shown in the table.

5.1.9 Conclusions

This work demonstrates that a triple factorization in coupled dictionary learning achieves better results than some of the more basic pansharpening techniques. Other variations of the proposed data fusion model are suitable to different problems and will be the subject of future work. One future avenue to explore is different methods of regularization for the linear transformations to each signal space.

5.2 Hierarchical Convolutional Dictionary Learning

5.2.1 Introduction

Computer vision and object detection is an active developing field with far reaching applications. There are a variety of object detection approaches including histograms of oriented gradients [39], scale invariant feature transform [104], and convolutional neural networks [82].

The use of dictionary learning for object detection has also been studied before with varying

degrees of success, [158]. One particular challenge is that dictionaries trained on sufficiently complex images are typically comprised of lines, gradients, and very simple patterns. For computer vision and object detection, dictionaries with more complex objects may be desirable. The inability to learn complex objects from images in dictionary learning poses a significant challenge that limits its efficacy for object detection. In this section we explore a hierarchical variation of convolutional dictionary learning with the objective of learning more complex features for an object bank.

5.2.2 Convolutional Dictionary Learning and Max Pooling Background

Convolutional dictionary learning is a translation-invariant variety of dictionary learning that finds single valued reconstructions [162]. The convolutional dictionary learning problem can be formulated as an optimization problem,

$$\min_{D_i, X_i} \| \sum_i D_i \star X_i - S \|_F^2 + \lambda \sum_i^n \|X_i\|_1 \tag{5.1}$$

where $\{D_i\}$ is a set of n matrices, or dictionary filters, $\star$ is the convolution operator, $\{X_i\}$ is a set of n coefficient maps, and S is a signal. This optimization aims to represent the signal S as a sum of convolutions of dictionary filters with sparse matrices.

Efficient algorithms for solving the convolutional dictionary learning problem alternatively update $\{D_i\}$ and $\{X_i\}$ until convergence [155]. To handle the convolution, the problem can be transformed with the Discrete Fourier Transform (DFT) so that the convolution becomes a Hadamard product in the frequency domain. The problem 5.1 is solved by alternating optimizion of the two sub problems,

$$\min_{X_i} \| \sum_i D_i \star X_i - S \|_F^2 + \lambda \sum_i^n \|X_i\|_1, \tag{5.2}$$

$$\min_{D_i} \| \sum_i \hat{D}_i \odot \hat{X}_i - \hat{S} \|_F^2 \tag{5.3}$$

where $\hat{A} = \text{DFT}(A)$ is used to denote the DFT of a variable. For both updates, Alternating Direction Method of Multipliers (ADMM) can be utilized to solve the optimization, see [155] for details.

Preprocessing for convolutional dictionary learning typically includes applying a high pass filter to the signal. This step is analogous to centering the signals in patch based dictionary learning to avoid ringing effects. One technique to accomplish this uses Tikhinov regularization to do a low pass filtering, which can then be used to find a high pass component. Another challenge in convolutional dictionary learning is boundary handling. One approach to avoid undesirable boundary effects pads the images and applies a mask to the fidelity function [154].

Pooling operations are generally nonlinear transformations that down sample an image. Max pooling segments an image into non-overlapping squares and each segment is represented by the

value in the segment that is largest in magnitude. In convolutional neural networks max pooling has been found to outperform other pooling methods [129].

5.2.3 Hierarchical Convolutional Dictionary Learning

In this section we propose using a hierarchy of convolutional dictionary learning with max pooling to construct successively more complex filters. Though this proposed method easily extends to an arbitrary number of hierarchies, for simplicity of proof of concept we study the case of just two hierarchies. The first hierarchy learns a dictionary and sparse representation with the convolutional dictionary learning formulation applied to the signal,

$$\min_{D_i^{(1)},X_i^{(1)}} \|\sum_{(i)} D_i^{(1)} \star X_i^{(1)} - S\|_F^2 + \lambda \sum_i^n \|X_i^{(1)}\|_1. \tag{5.4}$$

A second hierarchy also does convolutional dictionary learning, but is applied to the sparse representation found in the first hierarchy, $X_{(i)}^1$, after a step of max pooling,

$$\min_{D_i^{(2)},X_i^{(2)}} \|\sum_{(i)} D_i^{(2)} \star X_i^{(2)} - \text{maxpool}(X^{(1)})\|_F^2 + \lambda \sum_i^n \|X_i^{(2)}\|_1. \tag{5.5}$$

Applying the machine learning technique to the pooled sparse representation, $X^{(1)}$, learns the relationships between the first hierarchy dictionary filters.

To demonstrate this method we apply our dictionary learning and max pooling scheme to images from the Extended Yale Face Database B [62]. This database includes 28 individuals, each with several different expressions and lighting angles. For our experiment we chose a collection of 10 images of subjects to apply in our dictionary learning scheme.

5.2.4 Results and Discussion

The first hierarchy of our proposal constructs a dictionary of 8 filters of size 8 by 8. Figure 5.4 depicts $D_i^{(1)}$, the collection of 16 filters trained on 10 faces from the database. We can see that in the first hierarchy the atoms are primitive lines and gradients that can be used to sparsely represent the training images.

A second hierarchy is then learned on the pooled sparse representations corresponding to $D^{(1)}$. The second hierarchy dictionary captures the relationships of filters used in combination. The pooling provides flexibility in the spatial use of the filters. Figure 5.5 depicts the convolution of down-sampled $D^{(1)}$ with $D^{(2)}$.

Visually we can see the emergence of more complex structures than those obtained in just a single layer of dictionary learning. The formation of these more complex dictionary filters can

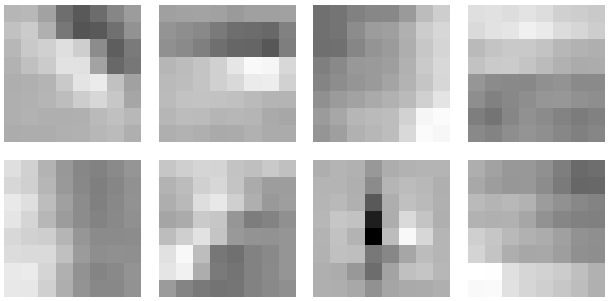


Figure 5.4 Dictionary of 8 by 8 filters trained on faces from the Extended Yale Face Database B.

assist in object recognition and detection. In particular we can see the structure of an eye in the second hierarchy of dictionary learning. This hierarchical dictionary atom can be used to highlight corresponding sections in the training images by consideration of the sparse representation. The contribution of the dictionary atom corresponding to an eye is highlighted in red in Figure 5.6. In this figure we can see that this atom plays a prominent role in the sparse representation of the eye features.

This proof of concept demonstrates that a hierarchical variation of dictionary learning can produce dictionary atoms that can be used directly to sparsely represent more complex objects. Though the demonstrated hierarchical dictionary learning consisted of just two hierarchies, additional layers of maxpooling and sparse representation can easily be appended. Each layer of maxpooling shrinks the sparse representation of the previous sparse coding step, adding slight rotational and scale flexibility. The subsequent sparse coding and dictionary learning step learns the underlying patterns in the pooled sparse representation. Each additional layer of maxpooling and sparse representation will allow for more complex objects. In the future this approach may be extended to constructing a versatile object bank.

5.2.5 Future Work

One of the difficulties of the proposed method is encountered with different scales and rotations of objects in the signals. The Extended Yale Face Database B contains similarly cropped, rotated, and scaled faces. The low scale and rotation variance nature of this database allows for reasonable representations with relatively few dictionary filters. More complex training signals require more training data as well as more filters, so computer memory quickly becomes a limitation. Additionally with more scale and rotation variance in the signals it becomes unclear what objects the hierarchical dictionary will learn.

To demonstrate the complexity that emerges with scale variance, we construct a synthetic signal comprised of different sized squares. Our synthetic image is sparsely representable with the

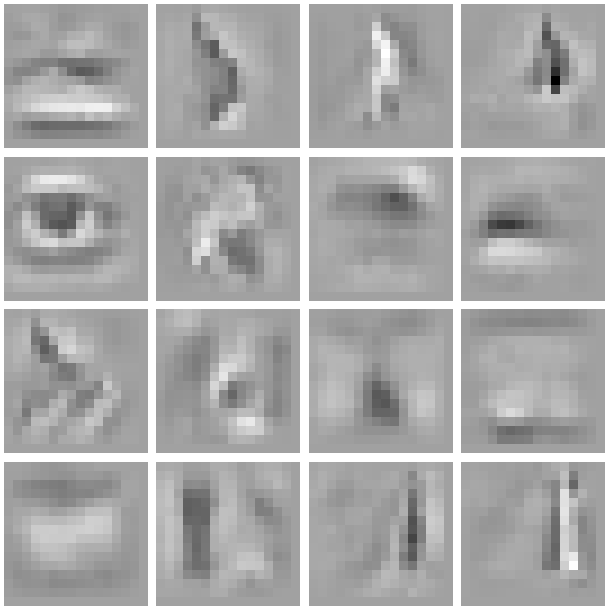


Figure 5.5 Convolution of two hierarchies of dictionaries trained on the Extended Yale Face Database B.

perimeters of two squares of different sizes, depicted in Figure 5.7. In this case one would like to learn two different filters, one filter to represent each size square. Applying the hierarchical dictionary learning to the synthetic image results in the dictionaries seen in Figure 5.8.

There are a few surprising results in the trained dictionaries. In the first hierarchy we observe that there is one filter that appears as a cross, this is surprising as the pattern never appears in the training image. In the second hierarchy we observe that the two different atoms do not represent the two different sized squares in the training image. Instead it appears that each filter represents two sides of a single square. These filters are used in conjunction to sparsely represent both size squares.

The inability of the second hierarchy to learn each square independently in such a simple noise free case shows that the interaction between different scales of the same object complicates the dictionary learning. This is because in the first hierarchy some parts of the sparse representations will be similar, while other parts of the representations will be very different. Additional layers in the hierarchy may address this problem, allowing for more and more complex features, but at the cost of computation time. Future study with hierarchical convolutional dictionary learning will aim to make this method more suitable and computationally feasible on complex databases with significant scale and rotation variance.

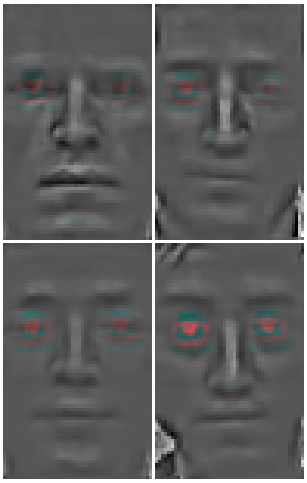


Figure 5.6 Reconstruction of the high pass training images where the contribution of the hierarchical dictionary filter corresponding to an eye is portrayed in red.

5.2.6 Conclusion

The ability to construct complex features useful for object detection in dictionary learning is of extrordinary interest for applications in computer vision. This work demonstrates the feasibility of a hierarchy of dictionaries to generate more complex features as the next step towards object detection with dictionary learning. The emergence of complex features in the hierarchy of dictionaries is a complex and useful phenomena which we hope to study and expand upon in the future.

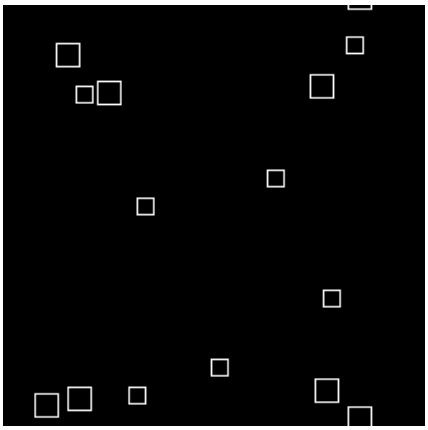
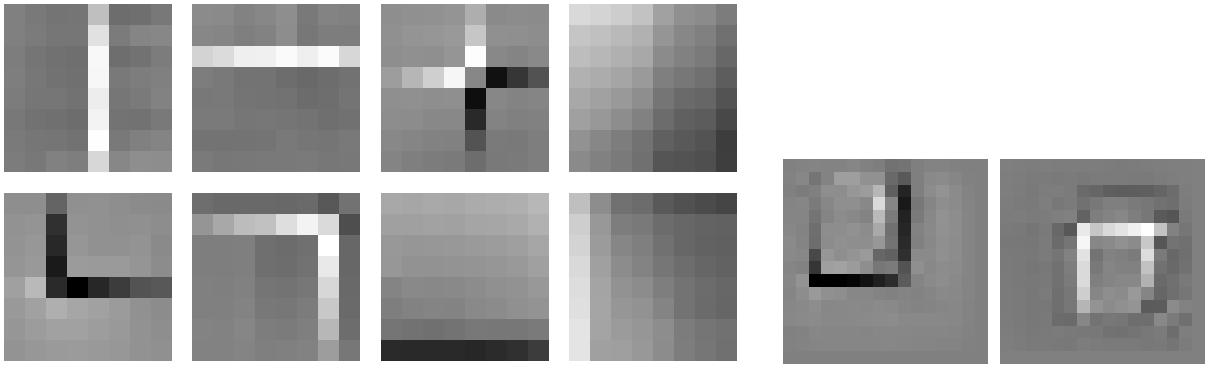


Figure 5.7 Synthetic image comprised of square perimeters.



(a) First layer of a hierarchical dictionary. (b) Convolution of two hierarchies of dictionaries.

Figure 5.8 Hierarchical dictionaries trained on a synthetic image of squares.

CHAPTER

6

FUSING HETEROGENEOUS DATA: A CASE STUDY FOR REMOTE SENSING AND SOCIAL MEDIA

6.1 Introduction

The topic of information fusion is broad and shares a diverse body of disciplines. Interest in fusion often arises in inference problems with a limited number of features obtained from a single sensor, hence unable to fully characterize the physical phenomenon at hand. Different types of fusion have been considered from lower to higher levels, e.g., sensor/data level [160], the feature level [160], and the knowledge/decision level [47].

The breadth and variety of data fusion problems necessitate different strategies to addressing different problems at hand. One of the well known data fusion frameworks [40] splits data fusion up further into five different levels, including Data In-Data Out (DAI-DAO), Data In-Feature Out (DAI-FEO), Feature In-Feature Out (FEI-FEO),Feature In-Decision Out (FEI-DEO) and Decision In-Decision Out (DEI-DEO).

The DAI-DAO problems often concentrate on combining multiple sensor readings with the objective of minimizing the uncertainty of one of those measurements. DAI-DAO level fusion is applied directly to raw sensor readings and the output should also be thought of as raw data.

There are a variety of sensor fusion techniques depending on if the sensors are complementary, redundant, or cooperative. One example of a DAI-DAO fusion problem is image super resolution [118]. Image super resolution techniques aim to use many low resolution aliases of a scene to construct a single high resolution image. One common approach to super resolution entails solving maximum likelihood optimization problems such as those in [22, 113, 134], while other works use other optimization schemes such as sparse representation [52].

The objective of the FEO levels of data fusion is to combine the input data or features to generate a desirable feature. There are a variety of objectives that may go with these levels of data fusion, depending on the specific situation. One problem in the FEO classes is the multi-modal image fusion problem [136]. This problem aims to synthesize a single clear image containing all the pertinent information for the specific application from a variety of modalities. There are numerous medical and military applications for this type of data fusion. One military application solved an entropy optimization problem to combine infrared and low light sensor data to construct a single image in [9].

Decision level fusion aims to combine multiple variables to make a new decision, or improve upon an existing decision. A common guiding principle to these data fusion problems is to maximize classification performance. An example of a DEO problem is target detection or classification [74]. These problems aim to predict the presence or absence of a specific object in a given image or region. A common approach for object classification uses parsimony or sparsity as a guiding principle, such as in the facial recognition work [157].

Our framework is closest to the DAI-FEO fusion in Dasarathy's classification. But more generally we would like to consider fusion of heterogeneous data with output for higher level decision making. Nonetheless, there is not yet a well-received principled fusion scheme requiring feature in and decision out. This is mainly because of the numerous required tasks including the necessity to homogenize heterogeneous data, to determine a suitable criterion for fusion, and to design a principled fusion scheme to target the criterion.

In our case study, we consider the fusion of drastically different data modalities: satellite images and social media data. Our aim in this particular case is to fuse such diverse modality data through a well-formulated mathematical framework in order to produce a hazard level estimation, i.e., providing guidance for decision making.

Our main contribution is a successful showcase of our information-based optimization fusion framework. In this framework constraints are given by structured feature maps from remote sensing imagery and heterogeneous data from social media.

The following sections proceed as described.

Section 6.2 is related work concentrating on the use of social media data.

Section 6.3 describes the available data for our application. We describe the satellite imagery, the available social media data, and information provided by the city of Boulder, CO. We describe

the maximum entropy model used in our case study.

Section 6.4 describes an equivalent formulation to the maximum entropy problem followed by an algorithmic discussion.

Section 6.5 discusses the outcome of various optimization instances. We discuss the necessary refinements to the social media data and compare the results to those achieved with more structured data.

Finally we conclude in Section 6.6. Here we discuss the major contributions and results of the work, as well as discuss potential future developments.

6.2 Related works

A central question to our investigation is the possibility of enhancing common data features with other modalities such as social media data. Using social media to estimate the extent of natural disasters or events is a relatively new area of research. In the past several years, several different approaches have been taken to employ social media data with varying results. Here we survey just a few of them.

Spatial-temporal analysis of tweets was used to pinpoint the movement of a forest fire that broke out in France in 2009 in [45]. They employ temporal analysis for comparison with a timeline of major events in the forest fire. Spatially, they use both textual place names, as well as user locations. They conclude that for their purpose the Twitter information was temporally quite accurate, but journalists were often a better source for developing information, especially at the beginning of the event. Their spatial study leads them to conclude that the mixture of primary and secondary information coming from Twitter users makes analysis exceedingly difficult.

Another study looking to mine tweets considers both a wildfire and flood event [150]. They do significant statistical analysis of the tweets, finding relations between geotagged information and primary tweeters or retweeters. They also look at how likely geographic information is included in the content of the tweets during the flood event and the wildfire event.

There has also been some research into using tweets to track city activities by police departments, such as that shown in [68]. They analyze the content of tweets directed to police stations and classify them into major categories. They conclude that typically information is one directional, from the police to citizens, and that rarely do citizens alert police to events of interest via Twitter. The Twitter dialog has significant unused potential in two way communication with police departments.

Another group aiming to combine social media information with other geographic information studies a flooding event in Germany [3]. They classify tweets into several flood related categories and demonstrate spatial correlations between some sets of tweets and the flooding event.

When one tries to incorporate information from social media data, the domain adaptation problem and transfer learning [116] is a subject of relevant interest. In such context, the discrepancy

between the probability density functions (PDF) of source domain and target domain poses the main challenge for learning a desirable fusion function. However, most domain adaptation methods are based upon assumptions of marginal distributions of source and target domain. These assumptions are hardly met in our case study of flood estimation dealing with the noisy, often unreliable labeling data from the social media.

6.3 Problem Statement and Formulation

Our primary objective is to propose a principled optimized framework to homogenize heterogeneous large scale data and provide a quantitative strategy to estimate or predict a scenario. In this case, the estimate pertains to floods over a geographical region. Our hazard estimation of inundated areas using multi-modal data with the maximum entropy criterion for estimating the field distribution of interest. The maximum entropy principle is a well known statistical model dating back to the 1950's [79]. The rationale of maximum entropy is to seek the most unassuming and hence maximally random mode. A more recent work related to our research utilizes maximum entropy to estimate species distribution [120].

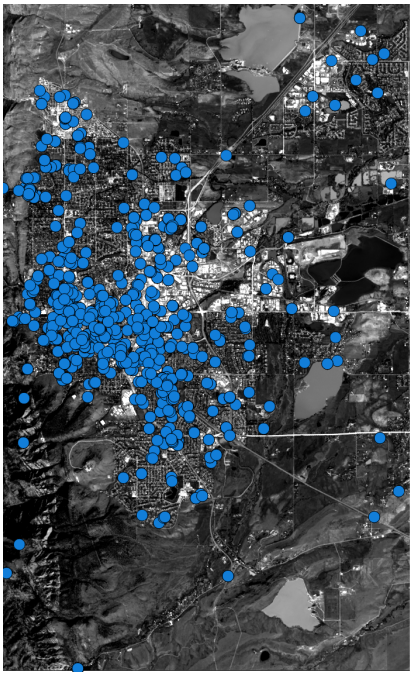
6.3.1 Weather-related Data: Boulder Colorado

This work focuses on the 2013 Boulder Colorado Flood, which was declared as a natural disaster starting in September 2013. From September 9, to September 15, some places in Boulder County saw up to 17 inches of rainfall, which is comparable to the annual average of approximately 20 inches. The flood has been attributed to at least eight deaths, with several more missing or unaccounted for. Financially, the flood has been estimated to cause $2 billion in damage.

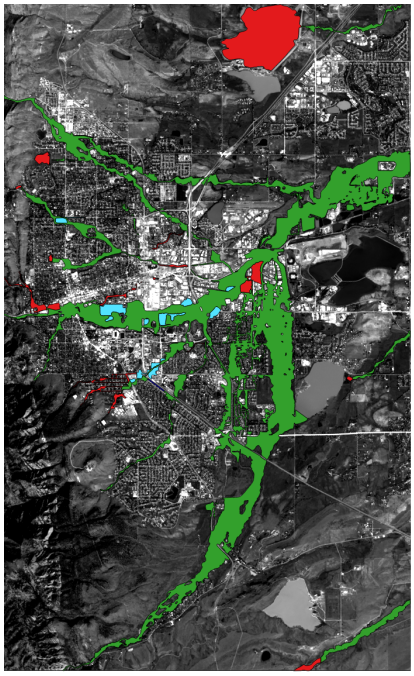
Satellite imagery is one of the principal tools to analyze many events, including floods. Landsat 8 is an American satellite, operated by NASA and the USGS, and was launched in Feburary of 2013. The satellite carries two sensors, an Operational Land Imager (OLI) and a Thermal InfraRed Sensor (TIRS). The OLI is a multispectral sensor that collects images in 9 shortwave bands. Table 6.1 shows the spectral information for the OLI sensor. Shortly after the flood, on the 17th of September, the Landsat 8 satellite passed over Boulder Colorado capturing multispectral images of the aftermath of the flood. There is also a photo of this region prior to the flood taken on Aug. 25, 2013. The images from this satellite are freely available to the public. Figure 6.1 (a) is a panchromatic image of the City of Boulder taken by Landsat on Sept 17, 2013.

Throughout the flooding event, social data was collected from the popular social media service Twitter. Twitter allows users to send and read short messages called "tweets." Often times users use hashtags to somewhat contextualize their messages. During the flood, several hashtags emerged to indicate a connection to the 2013 Boulder Flood, such as "#LongmontFlood,", "#boulderflood,"

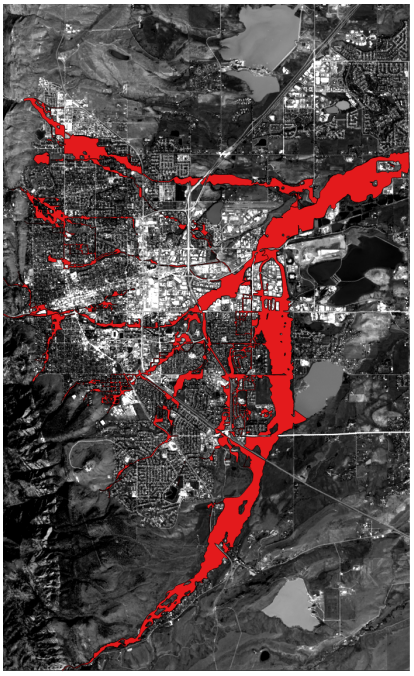
(a) A panchromatic image of the City of Boulder, CO.



(b) An overlay of Geotagged Tweets.



(c) An overlay of the SFHA zones.



(d) An overlay of the UFE area.

Figure 6.1 Data from the 2013 Boulder, CO flood.

Table 6.1 Details of the Landsat 8 OLI sensor.

Spectral Band	Wavelength	Resolution	Solar Irrandiace
Band 1 - Coastal/Aerosol	$0.433-0.453\mu m$	$30m$	$2031W/(m^2\mu m)$
Band 2 - Blue	$0.450-0.515\mu m$	$30m$	$1925W/(m^2\mu m)$
Band 3 - Green	$0.525-0.600\mu m$	$30m$	$1826W/(m^2\mu m)$
Band 4 - Red	$0.630-0.680\mu m$	$30m$	$1574W/(m^2\mu m$
Band 5 - Near Infrared	$0.845-0.885\mu m$	$30m$	$955W/(m^2\mu m)$
Band 6 - Short Wavelength Infrared	$1.560-1.660\mu m$	$30m$	$242W/(m^2\mu m)$
Band 7 - Short Wavelength Infrared	$2.100-2.300\mu m$	$30m$	$82.5W/(m^2\mu m)$
Band 8 - Panchromatic	$0.500-0.680\mu m$	$15m$	$1739W/(m^2\mu m)$
Band 9 - Cirrus	$1.360-1.390\mu m$	$30m$	$361W/(m^2\mu m)$

and "#coflood." In addition, depending on the users preferences, several tweets were geotagged with the metadata of the latitude and longitude of the phone at the time of posting. In Figure 6.1 (b) we can visualize the distribution of people tweeting about the Boulder flood in a region containing Boulder City and some of its surroundings.

Special Flood Hazard Areas (SFHAs) are flood hazard areas identified on Flood Insurance Rate Maps (FIRMs). SFHA regions have a probability of being flooded greater than 1% each year. These areas can be expected to flood at least once every 100 years. SFHAs consist of several different types of zones, some of which are: Zone A, Zone AE, Zone AH, and AO. These different zones correspond to different methods of calculation, or different expected types of flooding events. For instance, Zone A regions are generally determined using approximate methodologies, while Zone AE regions are determined with detailed methodologies. AH Zones correspond to areas expected to undergo shallow pooling with depths less than three feet, while AO Zones correspond to regions expected to undergo flood sheets on an incline with depths less than three feet. The FIRMs for Boulder Colorodo as they were defined prior to the flood, are also publicly available and can be seen in Figure 6.1 (c).

In the weeks and months following the flood, the city of Boulder created a map of inundated areas to better understand the flood as well as to potentially revise the SFHAs. With the help of hand held GPS devices, workers geotagged high water locations and carefully produced an accurate account of the inundated areas. With additional information from community provided photo evidence the extent of the flooding in some regions was also obtained. This information was conglomerated to construct a Urban Flood Extent map (UFE). The UFE was graciously provided to us by the City of Boulder. The UFE provides a semblance of ground truth, as it is a reasonable approximation of the inundated areas. The UFE may be seen in Figure 6.1 (d).

6.3.2 Problem Formulation

We would like to consider the fusion of the following feature maps $\{f_i\}_{i=1,\dots,K}$ with non-homogenized domains and values:

$$f_i : D_i \to \mathbb{V}_i \,,$$

where D_i is the domain associated with feature map f_i and $\mathbb{V}_i$ can be feature values in $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{Z}_p$, etc.

Given a fusion domain D_f, and transformations $\phi_i : D_f \to D_i$, our fusion function is then defined implicitly as an optimizer of an information-based functional with constraints given by

$$F(f_1(\phi_1(x)), \dots, f_K(\phi_K(x))) \in C(x) \,,$$

where $C(x)$ is a constraint set associated with each x in the fusion domain D_f.

In our case study of flood estimation, the social media data is fused as labels of geo-locations of flood, namely, a feature map with values in $\{0,1\}$ defined on a subset S of the fusion domain $D_f \in \mathbb{R}^2$.

The first step to formulating an optimization problem that can use different data modalities is mapping data from all of the modalities onto a common space. We pick a spatial domain D containing the region of interest. We segment D into a set of cells in a grid, X, where each $x \in X$ is approximately $30m$ by $30m$. Our objective is to map each data source to a function defined on X, and represent the information from the modality.

Using the social media data, we define an empirical distribution to represent the tweets on X. The empirical distribution is defined as:

$$\begin{aligned} &\tilde{p} : X \to [0,1] \\ &p(x) = \frac{|\text{observations originating from x}|}{|\text{all observations}|} \cdot \end{aligned}$$

The defined empirical distribution is a proper probability density function since it integrates to one, and is nonnegative. Since the observations were a sampling of where people tweet about the flood, the empirical distribution can be thought of as an approximation of this distribution. In fact, as the number of observations tend to infinity, the empirical distribution should approach the parent distribution.

To represent the OLI Landsat images as a function defined on X, we chose the very natural method of resampling each band on the grid X. For each band provided by the OLI sensor, we resample the image on our grid to naturally obtain a function $f_i : X \to \mathbb{R}$, where i is an index for different bands.

Both the SFHA and UFE data can similarly be mapped to a function on X. For each of these

regions indexed by the modalities we define the indicator function,

$$g_i : X \to \{0,1\}$$
$$x = \begin{cases} 1 & \text{if } x \cap \text{region}_i \neq \emptyset \\ 0 & \text{if } x \cap \text{region}_i = \emptyset. \end{cases}$$

These functions can then also be normalized so that they can be viewed as probability density functions on X.

6.3.2.1 Maximum Entropy Model

With the above defined, our maximum entropy formulation may be written as,

$$\begin{aligned} \underset{q(x)}{\arg\max} \quad & -\sum_{x \in X} q(x) \ln\big(q(x)\big) \\ \text{subject to} \quad & \begin{cases} \mathrm{E}_{q(x)}\big[f_i(x)\big] = \mathrm{E}_{\hat{p}(x)}\big[f_i(x)\big],\ i = 1, \ldots, m \\ q(x) \geq 0 \quad \forall x \in X \\ \sum_{x \in X} q(x) = 1 \end{cases} \end{aligned}$$

where $E_{p(x)}[p(x)] = \sum_{x \in X} p(x) f(x)$ is the expected value of f under the distribution p. This formulation amounts to seeking the maximum entropy distribution such that the expected values of our environmental variables $f_i(x)$ under our estimation match the expected values of the environmental variables under an empirical distribution.

As widely noted [120], such strong constraints may not be reasonable, and should be replaced with relaxed constraints enforcing that the expected values be close. To that end, we relax the above formulations to,

$$\begin{aligned} \underset{q(x)}{\arg\max} \quad & -\sum_{x \in X} q(x) \ln\big(q(x)\big) \\ \text{subject to} \quad & \begin{cases} \left|\mathrm{E}_{q(x)}\big[f_i(x)\big] - \mathrm{E}_{\hat{p}(x)}\big[f_i(x)\big]\right| \leq \beta_i,\ i = 1, \ldots, m, \\ q(x) \geq 0 \quad \forall x \in X, \\ \sum_{x \in X} q(x) = 1, \end{cases} \end{aligned} \tag{6.1}$$

with some $\beta_i \geq 0$ for $1 \leq i \leq m$. This relaxation allows the expected values to vary inside of a bounding box centered at the expected values under the empirical distribution.

6.4 Algorithmic Fusion Solution

6.4.1 A Maximum Likelihood Equivalence

An equivalent formulation of the maximum entropy model in (6.1), can be shown to be a maximum likelihood model with a Gibbs distribution [46]. This formulation can be written as

$$\begin{aligned} \underset{q(x),\lambda_i}{\arg\min} \quad & \hat{p}(x)\ln\left(\frac{\hat{p}(x)}{q(x)}\right)+\sum_{i=1}^{m}\beta_i|\lambda_i| \\ \text{subject to} \quad & q(x)=\frac{e^{\sum_{i=1}^{m}\lambda_i f_i(x)}}{\sum_{z\in X} e^{\sum_{i=1}^{m}\lambda_i f_i(z)}}. \end{aligned} \tag{6.2}$$

This formulation aims to find the maximum likelihood Gibbs distribution that minimizes the relative entropy to the empirical distribution. From (6.2), it is evident that the optimal solution takes the form of a normalized function on X of an affine combination of the environmental features. There are several options available to solve this constrained convex problem. The advantages and drawbacks of several different methods are discussed in [108]. For the sake of simplicity in algorithm design, we follow the work of [120] and use an iterative scaling algorithm to solve this optimization problem. We next describe the procedure used in solving the maximum entropy problem.

We concentrate on solving the maximum entropy problem,

$$\begin{aligned} \underset{x\in P}{\arg\max} \quad & -x^T\log(x/x_0) \\ \text{subject to} \quad & Ax=Ay \end{aligned}$$

by using its equivalent maximum likelihood formulation,

$$\begin{aligned} \underset{x,\lambda}{\arg\max} \quad & -y^T\log(x) \\ \text{subject to} \quad & x=\frac{x_0 e^{A\lambda}}{\|x_0\odot e^{A\lambda}\|_1} \end{aligned}$$

which can be condensed into

$$\underset{\lambda}{\arg\max} \quad \log(x_0^T e^{A\lambda})-y^T\log(x_0\odot e^{A\lambda}).$$

To proceed, we take a coordinate descent approach on λ by maximizing the difference in change of the objective function. If $Q(\lambda)$ is the objective value then we calculate

$$
\begin{aligned}
Q(\lambda+\delta)-Q(\lambda) &= \log(x_0^T e^{A(\lambda+\delta)}) - y^T \log(x_0 e^{A(\lambda+\delta)}) - \log(x_0^T e^{A\lambda}) - y^T \log(x_0 e^{A\lambda}) \\
&= -y^T A\delta + \log(\frac{x_0^T e^{A(\lambda+\delta)}}{x_0^T e^{A\lambda}} \\
&= -y^T A\delta + \log(\frac{x_0^T e^{A\lambda} e^{A\delta})}{x_0^T e^{A\lambda}}) \\
&\leq -y^T A\delta + \log(\frac{(x_0 \odot e^{A\lambda})^T (1+(e^{\delta_j - 1})Ae_j)}{x_0^T e^{A\lambda}})
\end{aligned}
$$

The last inequality results from Jenson's inequality on e^x with $x \in [0,1]$. The update step is obtained by differentiating $-y^T A\delta + \log(\frac{(x_0 \odot e^{A\lambda})^T (1+(e^{\delta_j - 1})Ae_j)}{x_0^T e^{A\lambda}})$ with respect to δ and solving for the critical point. When working with the bounding box regularization, there are several critical points. The critical point that maximized the objective value corresponds to the update step.

6.5 Results and Discussion

6.5.1 Fusing Unstructured Data: Tweets

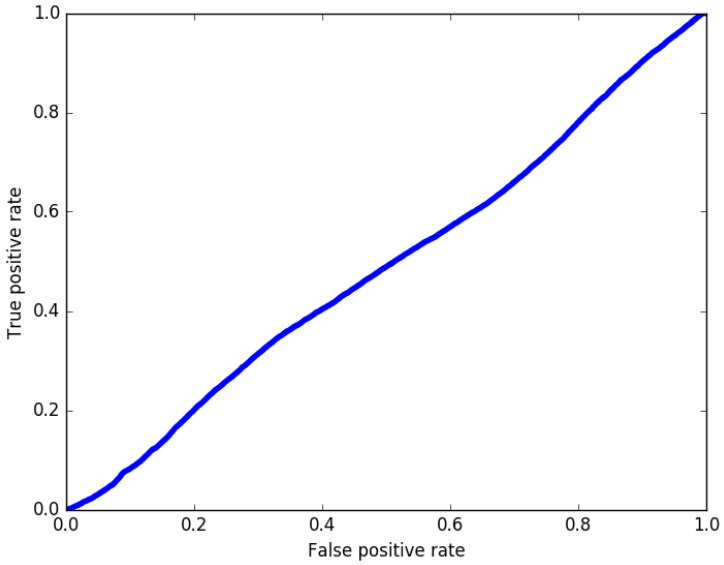
Using all post-flood landsat bands as environmental variables and the Twitter empirical distribution in the maximum entropy model produces the estimation seen in Figure 6.2 (a). In the resulting estimation, the regions depicted with lighter cells correspond to higher estimated likelihoods. Unfortunately, visual inspection of the estimation suggest that the produced feature is not representative of the UFE depicted in Figure 6.1 (d). Instead, it appears that the feature is representative of man made infrastructure.

Using the UFE as the ground truth, we obtain a performance evaluation of our estimation by constructing a Receiver Operator Characteristic (ROC) curve. As may be seen in Figure 6.2 (b), this rudimentary estimation performance is no better than a random guess.

6.5.2 Unstructured Data: Manual Classification

One hypothesis as to why the rudimentary estimation is less than optimal is the positional uncertainty of the collected tweets. To explore this possibility, we undertook manual classification of the tweets as indicative of flooded or not flooded in the immediate vicinity of the GPS coordinates. Approximately 25% of the tweets were classified as indicative of flooding close their GPS coordinates, suggesting that 75% of the tweets have a position uncertainty significantly greater than the 30m resolution of the satellite imagery. These tweets may contain useful information, but should not be attributed to the cell of their GPS coordinates as their positional uncertainty is too large. The subset of tweets manually classified as indicating flooding in the immediate vicinity is depicted in Figure

(a) Maximum entropy estimation.



(b) ROC curve analysis

Figure 6.2 The results of a maximum entropy estimation using the tweet empirical distribution and the landsat photos taken after the flood.

6.3.

Using the Landsat features taken after the flood in conjunction with an empirical distribution generated from the subset of tweets produces the estimation seen in Figure 6.4 (a). From the corresponding ROC curve in Figure 6.4 (b), we conclude that estimation using the manually classified tweets is better than random. The manual classification of the tweets resolves the positional uncertainty, not by correcting the location of the tweets, but by eliminating positionally uncertain tweets from the estimation. This result indicates that some information is contained in the social media data, but additional work is needed to extract it.

6.5.3 Unstructured Data: Location Correction

To improve the geo-location information from the social media data, we make positional corrections with the use of expert information. The manual classification of the tweets into two groups revealed that for up to 75% of observations, the tweeter may have been indirectly affected by the flooding, but are not necessarily at a flooded location. These social media observations should not be applied to the cell containing the observation, but should be attributed to nearby pixels that are likely to be flooded. Even the observations indicating flooding in the immediate vicinity may need a positional

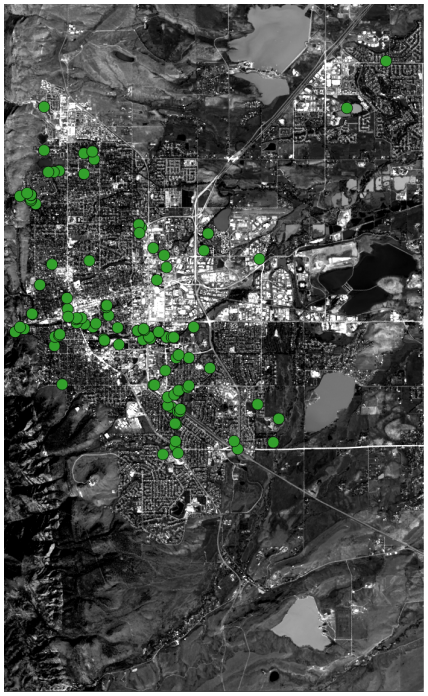


Figure 6.3 An overlay of tweets classified as indicative of flooding.

correction, as being flooded at those precise GPS coordinates would require the observer standing in or above water.

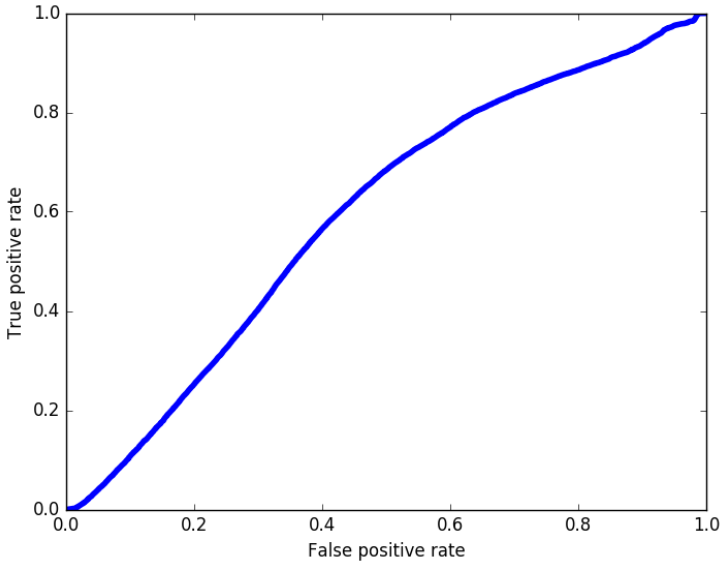
To address this issue we attribute each social media observation located to the closest cell likely to be flooded using prior information. For our choice of prior we use the Modified Normalized Difference Water Index (MNDWI). The Normalized Difference Water Index (NDWI) is a method to estimate the presence of water from multispectral imagery [60]. The NDWI was then modified a decade later to the MNDWI [159]. The calculation for the MNDWI relies on two landsat bands, the green band, and the Short Wavelength InfraRed (SWIR) band. The equation for the MNDWI is a simple normalized difference,

$$\begin{aligned} f_{MNDWI}(x) &= \frac{f_{green}(x) - f_{SWIR}(x)}{f_{green}(x) + f_{SWIR}(x)} \\ &= \frac{f_{\text{Band 3}}(x) - f_{\text{Band 6}}(x)}{f_{\text{Band 3}}(x) + f_{\text{Band 6}}(x)}. \end{aligned}$$

The resulting value gives a water index, or some measure of how much water was contained in that pixel. Figure 6.5 depicts the MNDWI calculated from the Boulder, CO data.

With the available satellite imagery, the use of the MNDWI to estimate the inundated areas directly will be undesirable as the flooding took place several days prior to the satellite passing. The

(a) Maximum entropy estimation.



(b) ROC curve analysis

Figure 6.4 The results of a maximum entropy estimation using the manually classified tweets and the landsat photos taken after the flood.

use of the MNDWI directly may also be problematic as the spectral properties of shadows from mountains, buildings, and clouds may cause missclassification [56, 148], but it provides a reasonable prior that will be correlated with the inundated regions.

We use the MNDWI to attribute each tweet, originating in cell x, to the closest cell with a MNDWI above a threshold. The MNDWI has the advantage of having a intuitive and physical meaning for choosing the threshold as discussed in [159]. From this physical interpretation we chose a value of 0.0 to threshold the MNDWI into regions likely to have flooded, and not likely to have flooded and solved the optimization problem,

$$\begin{aligned} \underset{y \in X}{\arg\min} \quad & \operatorname{dist}(x, y) \\ \text{subject to} \quad & \operatorname{MNDWI}(y) \geq 0.0. \end{aligned}$$

This relocation method uses the MNDWI as a prior of the flooded region to address the positional uncertainty of the tweets. Figure 6.6 (a) depicts the resulting estimation. The resulting ROC curve of this estimation can be seen in Figure 6.6 (b) in conjunction with an estimation using a normalized MNDWI as the guiding distribution for a comparative control.

Figure 6.5 The calculated MNDWI for Boulder, CO.

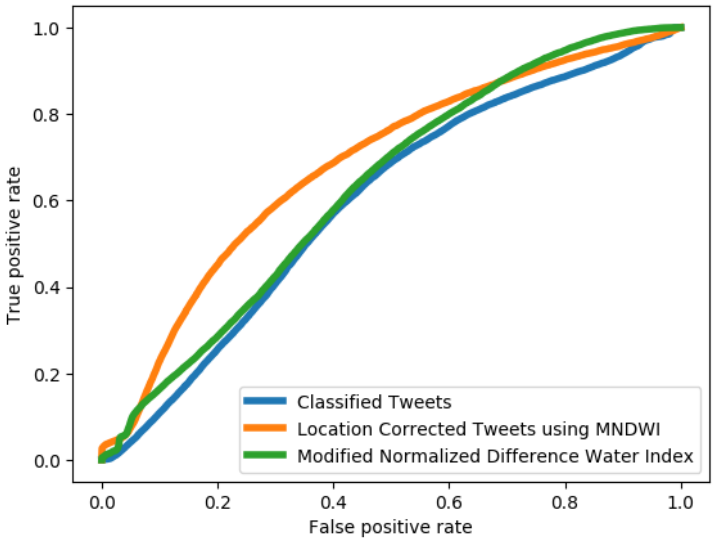
We see that this method of attributing the tweets to nearby likely regions of flood drastically improves the quality of the social media data. Importantly, this method outperforms the prior distribution based on the MNDWI directly, indicating that the social media data is providing valuable input into the maximum entropy model.

6.5.4 Structured Data: SFHA

As a control against using social media data, we also consider using the SFHA instead of an empirical distribution. The SFHA contains expert information of regions with high chances of flooding. There are a variety of methods used to calculate the SFHA including the use of digital elevation models (DEM) and computer simulations. The SFHA acts as a normalized indicator function returning different values depending on if the cell is in a flood zone or not. By using the SFHA instead of an empirical distribution, we get a comparison of how informative social media information is to structured data. Figure 6.7 (a) depicts the estimation when using the SFHA to supply our optimization constraints with all the post flood Landsat bands.

By considering the ROC curve of the resulting estimation in Figure 6.7 (b) we conclude that the estimation using SFHA is comparable to that using the manually classified or the location corrected estimations.

(a) Maximum entropy estimation.



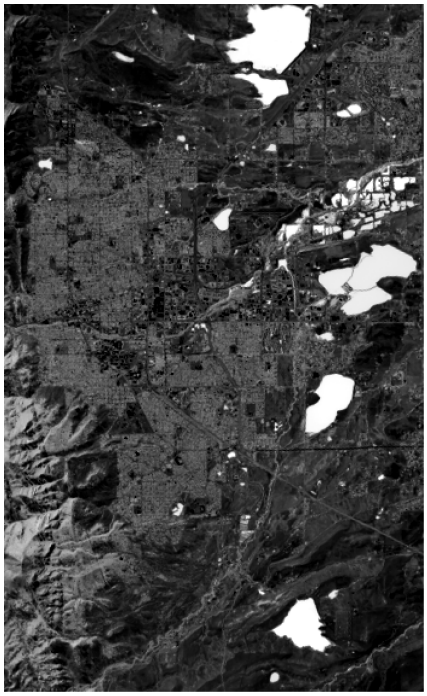
(b) ROC curve analysis

Figure 6.6 The results of a maximum entropy estimation using the nearest neighbors relocated tweets and the landsat photos taken after the flood.

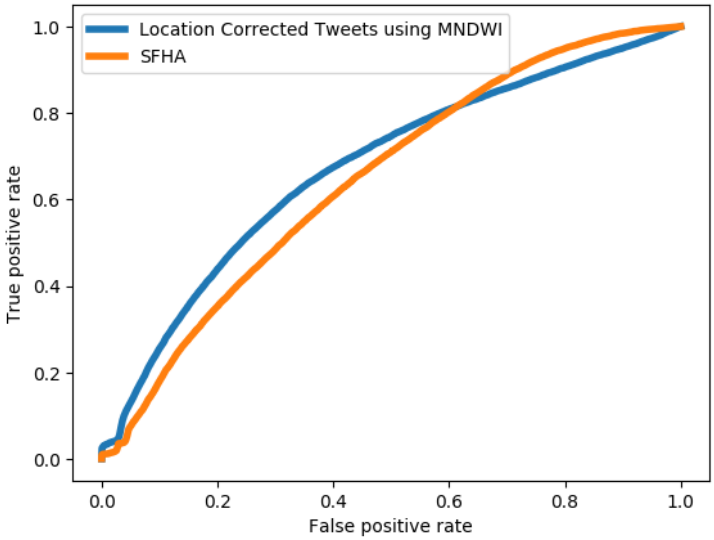
6.5.5 Feature Generation

Generating good and efficient features is central to any estimation problem. One way to potentially improve the method is by producing new features. Every new feature included in the problem introduces a new degree of freedom when solving the maximum likelihood formulation.

One idea to generate features is to apply dictionary learning in the spectral domain. This approach learns a set of atoms that sparsely represent the spectral signatures of different types of materials, as well as representations for each pixel in this new basis. In our experimental validation, we learned a complete dictionary and used the sparse representations as our new input features. Figure 6.8 depicts some of the sparse representations obtained after dictionary learning. Visually the dictionary seems to work well at segmenting the image into regions of different materials. We can perceive the roads in some bands, and potentially inundated regions in others. This also seems to yield a slightly improved estimate to that of using the features that the dictionary was learned on as may be seen in Figure 6.9.



(a) Maximum entropy estimation.



(b) ROC curve analysis

Figure 6.7 The results of a maximum entropy estimation using the SFHA and the landsat photos taken after the flood.

6.5.6 Robustness to Malicious Information

Another issue we can explore with this data is the sensitivity of the model to the injection of malicious information. We can evaluate the robustness of this model to the injection of malicious information by using the two classified sets of tweets as true signal and malicious signal. Assume that the tweets indicating flooding in the immediate vicinity comprise the true signal and come from a distribution p_1. The remaining tweets, p_2 can be injected into the signal with different quantities to study their effect on the estimation. We model the injection of malicious information with a mixture distribution. By considering a mixture distribution we can analyze the impact of such an attack on the maximum entropy formulation.

Suppose our sampling comes from a convex combination of the two different distributions, $p = \alpha p_1 + (1-\alpha)p_2$. Given that the expected values of our features are defined under the distributions p_1 and p_2, we can then consider the resulting constraints $E_q[f_i] = E_p[f_i]$.

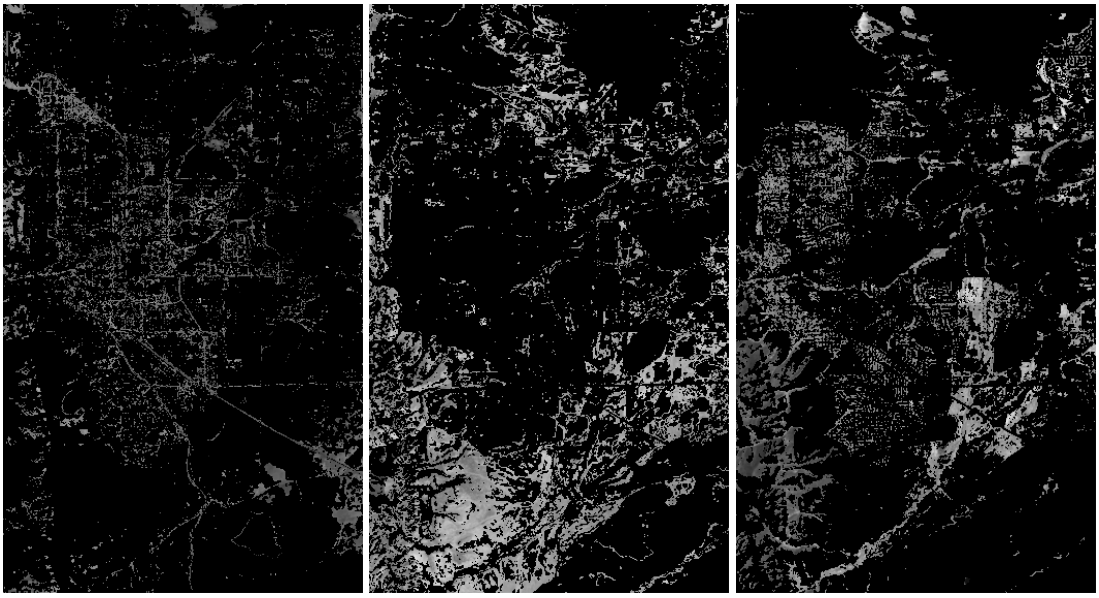


Figure 6.8 The sparse representations after dictionary learning.

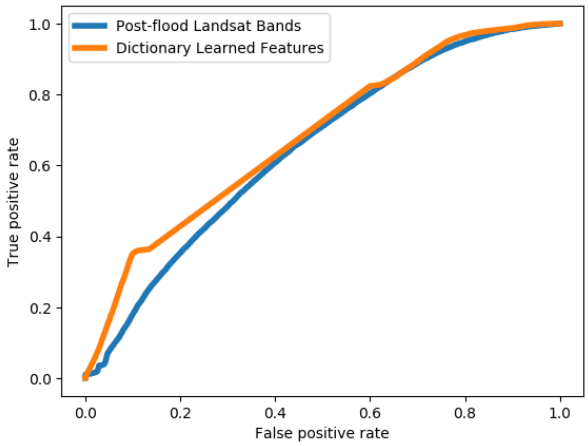


Figure 6.9 ROC curve comparison of dictionary learned estimation and the estimation from the features the dictionary was trained on.

$$
\begin{aligned}
E_q[f_i] &= E_p[f_i] \\
&= \sum_{x \in X} f_i(x) p(x) \\
&= \sum_{x \in X} f_i(x)(\alpha p_1(x) + (1-\alpha) p_2(x)) \\
&= \alpha \sum_{x \in X} f_i(x) p_1(x) + (1-\alpha) \sum_{x \in X} f_i(x) p_2(x) \\
&= \alpha E_{p_1}[f_i] + (1-\alpha) E_{p_2}[f_i]
\end{aligned}
$$

This indicates that the new expected values are a convex combination of the old expected values with the same weights as that of the mixture. Varying the mixture parameter where p_1 is the distribution of tweets indicating flood and p_2 are the remaining tweets, we get the estimations shown in Figure 6.10. The ROC curves from the estimations behave as expected, varying from our ideal estimation to malicious information as the mixture parameter varies as seen in Figure 6.11.

6.6 Conclusion

Fusing information from social media and satellite imagery for hazard extent estimation is an interesting problem with many approaches. We demonstrate that maximum entropy is one feasible approach. Unfortunately additional filtering of any social media data is required for it to be a suitable information source. Filtered social media information is not as good as structured expert information.

A mixture model is the ideal tool to consider the effect of malicious information being injected into the data. By studying the mixture of two different distributions we can see how stable our model is to attack. Expanding upon these results to make the model more resistant to attack will be pursued in the future.

There are several potential extensions to this model that can be considered for future work. Feature generation has shown some promise and there are several other feature generation approaches that one could pursue. How to incorporate any non-linearity of other methods to add additional freedom into the problem is of interest for future work.

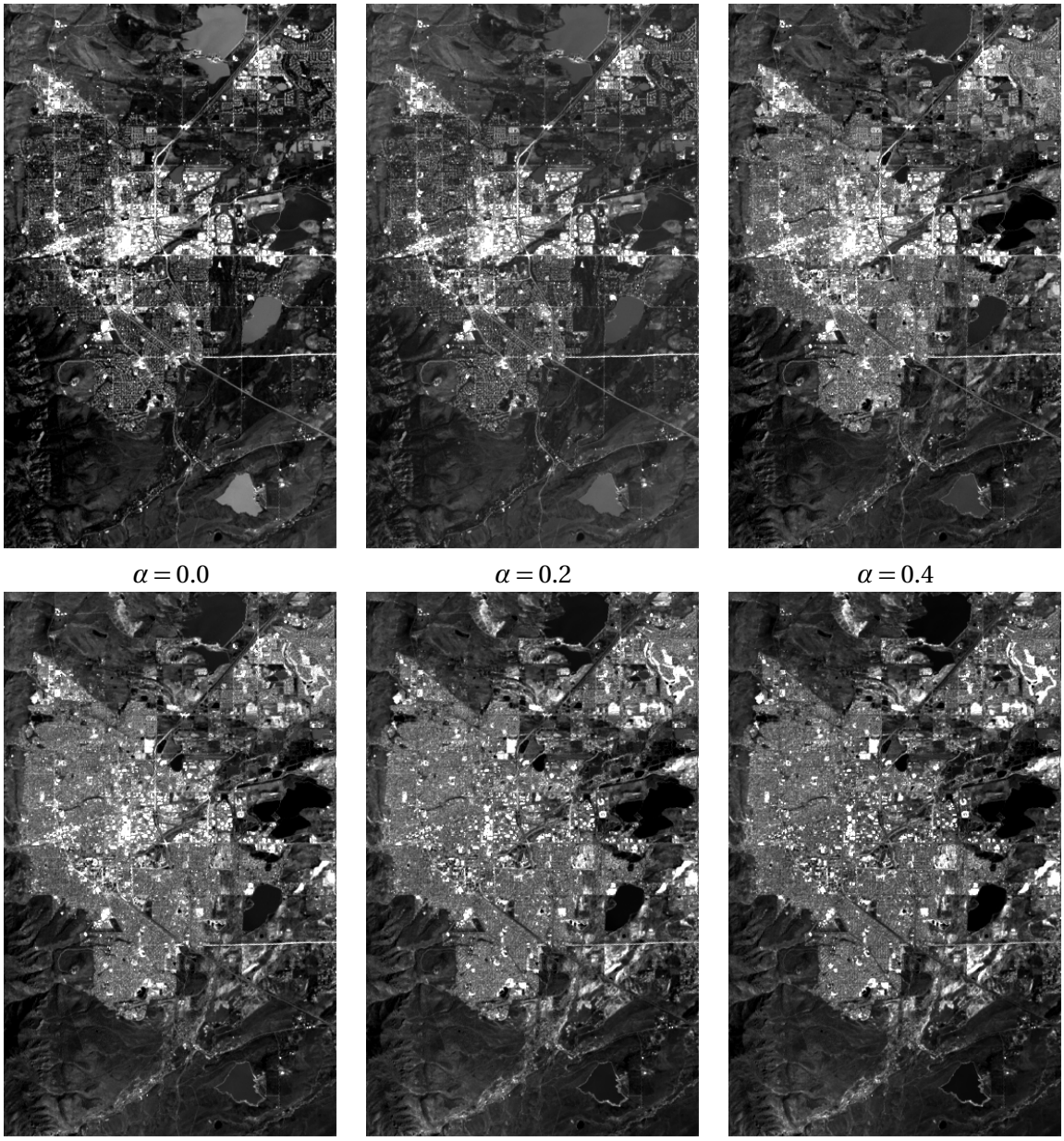


Figure 6.10 Maximum entropy estimation with mixture distributions

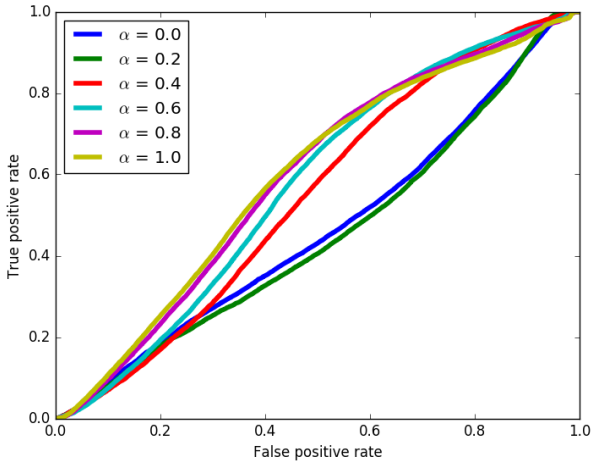


Figure 6.11 ROC curves comparing different mixture levels of the empirical tweet and the SFHA distributions.

CHAPTER

7

CONCLUSION

In this dissertation we explore the relaxations of sparse optimization problems in a variety of applications. Relaxation techniques ensure computational feasibility to approximate solutions with parsimonious characteristics in optimization problems.

We study and use these techniques in a variety of ways. We pose the Gramian tensor decomposition problem as a rank minimization problem and study the corresponding relaxation. Specifically of interest are cases where the optimal solutions of the relaxations will generically be optimal in the original problems, or will never be optimal in the original.

Natural image patches lie on a complex manifold. We use sparse approximation techniques to represent this manifold and study the adherence of real data to the sparse representation model. Through a geometric and statistical analysis of the representations we propose a modification that allows more diverse application of the sparse representation model.

Variations of dictionary learning can be applied to a variety of tasks besides noise reduction and image restoration. Triple factorization coupled dictionary learning was demonstrated to work well for pansharpening. Hierarchical convolutional dictionary learning is one approach to use dictionary learning to generate an object bank that can be used for object detection.

Data fusion is a rapidly growing field with a variety of applications. We explore the feasibility of using social media data with satellite imagery to estimate the extent of a natural disaster in a case study. Here parsimony plays a crucial role in avoiding over-fitting of data and allowing for positive results.

Future research will certainly apply the concept of parsimony in a variety of applications, as this underlying concept is fundamental to so many applications.

BIBLIOGRAPHY

[1] Aharon, M. et al. “K -SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation”. *Signal Processing, IEEE Transactions on* **54**.11 (2006), pp. 4311–4322.

[2] Albera, L. et al. “Blind identification of overcomplete mixtures of sources (BIOME)”. *Linear Algebra Appl.* **391** (2004), pp. 3–30.

[3] Albuquerque, J. P. de et al. “A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management”. *International Journal of Geographical Information Science* **29**.4 (2015), pp. 667–689.

[4] Alexander, J. & Hirschowitz, A. “Polynomial interpolation in several variables”. *J. Algebraic Geom.* **4**.2 (1995), pp. 201–222.

[5] Alparone, L. et al. “Comparison of Pansharpening Algorithms: Outcome of the 2006 GRS-S Data-Fusion Contest”. *Geoscience and Remote Sensing, IEEE Transactions on* **45**.10 (2007), pp. 3012–3021.

[6] Appellof, C. J. & Davidson, E. R. “Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents”. *Anal. Chem.* **53**.13 (1981), pp. 2053–2056.

[7] Bardet, M. et al. “On the complexity of Gröbner basis computation of semi-regular overdetermined algebraic equations”. *Proceedings of the International Conference on Polynomial System Solving*. ICPSS’04. Paris, France, 2004, pp. 71–75.

[8] Bardet, M. et al. “On the complexity of solving quadratic Boolean systems”. *J. Complexity* **29**.1 (2013), pp. 53–75.

[9] Ben Hamza, A et al. “A multiscale approach to pixel-level image fusion”. *Integrated Computer-Aided Engineering* **12**.2 (2005), pp. 135–146.

[10] Bernardi, A. et al. “General Tensor Decomposition, Moment Matrices and Applications”. preprint, arXiv:1103.0203. 2011.

[11] Bernardi, A. et al. “Multihomogeneous polynomial decomposition using moment matrices”. *ISSAC 2011—Proceedings of the 36th International Symposium on Symbolic and Algebraic Computation*. ACM, New York, 2011, pp. 35–42.

[12] Bernardi, A. et al. “A comparison of different notions of ranks of symmetric tensors”. *Linear Algebra Appl.* **460** (2014), pp. 205–230.

[13] Blekherman, G. “Nonnegative polynomials and sums of squares”. *Journal of the American Mathematical Society* **25**.3 (2012), pp. 617–635.

[14] Bovik, A. C. *Handbook of image and video processing*. Academic press, 2010.

[15] Boyd, S. et al. “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers”. *Found. Trends Mach. Learn.* **3**.1 (2011), pp. 1–122.

[16] Boyd, S. & Vandenberghe, L. “Semidefinite programming relaxations of non-convex problems in control and combinatorial optimization”. *Communications, Computation, Control, and Signal Processing*. Springer, 1997, pp. 279–287.

[17] Boyd, S. et al. “Distributed optimization and statistical learning via the alternating direction method of multipliers”. *Foundations and Trends® in Machine Learning* **3**.1 (2011), pp. 1–122.

[18] Brachat, J. et al. “Symmetric tensor decomposition”. *Linear Algebra Appl.* **433**.11-12 (2010), pp. 1851–1872.

[19] Candès, E. J. & Recht, B. “Exact matrix completion via convex optimization”. *Found. Comput. Math.* **9**.6 (2009), pp. 717–772.

[20] Candès, E. J. & Tao, T. “The power of convex relaxation: near-optimal matrix completion”. *IEEE Trans. Inform. Theory* **56**.5 (2010), pp. 2053–2080.

[21] Cao, T. et al. “Robust multimodal dictionary learning”. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2013, pp. 259–266.

[22] Capel, D. & Zisserman, A. “Super-resolution enhancement of text image sequences”. *Pattern Recognition, 2000. Proceedings. 15th International Conference on*. Vol. 1. IEEE. 2000, pp. 600–605.

[23] Carlsson, G. et al. “On the local behavior of spaces of natural images”. *International journal of computer vision* **76**.1 (2008), pp. 1–12.

[24] Carroll, J. D. & Chang, J.-J. “Analysis of individual differences in multidimensional scaling via an n-way generalization of “Eckart-Young” decomposition”. *Journal Psychometrika* **35**.3 (1970), pp. 283–319.

[25] Carroll, J. D. et al. “CANDELINC: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters”. *Journal of Psychometrika* **45**.1 (1980), pp. 3–24.

[26] Chaumette, E. et al. “ICA-based technique for radiating sources estimation: application to airport surveillance”. *Radar and Signal Processing, IEE Proceedings F* **140**.6 (1993), pp. 395–401.

[27] Chavez, P. & Kwarteng, A. “Extracting Spectral Contrast in Landsat Thematic Mapper Image Data Using Selective Principal Component Analysis” (1989).

[28] Chen, S. et al. “Pan-Sharpening via Coupled Unitary Dictionary Learning”. English. *Image and Graphics*. Ed. by Zhang, Y.-J. Vol. 9219. Lecture Notes in Computer Science. Springer International Publishing, 2015, pp. 1–10.

[29] Chen, S. S. et al. “Atomic decomposition by basis pursuit”. *SIAM review* **43**.1 (2001), pp. 129–159.

[30] Chiantini, L. & Ciliberto, C. “On the concept of k-secant order of a variety”. *J. London Math. Soc. (2)* **73**.2 (2006), pp. 436–454.

[31] Comon, P. & Mourrain, B. “Decomposition of quantics in sums of powers of linear forms”. *Signal Processing* **53**.2-3 (1996), pp. 93–107.

[32] Comon, P. et al. “Symmetric tensors and symmetric tensor rank”. *SIAM Journal on Matrix Analysis and Applications* **30**.3 (2008), pp. 1254–1279.

[33] Comon, P. “Independent Component Analysis”. *Higher Order Statistics*. Ed. by Lacoume, J.-L. Amsterdam: Elsevier, 1992, pp. 29–38.

[34] Comon, P. & Ottaviani, G. “On the typical rank of real binary forms”. *Linear Multilinear Algebra* **60**.6 (2012), pp. 657–667.

[35] Curto, R. E. & Fialkow, L. A. “Solution of the truncated complex moment problem for flat data”. *Mem. Amer. Math. Soc.* **119**.568 (1996), pp. x+52.

[36] Curto, R. E. & Fialkow, L. A. “Flat extensions of positive moment matrices: recursively generated relations”. *Mem. Amer. Math. Soc.* **136**.648 (1998), pp. x+56.

[37] Curto, R. E. & Fialkow, L. A. “Truncated K-moment problems in several variables”. *J. Operator Theory* **54**.1 (2005), pp. 189–226.

[38] Dabov, K. et al. “Image denoising by sparse 3-D transform-domain collaborative filtering”. *IEEE Transactions on image processing* **16**.8 (2007), pp. 2080–2095.

[39] Dalal, N. & Triggs, B. “Histograms of oriented gradients for human detection”. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE. 2005, pp. 886–893.

[40] Dasarathy, B. V. “Sensor fusion potential exploitation-innovative architectures and illustrative applications”. *Proceedings of the IEEE* **85**.1 (1997), pp. 24–38.

[41] De Lathauwer, L. et al. “Fetal Electrocardiogram Extraction by Blind Source Subspace Separation ”. *IEEE Transactions on Biomedical Engineering BME* **47**.5 (2000), pp. 567–572.

[42] De Lathauwer, L. “Signal Processing based on Multilinear Algebra”. PhD thesis. Katholieke Universiteit, Leuven, 1997.

[43] De Lathauwer, L. “A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization”. *SIAM J. Matrix Anal. Appl.* **28**.3 (2006), 642–666 (electronic).

[44] De Lathauwer, L. et al. “Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition”. *SIAM J. Matrix Anal. Appl.* **26**.2 (2004/05), 295–327 (electronic).

[45] De Longueville, B. et al. “"OMG, from Here, I Can See the Flames!": A Use Case of Mining Location Based Social Networks to Acquire Spatio-temporal Data on Forest Fires”. *Proceedings of the 2009 International Workshop on Location Based Social Networks*. LBSN ’09. New York, NY, USA: ACM, 2009, pp. 73–80.

[46] Della Pietra, S. et al. “Inducing features of random fields”. *IEEE transactions on pattern analysis and machine intelligence* **19**.4 (1997), pp. 380–393.

[47] Dong, X. L. et al. “From data fusion to knowledge fusion”. *Proceedings of the VLDB Endowment* **7**.10 (2014), pp. 881–892.

[48] Donoho, D. L. & Logan, B. F. “Signal recovery and the large sieve”. *SIAM Journal on Applied Mathematics* **52**.2 (1992), pp. 577–591.

[49] Donoho, D. L. & Stark, P. B. “Uncertainty principles and signal recovery”. *SIAM Journal on Applied Mathematics* **49**.3 (1989), pp. 906–931.

[50] Elad, M. & Aharon, M. “Image denoising via sparse and redundant representations over learned dictionaries”. *IEEE Transactions on Image processing* **15**.12 (2006), pp. 3736–3745.

[51] Elad, M. & Bruckstein, A. M. “A generalized uncertainty principle and sparse representation in pairs of bases”. *IEEE Transactions on Information Theory* **48**.9 (2002), pp. 2558–2567.

[52] Elad, M. & Feuer, A. “Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images”. *IEEE transactions on image processing* **6**.12 (1997), pp. 1646–1658.

[53] Engan, K. et al. “Method of optimal directions for frame design”. *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. Vol. 5. 1999, 2443–2446 vol.5.

[54] Fazel, M et al. “A rank minimization heuristic with application to minimum order system approximation”. *Proceedings of the American Control Conference*. Arlington, VA: IEEE, 2001, pp. 4734–4739.

[55] Fazel, M. “Matrix Rank Minimization with Applications”. PhD thesis. Stanford University, Stanford, CA, 2002.

[56] Frey, H. et al. “Automated detection of glacier lakes based on remote sensing in view of assessing associated hazard potentials”. *Grazer Schriften der Geographie und Raumforschung* **45** (2010), pp. 261–272.

[57] Fröberg, R. “An inequality for Hilbert series of graded algebras”. *Math. Scand.* **56**.2 (1985), pp. 117–144.

[58] Fröberg, R. et al. “On the Waring problem for polynomial rings”. *Proc. Natl. Acad. Sci. USA* **109**.15 (2012), pp. 5600–5602.

[59] Gandy, S. et al. “Tensor completion and low-n-rank tensor recovery via convex optimization”. *Inverse Problems* **27**.2 (2011), pp. 025010–19.

[60] Gao, B.-C. “NDWI-A normalized difference water index for remote sensing of vegetation liquid water from space”. *Remote sensing of environment* **58**.3 (1996), pp. 257–266.

[61] Geladi, P. “Analysis of multi-way (multi-mode) data”. *Chemometrics and intelligent laboratory systems*. Vol. 7. Scandinavian symposium on chemometrics (1988). Revue: PAYS-BAS, 1989, pp. 11–30.

[62] Georghiades, A. S. et al. “From few to many: Illumination cone models for face recognition under variable lighting and pose”. *IEEE transactions on pattern analysis and machine intelligence* **23**.6 (2001), pp. 643–660.

[63] Giannakis, G. B. & Heath Jr., R. W. “Blind identification of multichannel FIR blurs and perfect image restoration”. *IEEE Trans. Image Process.* **9**.11 (2000), pp. 1877–1896.

[64] Goemans, M. X. & Williamson, D. P. “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming”. *Journal of the ACM (JACM)* **42**.6 (1995), pp. 1115–1145.

[65] Goldfarb, D. & Qin, Z. “Robust low-rank tensor recovery: models and algorithms”. *SIAM J. Matrix Anal. Appl.* **35**.1 (2014), pp. 225–253.

[66] Grellier, O. et al. “Analytical blind channel identification”. *IEEE Trans. Signal Process.* **50**.9 (2002), pp. 2196–2207.

[67] Harshman, R. A. “Foundations of the PARAFAC procedure: Models and conditions for an “explanatory" multimodal factor analysis.” *UCLA Working Papers in Phonetics* 16 (1970), pp. 1–84.

[68] Heverin, T. & Zach, L. “Twitter for city police department information sharing”. *Proceedings of the American Society for Information Science and Technology* **47**.1 (2010), pp. 1–7.

[69] Hillar, C. J. & Lim, L.-H. “Most tensor problems are NP-hard”. *J. ACM* **60**.6 (2013), Art. 45, 39.

[70] Hitchcock, F. L. “Multiple invariants and generalized rank of a p-way matrix or tensor”. *J. Math. Phys,* 7 (1927), pp. 39–79.

[71] Hitchcock, F. L. “The expression of a tensor or a polyadic as a sum of products”. *J. Math. Phys,* 6 (1927), pp. 164–189.

[72] Howell, T. D. “Global properties of tensor rank”. *Linear Algebra Appl.* **22** (1978), pp. 9–23.

[73] Hu, H. et al. “Task-driven dictionary learning for inpainting”. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Florence, Italy, 2014, pp. 3543–3547.

[74] Hu, W. et al. “A survey on visual surveillance of object motion and behaviors”. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **34**.3 (2004), pp. 334–352.

[75] Huang, J., Zhang, T., et al. “The benefit of group sparsity”. *The Annals of Statistics* **38**.4 (2010), pp. 1978–2004.

[76] Huang, T et al. “A fast two-dimensional median filtering algorithm”. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **27**.1 (1979), pp. 13–18.

[77] Hyvrinen, A. et al. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* 1st. Springer Publishing Company, Incorporated, 2009.

[78] Iarrobino, A. & Kanev, V. *Power sums, Gorenstein algebras, and determinantal loci*. Vol. 1721. Lecture Notes in Mathematics. Berlin: Springer-Verlag, 1999.

[79] Jaynes, E. T. “Information Theory and Statistical Mechanics”. *Physical Review* **106** (1957), pp. 620–630.

[80] Kailath, T. “An innovations approach to least-squares estimation–Part I: Linear filtering in additive white noise”. *IEEE transactions on automatic control* **13**.6 (1968), pp. 646–655.

[81] Knuth, D. E. *The art of computer programming. Vol. 2: Seminumerical algorithms*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont, 1969, pp. xi+624.

[82] Krizhevsky, A. et al. “Imagenet classification with deep convolutional neural networks”. *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[83] Kruskal, J. B. “Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics”. *Linear Algebra and Appl.* **18**.2 (1977), pp. 95–138.

[84] Laben, C. & Brower, B. *Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening*. US Patent 6,011,875. 2000.

[85] Landsberg, J. M. “The border rank of the multiplication of 2×2 matrices is seven”. *J. Amer. Math. Soc.* **19**.2 (2006), 447–459 (electronic).

[86] Landsberg, J. M. “Geometry and the complexity of matrix multiplication”. *Bull. Amer. Math. Soc. (N.S.)* **45**.2 (2008), pp. 247–284.

[87] Landsberg, J. M. & Manivel, L. “Generalizations of Strassen’s equations for secant varieties of Segre varieties”. *Comm. Algebra* **36**.2 (2008), pp. 405–422.

[88] Landsberg, J. M. & Teitler, Z. “On the ranks and border ranks of symmetric tensors”. *Found. Comput. Math.* **10**.3 (2010), pp. 339–366.

[89] Landsberg, J. & Ottaviani, G. “Equations for secant varieties to Veronese varieties”. preprint, arXiv:1006.0180. 2010.

[90] Landsberg, J. & Ottaviani, G. “Equations for secant varieties via vector bundles”. preprint, arXiv:1010.1825. 2010.

[91] Lasserre, J. B. “Global optimization with polynomials and the problem of moments”. *SIAM J. Optim.* **11**.3 (2001), 796–817 (electronic).

[92] Lasserre, J. B. *Moments, positive polynomials and their applications*. Vol. 1. Imperial College Press Optimization Series. Imperial College Press, London, 2010, pp. xxii+361.

[93] Lasserre, J. B. et al. “Semidefinite characterization and computation of zero-dimensional real radical ideals”. *Found. Comput. Math.* **8**.5 (2008), pp. 607–647.

[94] Lasserre, J. B. et al. “A unified approach to computing real and complex zeros of zero-dimensional ideals”. *Emerging applications of algebraic geometry*. Vol. 149. IMA Vol. Math. Appl. New York: Springer, 2009, pp. 125–155.

[95] Lee, A. B. et al. “The nonlinear statistics of high-contrast patches in natural images”. *International Journal of Computer Vision* **54**.1-3 (2003), pp. 83–103.

[96] Lim, L.-H. & Comon, P. “Nonnegative approximations of nonnegative tensors”. *Journal of Chemometrics* **23**.7-8 (2009), pp. 432–441.

[97] Lim, L.-H. & Comon, P. “Multiarray signal processing: Tensor decomposition meets compressed sensing”. *Comptes Rendus Mecanique* **338**.6 (2010), pp. 311 –320.

[98] Lin, Z. et al. “Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix”. *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)* **61**.6 (2009).

[99] Lin, Z. et al. “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices”. *arXiv preprint arXiv:1009.5055* (2010).

[100] Liu, D. & Boufounos, P. “Dictionary learning based pan-sharpening”. *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. 2012, pp. 2397–2400.

[101] Liu, J et al. “Tensor completion for estimating missing values in visual data”. *Proc. of the 12th Int. Conf. Computer Vision (Kyoto, 2009)*. Kyoto, 2009, pp. 211–21.

[102] Loncan, L. et al. “Hyperspectral pansharpening: a review”. *IEEE Geoscience and Remote Sensing Magazine* (2015), Accepted.

[103] Lounici, K. et al. “Taking advantage of sparsity in multi-task learning”. *arXiv preprint arXiv:0903.1468* (2009).

[104] Lowe, D. G. “Object recognition from local scale-invariant features”. *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Vol. 2. Ieee. 1999, pp. 1150–1157.

[105] Mairal, J. et al. “Online dictionary learning for sparse coding”. *Proceedings of the 26th annual international conference on machine learning*. ACM. 2009, pp. 689–696.

[106] Mairal, J. et al. “Task-driven dictionary learning”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**.4 (2012), pp. 791–804.

[107] Mallat, S. & Hwang, W. L. “Singularity detection and processing with wavelets”. *IEEE transactions on information theory* **38**.2 (1992), pp. 617–643.

[108] Malouf, R. “A comparison of algorithms for maximum entropy parameter estimation”. *proceedings of the 6th conference on Natural language learning-Volume 20*. Association for Computational Linguistics. 2002, pp. 1–7.

[109] McCullagh, P. *Tensor Methods in Statistics*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1987.

[110] Mella, M. “Base loci of linear systems and the Waring problem”. *Proc. Amer. Math. Soc.* **137**.1 (2009), pp. 91–98.

[111] Mourrain, B. & Pan, V. Y. “Multidimensional structured matrices and polynomial systems”. *Calcolo* **33**.3 (1996), pp. 389–401.

[112] Navasca, C & De Lathauwer, L. “Low multilinear tensor approximation via semidefinite programming”. *Proc. of the 17th European Signal Processing Conference (EUSIPCO 2009)*. Glasgow, Scotland, 2009, pp. 520–524.

[113] Nguyen, N. et al. “A computationally efficient superresolution image reconstruction algorithm”. *IEEE transactions on image processing* **10**.4 (2001), pp. 573–583.

[114] Obozinski, G. et al. “Joint covariate selection and joint subspace selection for multiple classification problems”. *Statistics and Computing* **20**.2 (2010), pp. 231–252.

[115] Oeding, L. & Ottaviani, G. “Eigenvectors of tensors and algorithms for Waring decomposition”. preprint, arXiv:1105.1229. 2011.

[116] Pan, S. J. & Yang, Q. “A survey on transfer learning”. *IEEE Transactions on knowledge and data engineering* **22**.10 (2010), pp. 1345–1359.

[117] Pardue, K. “Generic sequences of polynomials”. *J. Algebra* **324**.4 (2010), pp. 579–590.

[118] Park, S. C. et al. “Super-resolution image reconstruction: a technical overview”. *IEEE Signal Processing Magazine* **20**.3 (2003), pp. 21–36.

[119] Pati, Y. C. et al. “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition”. *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE. 1993, pp. 40–44.

[120] Phillips, S. J. & Dudík, M. “Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation”. *Ecography* **31**.2 (2008), pp. 161–175.

[121] Recht, B. et al. “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization”. *SIAM Rev.* **52**.3 (2010), pp. 471–501.

[122] Reichstein, B. “On Waring’s problem for cubic forms”. *Linear Algebra Appl.* **160** (1992), pp. 1–61.

[123] Reznick, B. “Sums of even powers of real linear forms”. *Mem. Amer. Math. Soc.* **96**.463 (1992), pp. viii+155.

[124] Richardson, W. H. “Bayesian-based iterative method of image restoration”. *JOSA* **62**.1 (1972), pp. 55–59.

[125] Rudin, L. I. & Osher, S. “Total variation based image restoration with free local constraints”. *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*. Vol. 1. IEEE. 1994, pp. 31–35.

[126] Rudin, L. I. et al. “Nonlinear total variation based noise removal algorithms”. *Physica D: Nonlinear Phenomena* **60**.1-4 (1992), pp. 259–268.

[127] Santosa, F. & Symes, W. W. “Linear inversion of band-limited reflection seismograms”. *SIAM Journal on Scientific and Statistical Computing* **7**.4 (1986), pp. 1307–1330.

[128] Saunderson, J. et al. “Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting”. *SIAM J. Matrix Anal. Appl.* **33**.4 (2012), pp. 1395–1416.

[129] Scherer, D. et al. “Evaluation of pooling operations in convolutional architectures for object recognition”. *Artificial Neural Networks–ICANN 2010* (2010), pp. 92–101.

[130] Sidiropoulos, N. D. et al. “Parallel Factor Analysis in Sensor Array Processing”. *IEEE Transactions on Signal Processing* **48**.8 (2000), pp. 2377–2388.

[131] Signoretto, M. et al. “Nuclear Norms for Tensors and Their Use for Convex Multilinear Estimation”. preprint, Internal Report 10-186, ESAT-SISTA, K.U.Leuven (Leuven, Belgium). 2011.

[132] Silva, V. d. & Lim, L.-H. “Tensor rank and the ill-posedness of the best low-rank approximation problem”. *SIAM Journal on Matrix Analysis and Applications* **30**.3 (2008), pp. 1084–1127.

[133] Skau, E. et al. “Pansharpening via coupled triple factorization dictionary learning”. *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE. 2016, pp. 1234–1237.

[134] Smelyanskiy, V. et al. “Bayesian super-resolved surface reconstruction from images”. *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*. Vol. 1. IEEE. 2000, pp. 375–382.

[135] Smilde, A. K. et al. *Multi-way analysis with applications in the chemical sciences*. John Wiley and Sons., 2004.

[136] Stathaki, T. *Image fusion: algorithms and applications*. Academic Press, 2011.

[137] Strassen, V. “Rank and optimal computation of generic tensors”. *Linear Algebra Appl.* **52/53** (1983), pp. 645–685.

[138] Strassen, V. “Gaussian elimination is not optimal”. *Numer. Math.* **13** (1969), pp. 354–356.

[139] Sylvester, J. J. “Sur une extension d’un thèorème de Clebsch relatif aux courbes du quatrième degré.” *Comptes Rendus* **102** (1886), pp. 1532–1534.

[140] Taylor, H. L. et al. “Deconvolution with the ℓ 1 norm”. *Geophysics* **44**.1 (1979), pp. 39–52.

[141] Tibshirani, R. “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[142] Tillmann, A. M. “On the computational intractability of exact and approximate dictionary learning”. *IEEE Signal Processing Letters* **22**.1 (2015), pp. 45–49.

[143] Tropp, J. A. “Just relax: Convex programming methods for identifying sparse signals in noise”. *IEEE transactions on information theory* **52**.3 (2006), pp. 1030–1051.

[144] Tsai, A. et al. “Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification”. *IEEE transactions on Image Processing* **10**.8 (2001), pp. 1169–1186.

[145] Tucker, L. R. “Implications of factor analysis of three-way matrices for measurement of change”. *Problems in measuring change*. Ed. by Harris, C. W. University of Wisconsin Press, 1963, pp. 122–137.

[146] Tucker, L. R. “Some mathematical notes on three-mode factor analysis”. *Journal Psychometrika* **31**.3 (1966), pp. 279–311.

[147] Vandenberghe, L. & Boyd, S. “Semidefinite programming”. *SIAM review* **38**.1 (1996), pp. 49–95.

[148] Verpoorter, C. et al. “Automated mapping of water bodies using Landsat multispectral data”. *Limnol. Oceanogr. Methods* **10** (2012), pp. 1037–1050.

[149] Vese, L. A. & Chan, T. F. “A multiphase level set framework for image segmentation using the Mumford and Shah model”. *International journal of computer vision* **50**.3 (2002), pp. 271–293.

[150] Vieweg, S. et al. “Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness”. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’10. Atlanta, Georgia, USA: ACM, 2010, pp. 1079–1088.

[151] Wald, L. “Quality of high resolution synthesised images: Is there a simple criterion?” *Third conference ”Fusion of Earth data: merging point measurements, raster maps and remotely sensed images”*. Ed. by Ranchin, T. & Wald, L. Sophia Antipolis, France: SEE/URISCA, 2000, pp. 99–103.

[152] Wang, S. et al. “Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis”. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. 2012, pp. 2216–2223.

[153] Watanabe, J. “The Dilworth number of Artinian rings and finite posets with rank function”. *Advanced Studies in Pure Mathematics, Commutative Algebra and Combinatorics* **11** (1987), pp. 303–312.

[154] Wohlberg, B. “Boundary Handling for Convolutional Sparse Representations”. *Proceedings of IEEE International Conference on Image Processing (ICIP)*. Phoenix, AZ, USA, 2016, pp. 1833–1837.

[155] Wohlberg, B. “Efficient Algorithms for Convolutional Sparse Representations”. *IEEE Transactions on Image Processing* **25**.1 (2016), pp. 301–315.

[156] Wohlberg, B. & Rodriguez, P. “An ℓ 1-TV algorithm for deconvolution with salt and pepper noise”. *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE. 2009, pp. 1257–1260.

[157] Wright, J. et al. “Robust face recognition via sparse representation”. *IEEE transactions on pattern analysis and machine intelligence* **31**.2 (2009), pp. 210–227.

[158] Wright, J. et al. “Sparse representation for computer vision and pattern recognition”. *Proceedings of the IEEE* **98**.6 (2010), pp. 1031–1044.

[159] Xu, H. “Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery”. *International Journal of Remote Sensing* **27**.14 (2006), pp. 3025–3033.

[160] Yang, J. et al. “Feature fusion: parallel strategy vs. serial strategy”. *Pattern Recognition* **36**.6 (2003), pp. 1369–1381.

[161] Yuan, M. & Lin, Y. “Model selection and estimation in regression with grouped variables”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**.1 (2006), pp. 49–67.

[162] Zeiler, M. D. et al. “Deconvolutional networks”. *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE. 2010, pp. 2528–2535.

[163] Zhao, Y.-B. “An approximation theory of matrix rank minimization and its application to quadratic equations”. *Linear Algebra and its Applications* **437**.1 (2012), pp. 77–93.